



**HAL**  
open science

# Développement et comparaison d'approches QSPR-inverse

Philippe Gantzer

► **To cite this version:**

Philippe Gantzer. Développement et comparaison d'approches QSPR-inverse. Bio-informatique [q-bio.QM]. Sorbonne Université, 2021. Français. NNT : 2021SORUS254 . tel-03531086

**HAL Id: tel-03531086**

**<https://theses.hal.science/tel-03531086>**

Submitted on 18 Jan 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Sorbonne Université

Ecole doctorale 388 - Ecole doctorale de Chimie Physique et de Chimie

Analytique de Paris Centre

*IFP Energies nouvelles / Direction Physico-Chimie et Mécanique appliquées*

## **Développement et comparaison d'approches QSPR-inverse**

Par Philippe Gantzer

Thèse de doctorat de Physico-Chimie

Dirigée par Carlos Nieto-Draghi

Co-dirigée par Benoît Creton

Présentée et soutenue publiquement le 4 novembre 2021

Devant un jury composé de :

Pr. Ronan Bureau	CERMN – Université Caen Normandie	Rapporteur
Dr. James Devillers	CTIS	Rapporteur
Pr. Pascal Bonnet	ICOA – Université d'Orléans	Examineur
Pr. Jean-Loup Faulon	Micalis – INRAE	Examineur
Dr. Patricia Rotureau	INERIS	Examineur
Dr. Marianna Yiannourakou	Materials Design	Examineur
Dr. Carlos Nieto-Draghi	IFP Energies nouvelles	Examineur
Dr. Benoît Creton	IFP Energies nouvelles	Examineur





## Remerciements

Je voudrais remercier dans un premier temps IFP Energies nouvelles, grâce à qui j'ai pu réaliser cette thèse. Particulièrement, Benoît Creton et Carlos Nieto-Draghi m'ont permis grâce à leurs nombreuses idées, conseils et leur motivation à faire face aux défis de ces trois années. Merci aux autres membres de mon département, actuels et anciens, pour m'avoir accueilli à part entière dans la vie de l'équipe. Merci à Delphine Sinoquet pour sa précieuse aide dans l'utilisation de l'optimisateur HubOpt. Merci également aux autres doctorants et à l'ADIFP, ainsi qu'à toutes les personnes ayant participé de près ou de loin à mon intégration.

Je remercie également les membres du laboratoire de Chémoinformatique de Strasbourg, pour m'avoir permis de collaborer avec eux, et utiliser certains des outils développés dans leur laboratoire durant ma thèse. Leurs nombreux conseils m'ont également permis de progresser dans mon travail, notamment dans la comparaison des méthodes de génération.

Je tiens également à remercier les membres du jury, les rapporteurs Pr. Ronan Bureau et Dr. James Devillers, ainsi que les examinateurs Pr. Pascal Bonnet, Pr. Jean-Loup Faulon, Dr. Patricia Rotureau et Dr. Marianna Yiannourakou pour avoir accepté d'évaluer mon travail de thèse.

Ce travail n'aurait pas pu être réalisé dans les meilleures conditions sans le soutien de mes proches. Merci à vous.



# Sommaire

Remerciements .....	2
Sommaire .....	4
Chapitre 1. Introduction .....	8
1.1 Contexte.....	8
1.2 La Chémoinformatique et ses outils .....	8
1.2.1 La Chémoinformatique .....	8
1.2.2 Les bases de données.....	9
1.2.3 Les relations quantitatives structures à propriété (QSPR) .....	10
1.2.4 La génération de nouvelles molécules .....	12
1.3 Objectifs de la thèse.....	13
Chapitre 2. Modélisation QSPR.....	15
2.1 Méthodologie QSPR.....	15
2.1.1 Principes de QSPR .....	16
2.1.2 Flux opérationnel pour la création des QSPR .....	28
2.2 Modèles construits au cours de la thèse.....	30
2.2.1 Jeu de données.....	31
2.2.2 Modélisation par SVR.....	32
2.2.3 Représentation des données par ACP.....	36
2.2.4 Représentation des données et modélisation par GTM .....	38
2.3 Conclusions sur la partie modélisation QSPR.....	40
Chapitre 3. Méthodes pour la génération moléculaire et pour leur comparaison .....	42
3.1 Introduction sur les méthodes de génération et leur comparaison.....	42
3.2 Génération non guidée de molécules .....	43
3.3 Génération guidée des molécules .....	44

3.4	Génération de molécules par apprentissage profond.....	46
3.5	Evaluation des méthodes de génération.....	49
3.6	Conclusions sur la partie méthodes pour la génération .....	53
Chapitre 4. Génération virtuelle de molécules.....		55
4.1	Evaluation de la qualité des générations.....	55
4.2	Génération par assemblage de fragments .....	57
4.2.1	Assemblage de fragments simples .....	58
4.2.2	Assemblage de fragments SMF.....	61
4.2.3	Conclusions sur la génération par assemblage de fragments .....	64
4.3	Génération par modifications successives des structures .....	65
4.3.1	Ajout et suppression de fragments .....	67
4.3.2	Croisement de graphes .....	70
4.3.3	Mutation de liaisons et d'atomes.....	72
4.3.4	Cyclisation et ouverture de cycles.....	74
4.3.5	Conclusions sur la génération par modification successives des structures.....	76
4.4	Génération par apprentissage profond .....	78
4.4.1	Génération de molécules à partir d'une carte GTM construite avec les vecteurs latents	79
4.4.2	Génération de molécules à partir des vecteurs latents bruités.....	80
4.4.3	Conclusion sur la génération par apprentissage profond .....	82
4.5	Conclusions sur la génération par apprentissage profond .....	83
Chapitre 5. Comparaison des générations moléculaires .....		85
5.1	Données et comparaisons préliminaires .....	85
5.1.1	Méthodes de génération comparées .....	85
5.1.2	Comparaisons des distributions de propriétés.....	86
5.2	Indices pour la comparaison des générations .....	88

5.2.1	Utilisation de l'espace $\mathbb{C}$ pour la comparaison moléculaire .....	88
5.2.2	Couverture de l'espace .....	91
5.2.3	Représentativité de la couverture de l'espace .....	92
5.2.4	Uniformité de la couverture de l'espace .....	93
5.2.5	Spécificité de la génération .....	94
5.3	Comparaison des méthodes de génération à l'aide des nouveaux indices .....	95
5.3.1	Génération de molécules diverses .....	95
5.3.2	Génération de molécules diverses possédant leur point d'éclair dans un intervalle défini	100
5.4	Conclusions sur la comparaison des générations moléculaires .....	108
Chapitre 6. Conclusions et Perspectives .....		110
6.1	Conclusions .....	110
6.2	Perspectives .....	113
Annexe A. Molécules de la base de données sur le PE .....		115
Annexe B. Inverse-QSPR for <i>de novo</i> Design: A Review .....		145
Annexe C. Génération de molécules par assemblages de fragments et contraintes sur la propriété		146
C.1.	Méthodes F supplémentaires .....	146
C.1.1.	Méthode F2a .....	146
C.1.2.	Méthode F2b .....	147
C.2.	Performances des générations de F avec contraintes sur la propriété .....	149
Annexe D. Choix de la taille des cubes discrétisant $\mathbb{C}$ .....		152
D.1.	Couverture du sous-espace de $\mathbb{C}$ regroupant les molécules initiales .....	152
D.2.	Variation des valeurs d'indices pour une même méthode de génération .....	154
Chapitre 7. Bibliographie .....		157
Liste des figures .....		171
Liste des tableaux .....		174





# Chapitre 1. Introduction

---

## 1.1 Contexte

IFP Energies nouvelles (IFPEN) est un établissement public national à caractère industriel et commercial. Contribuant initialement à la recherche pour l'industrie pétrolière, il est aujourd'hui un acteur important du « développement des technologies et matériaux du futur » concernant « les domaines de l'énergie, du transport et de l'environnement ». <sup>1</sup> Ces domaines se traduisent en quatre axes de recherche : (i) la mobilité durable, (ii) les énergies renouvelables, (iii) les hydrocarbures responsables, et (iv) le climat, l'environnement et l'économie circulaire. <sup>2</sup> La chimie est impliquée dans chacun de ces axes : nous citerons par exemple des applications concernant les batteries des véhicules (axes i et ii), les biocarburants et la transformation de la biomasse (axe ii), la pétrochimie et la récupération assistée du pétrole (axe iii), ou encore la capture du CO<sub>2</sub> ainsi que le recyclage des plastiques et des métaux (axe iv).

Le développement et l'usage d'outils informatiques sont aujourd'hui devenus incontournables pour stocker et exploiter les informations. Des données sont en effet régulièrement produites, par exemple dans le domaine de la recherche : informations sur les méthodologies/procédés, sur le climat, l'économie, ou encore concernant la chimie. En chimie, des molécules sont régulièrement découvertes, identifiées, synthétisées, ou caractérisées. Ma thèse s'inscrit dans ce cadre.

## 1.2 La Chémoinformatique et ses outils

### 1.2.1 La Chémoinformatique

La Chémoinformatique est un domaine scientifique récent qui consiste en la résolution de problèmes de la chimie à l'aide des outils de l'informatique. La notion de Chémoinformatique est apparue la première fois dans la littérature en 1998. <sup>3</sup> Elle y est définie comme un moyen de « transformer les données en informations et les informations en connaissances », c'est « la discipline qui va réduire le temps nécessaire de plusieurs semaines à plusieurs jours » pour

rechercher des molécules répondant à des caractéristiques spécifiques.<sup>3</sup> Dans ses prémices, la Chémoinformatique a employé des méthodes « qui ne fonctionnent pas bien ensemble » : par exemple, l'utilisateur devait manuellement adapter le format des données d'un programme à l'autre.<sup>3</sup> La définition de normes, notamment sur le format des fichiers, était alors nécessaire pour pouvoir traiter et exploiter plus efficacement les données à travers plusieurs programmes; surtout que le nombre de données disponible ne cesse d'augmenter.<sup>4</sup> Aujourd'hui, selon la déclaration d'Obernai, les méthodes de la Chémoinformatique regroupent les outils pour : construire les bases de données chimiques, prédire les propriétés des molécules, concevoir des médicaments, élucider les structures chimiques, prédire la réactivité des molécules, et planifier des synthèses organiques.<sup>5</sup>

### **1.2.2 Les bases de données**

Des bases de données (BDD) régulièrement mises à jour sont utilisées pour regrouper les informations chimiques : structures moléculaires, caractéristiques et propriétés. Le nombre de bases disponibles à tous est difficile à quantifier, Apodaca<sup>6</sup> fait par exemple état d'une soixantaine de bases accessibles gratuitement sur Internet tandis que d'autres sources en listent plus du double, en incluant les bases payantes.<sup>7</sup> Parmi elles, les bases de données PubChem<sup>8</sup> et ZINC<sup>9</sup> contiennent chacune plus de 10<sup>8</sup> entrées. La base ChEMBL<sup>10</sup> contient quant à elle près de 2 millions de composés et référence plus de 15 millions de valeurs d'activités biologiques. Le contenu des bases de données varie en fonction des molécules et des propriétés référencées. Deux bases de données peuvent informer sur les mêmes molécules mais contenir des informations différentes (par exemple la première base peut informer sur les spectres RMN et la seconde sur des propriétés). Deux bases de données peuvent également décrire les mêmes propriétés mais contenir des molécules différentes. En effet, une étude récente a démontré la complémentarité chimique des bases PubChem et ChEMBL, qui possèdent chacune des molécules uniques malgré leur volonté d'être toutes deux exhaustives.<sup>11</sup> La base SciFinder<sup>12</sup> est à notre connaissance la base la plus complète avec plus de 142 millions de composés référencés<sup>13</sup>. L'exportation de ses données n'est possible que de façon limitée, restreignant son utilisation à des recherches manuelles. Il existe également des BDD privées. Ces bases sont utilisables par un nombre restreint de personnes, et de ce fait ne sont pas discutées dans ce manuscrit.

Du fait du grand nombre de BDD disponibles, de leur taille et de la diversité d'information présente dans chacune d'entre elles, il est difficile d'y chercher manuellement une molécule pour une application spécifique. Cela peut se comparer à chercher une aiguille dans une botte de foin. Cette difficulté s'accroît avec le nombre de contraintes sur les propriétés à respecter. La Chémoinformatique permet de faciliter cette recherche avec les relations quantitatives structures à propriétés, décrites ci-dessous.

### 1.2.3 Les relations quantitatives structures à propriété (QSPR)

Une des voies d'études de la Chémoinformatique est l'identification de corrélations entre la structure et les propriétés de composés chimiques. Des modèles reliant structures et propriétés (QSPR, pour « Quantitative Structure Property Relationship ») sont ainsi créés sur l'hypothèse de base que les molécules structurellement similaires possèdent des propriétés similaires. Nous incluons également les relations quantitatives structures à activité biologique (QSAR, pour « Quantitative Structure Activity Relationship ») et propriété à propriété (QSPP, pour « Quantitative Property Property Relationship ») dans le terme QSPR. En modélisation QSPR, les structures moléculaires peuvent être encodées à l'aide de descripteurs. Chaque descripteur traduit une information moléculaire ou structurelle. Les descripteurs ( $X$ ) sont ensuite utilisés par une méthode d'apprentissage automatique (MAA) qui agit comme une fonction ( $f$ ) pour prédire la propriété ( $Y$ ) (équation (1)).

$$Y = f(X) \quad (1)$$

#### 1.2.3.1 Les QSPR dans l'industrie

Les QSPR représentent des outils intéressants pour prédire la valeur des propriétés de molécules, notamment dans l'industrie. En effet, comparées aux manipulations expérimentales, les QSPR permettent un gain de temps, une réduction des coûts et la production de moins de déchets.<sup>14</sup> Les QSPR sont autant utilisées lors du référencement de nouvelles molécules (pour renseigner leurs valeurs de propriétés) que pour aider à la sélection de molécules intéressantes au sein des bases de données.

Le référencement des propriétés moléculaires est important pour pouvoir estimer la réactivité et/ou la toxicité des molécules en amont de leur utilisation et ainsi pouvoir les utiliser en toute sécurité. Par exemple, le potentiel d'oxydoréduction est une grandeur qui permet d'estimer la faisabilité d'une réaction d'oxydation ou de réduction d'une espèce chimique. Si une molécule réductrice possède un potentiel d'oxydoréduction inférieure à celle du couple H<sub>2</sub>O/H<sub>2</sub>, elle peut être oxydée par l'eau et produire du dihydrogène, un gaz inflammable. D'autre part, la réglementation européenne REACH (pour « Registration, Evaluation and Authorisation of Chemicals ») demande dorénavant d'enregistrer les substances chimiques utilisées ou produites dans des quantités supérieures à une tonne par an.<sup>15</sup> Cet enregistrement nécessite de renseigner certaines valeurs de propriétés physicochimiques et (éco)toxicologiques, qui pour certaines peuvent être prédites par des modèles QSPR.<sup>16</sup> Le système international de classification et étiquetage des produits chimiques autorise également l'usage des QSPR pour renseigner sur les propriétés.<sup>17</sup> L'article de Quintero et *coll.* passe en revue certains des modèles proposés dans ces optiques.<sup>18</sup> Bien que l'on trouve actuellement un nombre important de modèles QSPR utilisables sur Internet, des entreprises créent leurs propres modèles. Par exemple, elles utilisent leurs propres données expérimentales (privées), et les combinent aux données issues de bases publiques, pour concevoir des modèles « sur-mesure » capables de prédire efficacement les propriétés de molécules similaires à celles qu'elles utilisent.<sup>19,20</sup>

Le criblage virtuel est une des méthodes existantes pour rechercher des molécules intéressantes dans les bases de données. Cette méthode consiste à prédire les propriétés de toutes les molécules puis à ne sélectionner que celles dont les valeurs de propriété satisfont à un ou des critère(s). Le criblage virtuel peut être considéré comme un tri préliminaire pour écarter les molécules inintéressantes.<sup>21</sup> Nous présentons ci-après quelques propriétés couramment utilisées en criblage virtuel et prédictibles par QSPR. Les propriétés d'absorption, distribution, métabolisme, excrétion et toxicité (ADME-Tox) renseignent sur la métabolisation et la toxicité physiologique des molécules ; ces propriétés sont surtout employées par l'industrie pharmaceutique qui recherche des molécules adaptées à une utilisation biologique.<sup>22</sup> De manière plus générale, l'accessibilité synthétique évalue la difficulté de synthèse des molécules et son utilisation en criblage permet d'écarter les molécules difficilement synthétisables.<sup>23</sup>

### 1.2.3.2 Les QSPR à IFPEN

A IFPEN, la modélisation QSPR est d'ores et déjà employée comme alternative à certaines expérimentations et modélisations.<sup>24</sup> Un premier exemple de modélisation concerne

l'augmentation du taux de récupération du pétrole, devenu un enjeu important pour répondre à la demande croissante alors que les réserves de pétrole ne sont pas illimitées. L'extraction du pétrole peut se décomposer en trois étapes.<sup>25</sup> À l'issue des deux premières étapes, un taux d'extraction d'environ 45% peut être atteint. Des techniques de récupération « tertiaire » (EOR, Enhanced Oil Recovery) peuvent alors être appliquées pour améliorer la récupération de 2 à 20%.<sup>26</sup> L'EOR consiste à modifier les propriétés des fluides pour mobiliser les gouttelettes d'huile piégées dans les pores de la roche. Parmi elles, l'EOR par voie chimique consiste en l'injection de formulations tensioactives dont le but est de diminuer la tension interfaciale entre l'eau salée et l'huile et ainsi de dépiéger l'huile des pores. Il existe une dépendance de la tension interfaciale huile/eau en fonction de la teneur en sel de cette dernière. La salinité optimale  $S^*$  du système est définie comme la concentration en sel de la phase aqueuse pour laquelle la tension interfaciale est minimale, menant à une récupération optimale de l'huile. Un modèle QSPR a été construit à IFPEN pour aider à la conception des mélanges de tensioactifs ; il prédit la valeur  $S^*$  de ces mélanges.<sup>27</sup>

Le second exemple concerne les structures d'imidazolates zéolithiques (ZIF). Les ZIF sont des complexes métalliques permettant de sélectivement capturer le  $CO_2$  dans des mélanges de gaz,<sup>28</sup> évitant ainsi de le libérer dans l'atmosphère. Le modèle QSPR crée a permis d'étudier la personnalisation des ZIF par des ligands organiques pour améliorer le taux de capture du  $CO_2$ .<sup>29,30</sup>

Le troisième exemple de modélisation concerne les carburants alternatifs, dont les biocarburants font partie. Ces carburants doivent posséder des spécifications similaires à celles des carburants conventionnels pour permettre leur utilisation sans modifier le moteur utilisé. Des modèles QSPR ont été créés pour prédire les valeurs de plusieurs propriétés d'usage des biocarburants : point d'éclair, indice de cétane, point de fusion, chaleur nette de combustion, densité et viscosité, ces deux derniers en fonction de la température.<sup>31</sup>

#### **1.2.4 La génération de nouvelles molécules**

Des efforts considérables sont effectués pour concevoir des bases de molécules les plus complètes possibles, mais la quantité de molécules connues à ce jour ne représente qu'une petite partie de l'ensemble des structures moléculaires pouvant être théoriquement synthétisées. Si l'on considère tous les assemblages moléculaires réalisables contenant jusqu'à trente atomes de carbone, azote, oxygène ou soufre, Bohacek *et coll.* ont estimé que le nombre théorique de

molécules pouvant être construites est supérieur à  $10^{60}$ .<sup>32,33</sup> Des molécules candidates prometteuses sont alors inexorablement absentes lors du criblage virtuel de BDD composées de molécules existantes.

Une alternative au criblage virtuel des BDD consiste à proposer et cribler des structures innovantes, inconnues des BDD. Des techniques de génération virtuelle de molécules sont développées depuis moins d'une trentaine d'années dans cette optique.<sup>34</sup> L'utilisation de contraintes sur les valeurs de propriété permet aux algorithmes d'éviter de générer des molécules ne possédant pas les caractéristiques souhaitées. Par exemple, l'inversion de modèles QSPR (i-QSPR, pour « inverse-QSPR ») est une des techniques existantes permettant de générer et d'identifier des molécules répondant à une ou des propriété(s) donnée(s).

### 1.3 Objectifs de la thèse

Le travail présenté dans ce manuscrit de thèse concerne l'implémentation et la comparaison de méthodes de génération de molécules. Nous avons concentré nos travaux sur la génération de molécules dites « corps pur ». Les différentes parties du travail proposé sont schématisées sur la Figure 1.

Le deuxième chapitre du manuscrit décrit autant les améliorations effectuées sur la méthodologie QSPR existante pour la prédiction des propriétés que les approches utilisées pour représenter les BDD de molécules. Les outils construits (modèles QSPR et représentations) sont utilisés pour présenter nos données et leur diversité, pour prédire la valeur de propriété et/ou pour être inversés.

Dans le troisième chapitre, nous décrivons l'état de l'art sur la génération virtuelle de molécules. Ce travail bibliographique ayant été publié dans un journal au cours de la thèse,<sup>34</sup> nous n'en présentons ici qu'un résumé. Nous complétons cet état de l'art avec l'analyse de nouveaux travaux. La discussion suivant cette analyse de la littérature permet de dresser les choix de recherche pour nos travaux, c'est-à-dire : sur les méthodes de génération à utiliser, adaptées aux données disponibles à IFPEN, et sur le besoin de mettre en place une nouvelle méthode de comparaison des générations.

Le quatrième chapitre décrit les méthodes de génération sélectionnées, implémentées et améliorées au cours de la thèse : (i) par assemblage de fragments (F), (ii) par modifications

successives de structures (G) et (iii) par apprentissage profond (E). Pour chaque méthode et amélioration, les performances de génération sont étudiées et discutées.

Le cinquième chapitre est dédié à la comparaison des méthodes de génération. Une série d'indices a été développée pour permettre de comparer les capacités de génération de chaque méthode, en fonction de leur (i) couverture de l'espace chimique, de leur (ii) représentativité de l'espace chimique et de leur (iii) spécificité de génération quand une valeur ou un intervalle de valeurs de propriété est ciblé. Les capacités de génération des méthodes améliorées F et G sont évaluées à partir de ces indices.

Enfin, nous récapitulons les résultats obtenus au cours de la thèse et exposons nos conclusions et perspectives dans le dernier chapitre du manuscrit.

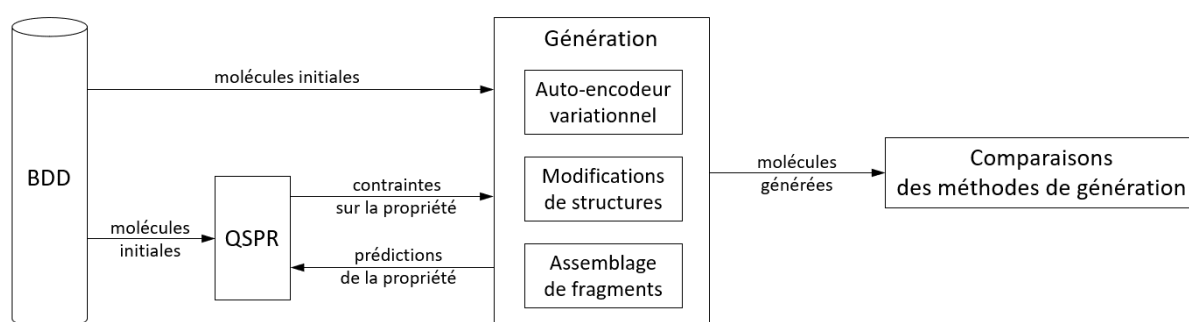


Figure 1 : Schéma représentatif des travaux présentés dans ce manuscrit de thèse.



## Chapitre 2. Modélisation QSPR

---

Ce chapitre est dédié au travail de modélisation QSPR, effectué en préambule de la génération de molécules. La première partie de ce chapitre expose les outils de modélisation utilisés ainsi que les méthodologies mises en place pour obtenir les modèles. Dans la seconde partie, nous présentons la base de données considérée, y appliquons la méthodologie précédemment exposée pour créer nos modèles, et analysons les modèles créés. Enfin, nous dressons les conclusions concernant la modélisation et les modèles développés, ces derniers nous serviront ensuite pour prédire les propriétés des molécules générées et guider les générations de structures.

### 2.1 Méthodologie QSPR

La partie « Méthodologie QSPR » regroupe les « Principes de QSPR » et le « Flux opérationnel pour la création des QSPR » utilisés. L'information donnée dans « Principes de QSPR » reprend les étapes de création d'un modèle QSPR en présentant les outils utiles au travail de thèse. Comme nous le verrons, chaque molécule est d'abord représentée et encodée en un vecteur de descripteurs pour la rendre interprétable par les programmes informatiques. Des méthodes d'apprentissage automatique (MAA) sont ensuite utilisées pour modéliser les propriétés des molécules à partir des vecteurs descripteurs. Différents outils, présentés à la suite, sont utilisés pour analyser et améliorer la qualité des modèles. Enfin, l'espace chimique dans lequel les prédictions par QSPR sont considérées fiables, appelé « Domaine d'Applicabilité », est défini. Dans « Flux opérationnel pour la création des QSPR », nous présentons les scripts qui ont été implémentés pour automatiser la création de trois types de QSPR différents.

## 2.1.1 Principes de QSPR

### 2.1.1.1 Représentations moléculaires

Il existe différentes manières de présenter une molécule : soit par des représentations textuelles avec leur nom, leur formule chimique, soit par leur transformation en objets tels qu'un graphe moléculaire ou une représentation 3D, ou encore à l'aide d'encodages spécifiques. Quelle que soit la représentation choisie, il est ensuite très souvent nécessaire de transformer les molécules en descripteurs pour être interprétables par les méthodes courantes de modélisation.

#### 2.1.1.1.1 Représentation des structures chimiques

La nomenclature chimique est une manière textuelle de représenter et cataloguer les molécules. Au XVIII<sup>e</sup> siècle, alors que la chimie inorganique connaît un essor important, Guyton-Morveau discute du fait que « la langue de la Chymie (sic) a besoin d'être réformée pour la plus grande partie », car « les plus célèbres Chymistes (sic) [...] n'ont cessé de déplorer la confusion, l'obscurité, la gêne qu'une foule de noms impropres portoit (sic) dans leur langage ». <sup>35</sup> En effet, les noms utilisés pour mentionner les substances chimiques étaient très souvent empruntés du latin, d'observations, voire de l'alchimie. Le carbone était nommé « carbonis » ou charbon ; le dioxyde de carbone était nommé soit « sylvestre spiritus », esprit sauvage, soit air fixe, en opposition à l'air vital qui désigne le dioxygène ; l'acétate de plomb était nommé « Sucre de Saturne » à cause de son goût sucré <sup>36</sup> et du rapprochement entre le plomb et la planète Saturne en alchimie. <sup>37,38</sup> Il proposa ainsi avec Lavoisier *et coll.* la *Méthode de Nomenclature Chimique*. <sup>37</sup> Ce traité propose des règles et codes de nomenclature qui permettent notamment de distinguer les éléments chimiques purs des composés chimiques. Les composés chimiques sont dès lors nommés en faisant référence aux éléments chimiques les constituants. La réforme de la nomenclature a été étendue à la chimie organique en 1889. Dès 1892, une commission dédiée à cette tâche s'est formée à Genève. <sup>39</sup> Depuis, l'union internationale de chimie pure et appliquée (IUPAC, pour International Union of Pure and Applied Chemistry) définit les règles internationales de nomenclature chimique. <sup>40</sup> Elles sont regroupées dans un « Red Book » pour la chimie inorganique, un « Blue book » pour la chimie organique et un « Purple book » pour la chimie des polymères. <sup>41</sup> Il existe aujourd'hui plusieurs autres nomenclatures, en fonction des variations linguistiques et des noms triviaux donnés aux molécules. Par exemple, la molécule composée d'une chaîne de deux atomes de carbone (ethan-) liés à une fonction alcool (-ol) est nommée : ethanol selon la nomenclature IUPAC, éthanol ou alcool éthylique en français,

Ethylalkohol en allemand.<sup>42</sup> Les codes de la nomenclature chimique nécessitent d'être connus des chimistes pour pouvoir visualiser rapidement les structures moléculaires mentionnées.

On peut également représenter les molécules en décrivant leur structure chimique. La formule chimique brute est la représentation qui indique uniquement le nombre de chaque type d'atomes. Dans les formules développées (ou semi-développées), l'arrangement des atomes (ou groupes d'atomes) est précisé et les liaisons formées entre ces atomes (groupes) sont explicitées. Les graphes moléculaires sont des figures représentant les molécules à partir de leur formule développée. Les sommets des graphes correspondent aux atomes et les arêtes aux liaisons chimiques. Dans de telles représentations, les isomères de constitution – molécules possédant la même formule chimique brute, mais un agencement différent de leurs atomes – sont distingués. Toute l'information sur la géométrie des molécules est obtenue si on ajoute aux graphes les informations sur l'agencement 3D ou si on utilise directement des représentations 3D. Dans ces cas, on différencie les isomères de configuration – les molécules qui possèdent le même graphe moléculaire, mais un agencement des atomes dans l'espace différent. Ces représentations sont majoritairement utilisées quand on a besoin de visualiser la structure des molécules, par exemple pour comprendre le mécanisme d'une réaction chimique. Elles ne sont pas adaptées pour cataloguer les molécules : elles possèdent soit peu d'information sur l'isomérisation (par exemple pour la formule brute ou semi-développée) ou ne sont pas textuelles (pour les autres représentations).

Une autre nomenclature est définie par le service CAS (« Chemical Abstracts Service ») de l'ACS (« American Chemical Society »). Ce service répertorie les substances publiées dans la littérature dans une base et leur assigne un numéro appelé numéro CAS.<sup>43</sup> Des corps purs comme des mélanges de molécules y sont présents. Cette nomenclature s'affranchit de la barrière linguistique et des noms triviaux, mais rends nécessaire d'avoir accès à la base de données CAS – ou d'un service listant les numéros CAS – pour passer d'une représentation avec numéros CAS à une autre, et vice-versa. En outre, toutes les molécules ne sont pas recensées dans la base (notamment les molécules non présentes dans la littérature) et certaines substances possèdent plusieurs numéros CAS (selon le dossier REACH de l'éthanol, cette molécule possède le numéro CAS 64-17-5, mais possédait également les numéros 121182-78-3, 8000-16-6 et 8024-45-1).<sup>42</sup>

Le langage SMILES (« Simplified Molecular Input Line Entry Specification ») est une nomenclature adaptée pour à la fois représenter les structures chimiques et les cataloguer. Il encode les structures avec une succession de caractères typographiques en explorant leur graphe moléculaire atome par atome. Son utilisation est très répandue du fait de sa compréhension tant par l'utilisateur que par la majorité des logiciels de Chémoinformatique. La conversion d'une chaîne SMILES en structure chimique et inversement ne nécessite pas de faire appel à une base de données. Un graphe moléculaire peut posséder plusieurs SMILES, dépendant du chemin par lequel le graphe est décrit (le prop-2-énal possède six notations SMILES : C=CC=O, O=CC=C, C(=O)C=C, C(C=C)=O, C(C=O)=C et C(=C)C=O). La notation canonique du langage SMILES commence par prioritariser les atomes (selon des règles pouvant varier, nous avons employé celles publiées par Schneider *et coll.*<sup>44</sup> et implémentées dans RDKit<sup>45</sup>) puis explore le graphe moléculaire selon les priorités calculées. De ce fait, un graphe moléculaire ne possède qu'un seul SMILES canonique pour un même jeu de règles de priorisation (le SMILES canonique du prop-2-énal est C=CC=O avec les règles de Schneider *et coll.*<sup>44</sup>). La notation sous forme de SMILES canonique possède cependant des limitations, et peut notamment rencontrer des difficultés à représenter des molécules aromatiques et polyaromatiques.<sup>46,47</sup>

Les représentations moléculaires décrites dans cette partie permettent de cataloguer les molécules, de les reconnaître, et/ou de connaître leur structure chimique. Cependant, une présentation de l'information chimique sous d'autres formes, telles que les descripteurs, est nécessaire pour que les algorithmes puissent distinguer les spécificités structurelles de chaque molécule.

#### 2.1.1.1.2 Descripteurs moléculaires

Un descripteur moléculaire est une représentation continue ou discrète *décrivant* une caractéristique moléculaire. Les structures moléculaires sont souvent représentées en Chémoinformatique par un ensemble de descripteurs, définissant un vecteur de descripteurs. Le Tableau 1 présente une des manières de catégoriser les descripteurs : ils y sont classés en fonction de la dimension chimique utilisée pour les définir.<sup>48</sup>

Dimension D	Niveau de description	Exemples de descripteurs
0D	Formule chimique, propriété	Nombre d'atomes, poids moléculaire
1D	Formule chimique développée	Dénombrement de fragments
2D	Graphe moléculaire	Indices topologiques
3D	Représentation 3D	Surface moléculaire, surface polaire
4D	Conformation des molécules	Energie potentielle des conformations

Tableau 1 : Classification des descripteurs moléculaires en fonction du niveau de description considéré.

Les descripteurs tels que les fragments ou les dénombrements de groupes fonctionnels représentent des sous-ensembles d'une molécule. A IFPEN, les descripteurs FGCD (Functional Group Count Descriptors) et SMF (Substructural Molecular Fragments) sont couramment utilisés.<sup>24</sup> Les SMF définissent plusieurs types de dénombrements substructuraux : fragments ou groupes définis par leurs atomes et/ou leurs liaisons. Le Tableau 2 présente les différents types de descripteurs SMF.

Chaque type de descripteur SMF peut être généré pour une longueur bornée d'atomes. La nomenclature des SMF se décompose telle que  $t_x l_y u_z$ , où x est le type de fragment, y sa longueur minimale et z sa longueur maximale. Par exemple, les descripteurs SMF appelés « t3l2u4 » encodent les fragments de 2 à 4 atomes de type 3 (séquences d'atomes et de liaisons). Pour une molécule d'éthane, le descripteur de type « t3l2u4 » nommé « C-H » dénombre le nombre de fragments éponyme dans la molécule et possède une valeur égale à 6. La suite logicielle In Silico design and Data Analysis (ISIDA) permet entre autres de calculer automatiquement les descripteurs SMF d'un jeu de molécules.<sup>49,50</sup> Les FGCD définissent quant à eux les groupes fonctionnels à l'aide du langage SMARTS. Le langage SMARTS permet de rechercher des motifs structuraux dans des molécules encodées par leurs SMILES.<sup>51</sup> Les motifs structuraux sont définis par l'utilisateur, leur longueur n'est pas limitée, et l'usage de caractères « jokers » est possible pour définir des atomes ou liaisons génériques.

Les graphes moléculaires permettent aussi de définir des matrices : la matrice de connectivité présente les voisins de chaque atome et la matrice de distance<sup>52</sup> renseigne du nombre de liaisons entre les paires d'atomes. De ces matrices découlent des descripteurs appelés Indices

Topologiques (TI)<sup>53-55</sup> tels que les indices de Kier<sup>56</sup>, d'Hosoya<sup>57</sup> ou de Randic<sup>58</sup>. Les TIs appartiennent à la famille des descripteurs 2D.

Type	Description	Exemples <sup>1</sup>
0	Dénombrement d'atomes	C,H,O
1	Séquences d'atomes	CC,CO,COH
2	Séquences de liaisons	-X-, -X=
3	Séquence d'atomes et de liaisons	C-C=O, C-C, C=C
4	Fragments centrés avec séquences d'atomes	C(CC), C(C)
5	Fragments centrés avec séquence de liaisons	X(=X)(-X-X-X)
6	Fragments centrés avec séquence d'atomes et liaisons	C(-C-O)(=C-C-O)
7	Fragments centrés avec séquence d'atomes de taille fixe	C(CC), C(CO)
8	Fragments centrés avec séquence de liaisons de taille fixe	X(=X-X)(-X-X)
9	Fragments centrés avec séquences d'atomes et de liaisons de taille fixe	C(-C-O)(=C-C)
10	Triplets	C2C5O3

<sup>1</sup>Les exemples sont donnés avec la notation SMILES, sauf pour le triplet C2C5O3 qui encode un atome X à une distance topologique de 2 d'un atome de carbone (C2), à une distance de 5 d'un autre atome de carbone (C5) et à une distance de 3 d'un atome d'oxygène (O3). X encode un atome inconnu.

Tableau 2 : Les différents types de descripteurs SMF considérés, avec description et exemple.

Les MCDs (Monotonically Changing Descriptors)<sup>59</sup> regroupent un ensemble de descripteurs variés. Ils possèdent la caractéristique commune de modifier leur valeur de façon monotone lors de l'ajout d'un atome à la structure chimique. Le dénombrement d'atomes et de liaisons, les descripteurs SMF et certains indices topologiques tels que l'indice de Randic<sup>58</sup> font partie des MCDs. Des descripteurs tels que ceux décrivant la surface polaire (Topological Polar Surface Area, TPSA) n'appartiennent pas aux MCDs : la surface polaire peut diminuer avec l'ajout d'un groupe apolaire. La propriété « logP » qui décrit la différence de solubilité d'une molécule entre des phases polaire (eau) et apolaire (octanol) n'est également pas un MCD.

Chaque descripteur possède un ordre de grandeur et une amplitude qui peuvent être différents de ceux des autres descripteurs. À titre d'exemple, l'ajout d'un atome de carbone dans une molécule augmente la valeur du descripteur « poids moléculaire » d'environ 12g.mol<sup>-1</sup> tandis que la valeur du descripteur « nombre d'atomes de carbone » augmente d'une unité. De ce fait, les descripteurs possédant de grandes amplitudes et/ou ordres de grandeur pourraient avoir plus d'importance que les autres au sein des vecteurs de descripteurs en modélisation.<sup>60</sup> Une mise à l'échelle des valeurs de descripteurs est alors effectuée pour éviter de tels problèmes. Nous avons choisi de travailler avec deux méthodes de mise à l'échelle. La normalisation MinMax redéfinit les valeurs de descripteurs de façon que leur amplitude soit égale à 1, en définissant leurs bornes telles que [0 ; 1]. Cela revient à diviser la différence entre la valeur du descripteur et sa valeur minimale par la différence entre sa valeur maximale et minimale (Equation (2)). La standardisation redéfinit les valeurs de descripteurs afin qu'ils possèdent une valeur de déviation standard égale à un. Les valeurs de descripteurs sont redéfinies en divisant la différence entre la valeur du descripteur et sa valeur moyenne par son écart type (Equation (3)).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

$$x' = \frac{x - \bar{x}}{\sigma} \quad (3)$$

Une fois les molécules converties en vecteurs de descripteurs, ces derniers peuvent être utilisés par les méthodes d'apprentissage automatique (MAA) pour établir les QSPR.

### 2.1.1.2 Méthodes d'apprentissage automatique (MAA)

Les MAA servent à construire des modèles pour représenter l'espace chimique et/ou prédire les propriétés des molécules. Quand la propriété n'est pas utilisée pour créer les modèles, l'on parle d'apprentissage non supervisé, au contraire de l'apprentissage supervisé où les modèles sont optimisés pour prédire correctement la propriété durant leur conception. Parmi les modèles prédictifs, les modèles de classification estiment des propriétés qualitatives (l'appartenance à une classe de molécules, par exemple). Les modèles de régression considèrent quant à eux des propriétés quantitatives.

L'apprentissage automatique est en plein essor et il existe actuellement une palette de MAA disponibles.<sup>61</sup> Trois MAA ont été employées durant la thèse et sont présentées ci-dessous : les machines à vecteurs de support, la cartographie topographique générative et les analyses en composantes principales.

#### 2.1.1.2.1 Machines à vecteurs de support (SVM, SVR)

Les machines à vecteurs de support<sup>62</sup> (SVM, pour « Support Vector Machine ») sont des MAA supervisées utilisées pour faire des modèles de classification. La Figure 2 présente une modélisation par SVM. Avec la modélisation par SVM, un hyperplan se charge de séparer les vecteurs appartenant à différentes classes dans l'espace des descripteurs. La distance orthogonale entre l'hyperplan et les vecteurs les plus proches est appelée marge. Les vecteurs présents à la marge sont appelés vecteurs de support. Idéalement, l'hyperplan est optimisé de façon à obtenir la plus grande marge sans effectuer d'erreur de séparation (SVM à marge dure). Cependant, dans la majorité des jeux de données, les instances de chaque classe ne sont pas séparables linéairement et une marge souple est employée. Le coût (C) est alors introduit pour pénaliser chaque vecteur séparé dans la mauvaise classe, proportionnellement à sa distance à la marge. Dans ce cas, l'hyperplan est optimisé de façon à obtenir la plus grande marge tout en réduisant les pénalités dues aux données mal classées. Pour des jeux de données pour lesquels une séparation linéaire n'est pas évidente dans l'espace chimique donné, un espace à plus haute dimension (supérieure au nombre de descripteurs) est utilisé pour projeter les données. Par exemple, avec une fonction radiale à la place d'une fonction linéaire, on parle de SVM à noyau « gaussien ». La fonction radiale fait intervenir un paramètre gamma qui contrôle la variance de la gaussienne utilisée ; plus sa valeur est faible plus la variance est élevée et plus deux points éloignés peuvent être considérés similaires.



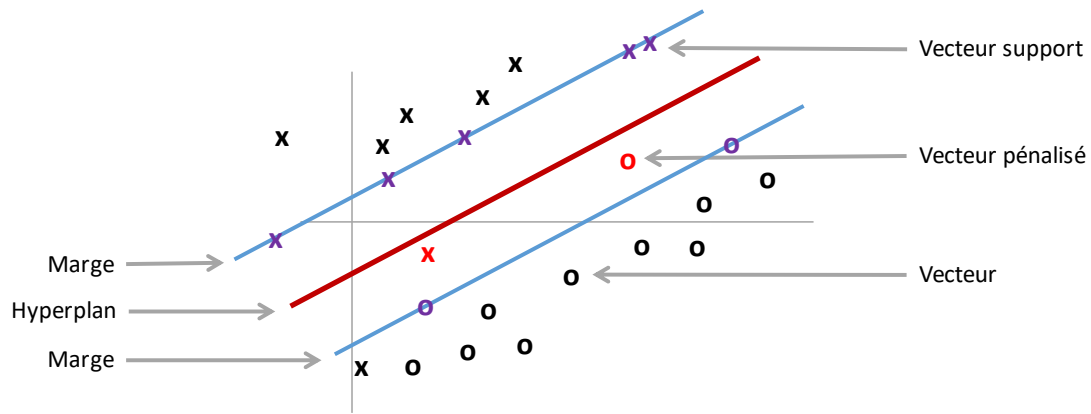


Figure 2 : Fonctionnement d'un modèle de classification par SVM. Les vecteurs de chaque classe sont représentés sous forme de "x" pour la première classe et de "o" pour la seconde.

Les machines à vecteur de support peuvent également être utilisées dans des problèmes de régression et sont alors appelées SVR (pour « Support Vector Regression »).<sup>63</sup> La Figure 3 illustre une modélisation SVR. En SVR, la fonction modélisant l'hyperplan modélise également la propriété. Une zone d'insensibilité est mise en place autour de l'hyperplan pour ne pas pénaliser les instances légèrement mal prédites. Avec la méthode  $\epsilon$ -SVR, la zone d'insensibilité est définie directement par le paramètre epsilon ( $\epsilon$ ).<sup>64</sup> Les vecteurs se trouvant à une distance de l'hyperplan plus élevée sont pénalisés par le coût proportionnellement à cette distance. L'optimisation d'une SVR consiste à placer l'hyperplan de façon à pénaliser un minimum de vecteurs.

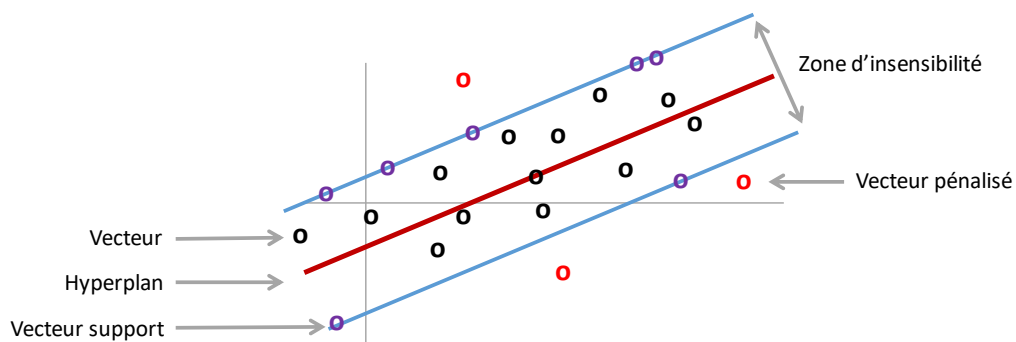


Figure 3 : Fonctionnement d'un modèle de régression par SVR.

#### 2.1.1.2.2 Analyse en Composantes Principales (ACP)

Des techniques de réduction de dimensions existent pour représenter les données de façon plus claire autant pour les algorithmes que pour l'utilisateur.<sup>65</sup> L'Analyse en Composantes Principales (ACP)<sup>66</sup> est une technique de réduction de dimensions. Elle génère des combinaisons non corrélées de descripteurs, appelées composantes principales (PC, pour

« Principal Component »). L'ACP est souvent utilisés pour présenter graphiquement la diversité de données ou pour obtenir de nouveaux descripteurs à utiliser avec une MAA de type prédictif. Selon les informations de la littérature, il n'est pas évident de classer l'ACP comme une méthode statistique<sup>67</sup> ou comme une MAA.<sup>48</sup> Bien que l'ACP ne soit pas un outil de prédiction en soi, nous la considérons comme une MAA car elle utilise l'information chimique – les descripteurs moléculaires – de façon non supervisée pour en proposer une nouvelle représentation.

### 2.1.1.2.3 Cartographie Topographique Générative (GTM)

La Cartographie Topographique Générative (GTM, pour « Generative Topographic Mapping ») est une MAA qui permet de représenter un espace chimique complexe en deux dimensions.<sup>68</sup> Pour ce faire une grille 2D, appelée « *manifold* », est insérée dans l'espace chimique initial multidimensionnel. Chaque nœud de la grille correspond à une fonction gaussienne de base radiale qui permet à la grille de s'agencer dans l'espace chimique multidimensionnel. Une fois l'agencement optimisé, chaque molécule est projetée sur chaque nœud avec une probabilité (appelée *responsabilité*), relative à sa distance à ce nœud. Le *manifold* est finalement déplié et devient la carte GTM. Chaque nœud est ensuite coloré en fonction de sa valeur moyenne de propriété. La valeur moyenne de propriété d'un nœud est définie comme la somme des valeurs de propriété des molécules, pondérées par leurs *responsabilités* dans le nœud. Aussi, l'opacité de la coloration renseigne sur la densité moyenne des nœuds : plus un nœud est représenté avec couleur opaque, plus le nombre de molécules pour lesquelles il possède une haute *responsabilité* est élevé. Les nœuds totalement transparents – en blanc – sont des nœuds non explorés par les données d'apprentissage. La méthode GTM est une méthode de représentation de l'espace chimique qui peut être adaptée pour la prédiction de propriétés : de nouvelles molécules peuvent être projetées sur la carte et leur propriété prédite.

### 2.1.1.3 Qualité des modèles

#### 2.1.1.3.1 Indices de performance

Les indices de performance évaluent la qualité de modèles prédictifs.<sup>69</sup> Parmi les différents indices disponibles, nous avons choisi de travailler avec les plus courants : l'erreur quadratique

et ses dérivées telles que le RMSE (de l'anglais « Root Mean Squared Error ») ou encore le coefficient de détermination ( $R^2$ ).

L'indice RMSE défini dans l'équation (4) est la racine carrée de la moyenne des écarts entre les valeurs prédites et valeurs réelles élevés au carré. Le coefficient de détermination de Pearson  $R^2$ , comme défini par l'équation (5), est un indice dont la valeur est comprise dans l'intervalle [0 ; 1]. Un modèle dit idéal, où la propriété est parfaitement prédite pour toutes les molécules considérées, possède une valeur  $R^2$  égale à 1 et une valeur de RMSE égale à 0.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^{pred} - y_i^{exp})^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i^{pred} - y_i^{exp})^2}{\sum_{i=1}^N (y_i^{exp} - \overline{y^{exp}})^2} \quad (5)$$

Où  $N$  est le nombre de données considérées,  $y_i^{pred}$  la valeur prédite de la propriété  $y$  pour le composé  $i$ ,  $y_i^{exp}$  la valeur expérimentale pour le composé  $i$ ,  $\overline{y^{exp}}$  la moyenne des valeurs expérimentales pour la propriété  $y$ .

#### 2.1.1.3.2 Validation des modèles

La validation est l'étape d'évaluation de la qualité de prédiction des QSPR. Les indices de performance sont utilisés dans ce but par deux types de validation : la validation interne et externe.<sup>70</sup>

La validation interne teste la capacité prédictive d'un modèle lors de sa construction, avec les données d'entraînement. Elle est utilisée pour choisir le jeu de paramètres de la MAA menant au modèle le plus performant. Pour cela, plusieurs jeux de paramètres sont testés pour modéliser la propriété et le jeu de paramètres menant au modèle aux meilleures performances est retenu.

La Validation Croisée (CV, pour « Cross-Validation ») est une des validations internes souvent utilisées : elle permet d'utiliser toutes les données d'entraînement autant pour créer des modèles que pour les valider. Lors d'une CV, les données d'apprentissage sont divisées en plusieurs paquets de taille similaire. Le nombre de paquets utilisé varie de 2 à  $m$ , où  $m$  est le nombre de

molécules initiales (si  $m$  paquets sont utilisés, il y a une molécule par paquet, c'est la validation « un contre tous »). Le choix du nombre de paquets reste une question ouverte, Afendras et Markatou proposent par exemple une série d'équations pour définir ce nombre,<sup>71</sup> tandis que Marcot et Hanea concluent qu'une division en cinq paquets donne des résultats satisfaisants.<sup>72</sup> Ensuite et itérativement, chacun des paquets est mis de côté alors que les données des paquets restants servent à créer un modèle d'entraînement. Ce modèle d'entraînement est appliqué pour prédire la propriété des données du paquet mis de côté et les indices de performance sont calculés.

La validation externe teste la capacité de prédiction d'un modèle déjà construit. Le modèle est appliqué sur des données de test non utilisées lors de sa construction. Les indices de performance sont ensuite calculés. La validation externe reflète les capacités prédictives du modèle en condition d'exploitation.

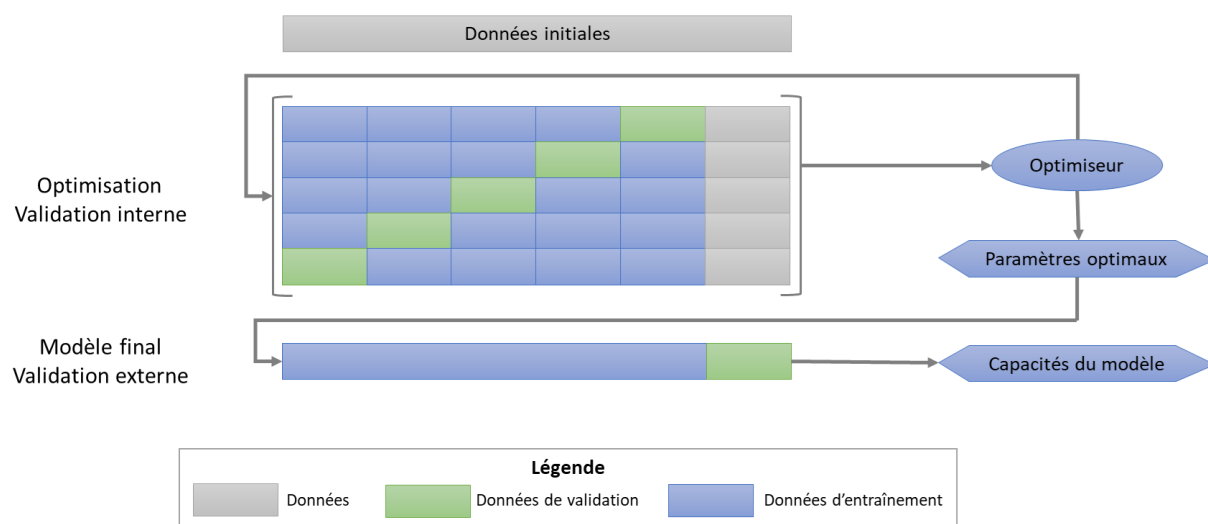


Figure 4 : Modèle de validation croisée proposé pour l'optimisation des paramètres QSPR.

La Figure 4 présente le fonctionnement conjoint des validations interne et externe lors de la création d'un modèle QSPR. Dans cet exemple, les données initiales sont divisées en six paquets. Cinq des six paquets sont utilisés pour optimiser les paramètres de la MAA par validation croisée. Un modèle final est ensuite conçu avec les paramètres optimaux et l'ensemble des données des cinq paquets d'apprentissage. La robustesse du modèle final est testée par validation externe avec les données du sixième paquet. Sur la Figure 4, nous pouvons remarquer qu'un « Optimiseur » se charge de superviser le choix des paramètres lors de la validation interne. Nous présentons plusieurs types d'optimiseurs dans la partie suivante.

#### 2.1.1.4 Optimisation des paramètres des modèles

Plusieurs méthodes d'optimisation existent pour fournir au processus de validation interne les valeurs de paramètres à tester. La méthode d'optimisation de base est l'énumération puis l'itération de la matrice des paramètres avec une certaine granularité. Cette méthode est lente, car chaque combinaison de paramètres est testée.

D'autres méthodes d'optimisation permettent de s'affranchir de tester les jeux de paramètres non pertinents, ce qui permet un gain de temps.<sup>73,74</sup> Ces méthodes d'optimisation plus avancées procurent tour à tour des ensembles de paramètres judicieusement choisis en fonction des performances observées avec les jeux de paramètres précédemment testés. Deux ensembles de méthodes d'optimisation sont présentés ci-dessous.

Les méthodes DFO (Derivative Free Optimization) telles que SQA<sup>75,76</sup> (Sequential Quadratic Approximation) optimisent une fonction sans utiliser de dérivées. SQA est une méthode séquentielle qui utilise des modèles quadratiques pour interpoler la fonction à optimiser. À chaque itération, la fonction est évaluée autour d'un point – jeu de paramètres – et la connaissance de l'influence des paramètres sur la fonction évaluée est consolidée. En peu d'itérations et sans usage de dérivées, un minimum de la fonction peut être atteint. La méthode SQA est implémentée dans l'optimisateur HubOpt interne à IFPEN et sert déjà à l'optimisation de modèles QSPR au laboratoire.

Les Méthodes Métaheuristiques (MM) sont des méthodes itératives et stochastiques qui font évoluer une fonction vers un minimum.<sup>77</sup> Parmi les méthodes MM, les Algorithmes Evolutionnaires (EA, pour « Evolutionary Algorithm ») considèrent l'ensemble des paramètres à optimiser comme un chromosome.<sup>78</sup> Un chromosome est constitué de plusieurs allèles, codant chacun un paramètre à optimiser. Ces chromosomes peuvent évoluer au fil des itérations par des opérations. Les mutations changent aléatoirement la valeur d'un allèle. Les croisements sélectionnent et segmentent aléatoirement deux chromosomes (A et B) et recombinent les segments issus de A avec ceux de B. L'Algorithme Génétique (GA, pour « Genetic Algorithm ») est un EA se basant sur le principe de l'évolution darwinienne, où les chromosomes parents sont remplacés par les chromosomes modifiés si les performances de ces derniers sont meilleures que celles des parents.<sup>79</sup>

### 2.1.1.5 Domaine d'applicabilité

Les prédictions des modèles QSPR ne sont garanties fiables que dans l'espace chimique défini par les molécules initiales, appelé Domaine d'Applicabilité (AD, pour « Applicability Domain »). Il est donc important de s'assurer que les molécules à prédire appartiennent à l'AD. Roy *et coll.* présentent dans leur revue la majorité des techniques existantes pour tester l'appartenance d'une molécule à l'AD.<sup>80</sup> Par exemple, on peut vérifier que la distance entre la molécule candidate et ses plus proches voisins du jeu d'apprentissage dans l'espace des descripteurs ne soit pas plus élevée que la distance moyenne entre les molécules du jeu d'apprentissage. On peut également se servir de variables aléatoires à densité (PDF, pour « Probability Density Function »).<sup>81</sup> Une PDF représente les probabilités pour une fonction d'atteindre des valeurs données. Dans ce cas, on vérifie que chaque valeur de descripteur de la molécule soit couverte par la PDF du descripteur dans le jeu d'apprentissage. On peut aussi diviser l'espace chimique en différentes sections et vérifier que la molécule appartient à une section dans laquelle des molécules du jeu d'apprentissage sont aussi présentes.

À IFPEN, nous avons utilisé une méthode basée sur les intervalles de descripteurs également décrite dans la revue de Roy *et coll.*<sup>80</sup> On vérifie pour cela que la molécule à prédire possède uniquement des descripteurs dont les valeurs sont comprises dans les bornes des descripteurs définies à partir des molécules du jeu d'apprentissage. De même, la présence de nouveaux fragments dans une molécule peut indiquer son éloignement de l'AD.<sup>49,82</sup> En effet, cela indique la présence de nouvelles caractéristiques moléculaires sur les molécules à prédire. L'influence de ces nouvelles caractéristiques sur la propriété modélisée est inconnue du modèle QSPR. Dans nos modélisations, nous avons considéré comme hors de l'AD les molécules possédant des fragments SMF inconnus des molécules du jeu d'apprentissage.

## 2.1.2 Flux opérationnel pour la création des QSPR

### 2.1.2.1 Modèles SVR

L'implémentation de la méthodologie d'optimisation des modèles SVR est représentée sur la Figure 5. Les données initiales sont d'abord transformées en descripteurs. Plusieurs jeux (espaces) de descripteurs SMF sont calculés avec le logiciel Fragmentor<sup>50</sup> (déjà utilisé

précédemment à IFPEN) ou directement procurés par l'utilisateur. Un script codé en Python se charge ensuite de diviser de la même manière les données de chaque espace de descripteurs en six paquets (cinq paquets pour une validation croisée, un paquet de test externe). Il parallélise ensuite l'optimisation d'un modèle SVR pour chaque espace de descripteur. Chacune de ces optimisations est contrôlée par un script codé en Bash, normalisant d'abord les valeurs descripteurs avec la méthode MinMax, puis utilisant libsvm<sup>83</sup> pour la création des modèles SVR, et conjointement la méthode SQA implémentée dans HubOpt pour optimiser les valeurs de paramètres. Une fois toutes les optimisations terminées, le script codé en Python choisit parmi tous les modèles optimisés celui obtenant les meilleures performances comme modèle final.

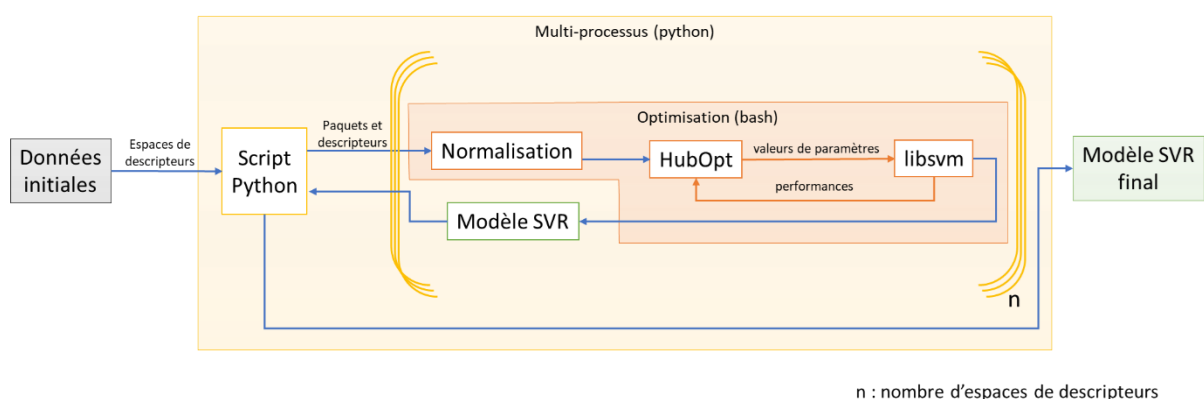


Figure 5 : Méthodologie employée pour la création et l'optimisation des SVR.

### 2.1.2.2 Représentation des données par ACP

La Figure 6 résume les étapes de génération des ACP : normalisation des valeurs de descripteurs selon la méthode MinMax, génération des combinaisons linéaires de descripteurs (sous forme de PC) réduisant la dimension de l'espace chimique, et représentation graphique de l'espace formé par les trois premières PC. Cette méthode a été implémentée dans un script codé en Python, utilisant la librairie scikit-learn<sup>84</sup> pour normaliser les données et effectuer l'ACP, la librairie matplotlib<sup>85</sup> pour représenter l'espace formé, la librairie SciPy<sup>86</sup> pour dessiner une enveloppe convexe autour des données.

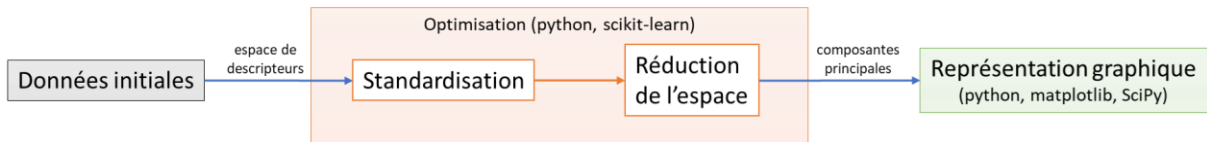


Figure 6 : Méthodologie employée pour la création et l’optimisation des ACP.

### 2.1.2.3 Représentation des données et modèles GTM

La Figure 7 présente la méthodologie que nous avons employée pour la création et l’optimisation des cartes GTM. Les espaces de descripteurs sont calculés avec le logiciel Fragmentor (ISIDA) ou directement procurés par l’utilisateur. Les espaces de descripteurs bruts et normalisés (par la méthode MinMax) sont testés. La création des cartes GTM a été effectuée avec le logiciel GTMapTool.<sup>87</sup> Nous avons utilisé le progiciel libsvm-GAconfig pour trouver les meilleures combinaisons de paramètres pour la carte GTM, c’est-à-dire choisir le jeu de descripteurs et les valeurs d’hyperparamètres (le nombre de nœuds dans la carte, le nombre de fonctions radiales, la largeur des fonctions radiales).<sup>88</sup> Ce progiciel utilise un algorithme génétique : chaque paramètre est encodé dans un allèle, l’ensemble des paramètres forme un chromosome, les chromosomes sont modifiés à l’aide d’opérations génétiques pour trouver les meilleures combinaisons de paramètres.

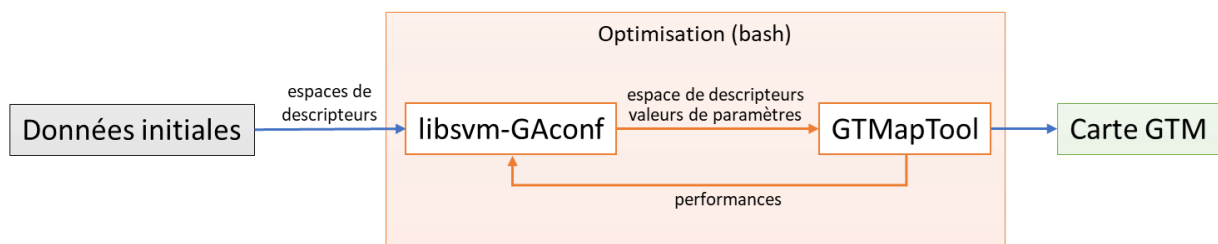


Figure 7 : Méthodologie employée pour la création et l’optimisation des cartes GTM.

## 2.2 Modèles construits au cours de la thèse

Nous décrivons dans cette partie les résultats de modélisation obtenus durant la thèse et comparons notre méthode d’optimisation à celle précédemment utilisée à IFPEN.



## 2.2.1 Jeu de données

À la suite des précédents travaux de modélisation QSPR à IFPEN, plusieurs jeux de données sont déjà disponibles au laboratoire. Parmi ces derniers, nous avons sélectionné celui sur la propriété de point d'éclair (PE). Le PE définit la température la plus basse à laquelle les vapeurs d'un composé peuvent s'enflammer en présence d'une source d'énergie calorifique.<sup>89</sup> Cette propriété est notamment utilisée pour définir l'inflammabilité des composés.<sup>90</sup> De ce fait, et contrairement à la majorité des jeux de données disponibles à IFPEN, l'usage du jeu de données sur le PE n'est pas limité au domaine de l'énergie. Ce jeu de données contient 814 molécules avec pour chacune la valeur expérimentale de point d'éclair exprimée en kelvin (K) et est issu des travaux de Saldana *et coll.*<sup>91</sup> Il a été construit à partir de données disponibles dans la littérature, notamment répertoriées dans la base DIPPR.<sup>92</sup> Les valeurs expérimentales de point d'éclair ont été obtenues par des tests dits « en coupelle fermée », jugés plus précis et reproductibles que les tests en « coupelle ouverte ».<sup>93</sup> Les incertitudes de mesure « coupelle fermée » reportées dans la littérature sont de l'ordre de 5 à 8 degrés.<sup>89</sup> Saldana *et coll.* ont déjà créé un modèle prédictif avec uniquement 625 molécules de cette base ; se limitant aux molécules pouvant être présentes dans les carburants.<sup>91</sup>

Dans ce travail de thèse, nous sommes partis de l'intégralité de la base. Les descripteurs SMF que nous avons employés, de type fragments, ne prennent pas en compte la stéréo-isomérie.<sup>49</sup> Nous avons donc enlevé des structures moléculaires les informations sur la stéréo-isomérie et supprimé les doublons. La base épurée contient alors 785 structures uniques possédant une valeur de PE comprise entre 135 et 533 K. Les molécules de cette base sont présentées avec leur notation SMILES en Annexe A. Cette base de données a été aléatoirement divisée en deux ensembles : d'apprentissage (599 molécules) et de test (186 molécules). La Figure 8 présente la distribution des valeurs de PE des molécules de la base entière, de celles de l'ensemble d'entraînement et de celles de l'ensemble de test. Pour chaque jeu, la majorité des molécules possèdent une valeur de PE comprise entre 200 et 500 K. L'ensemble d'entraînement possède un ratio légèrement plus important de molécules avec une valeur de PE comprise entre 250 et 340 K que l'ensemble de test. La Figure 9 présente quant à elle la diversité chimique des structures de la base. Environ 60% des molécules de la base sont des composés oxygénés et 40% des hydrocarbures. La répartition des molécules en fonction de leur famille chimique est similaire entre le jeu d'apprentissage et de test. Les molécules cycliques, dont font partie les naphthènes et les aromatiques, représentent près de 30% du nombre total de molécules. Notons

la présence importante des esters, qui représentent 21% de la base, à l'inverse des alcynes (seulement huit molécules) et des peroxydes (deux molécules).

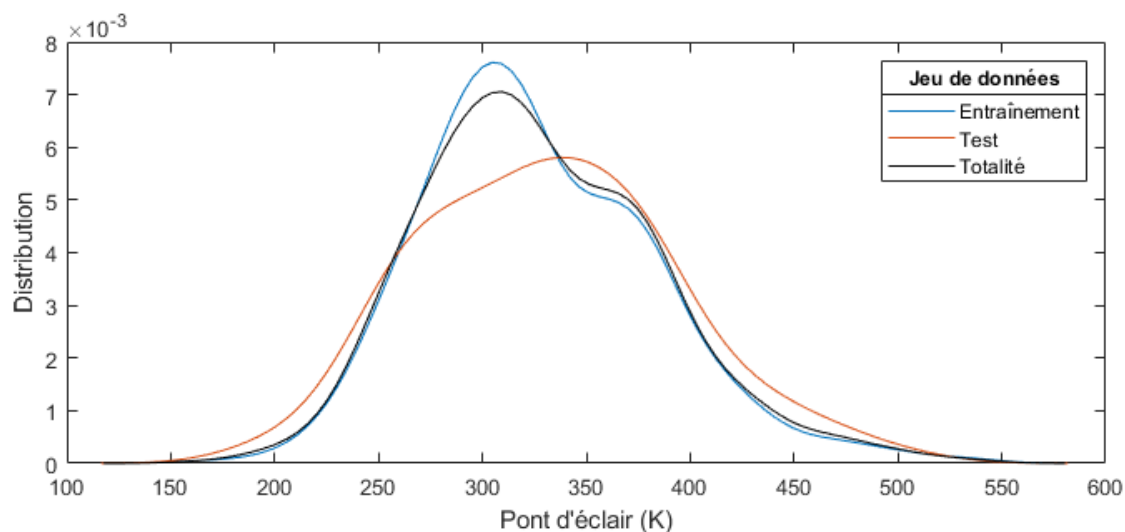


Figure 8 : Distributions des valeurs de point d'éclair dans le jeu de données considéré pour le travail de thèse, pour la totalité du jeu, le jeu d'entraînement, et le jeu de test.

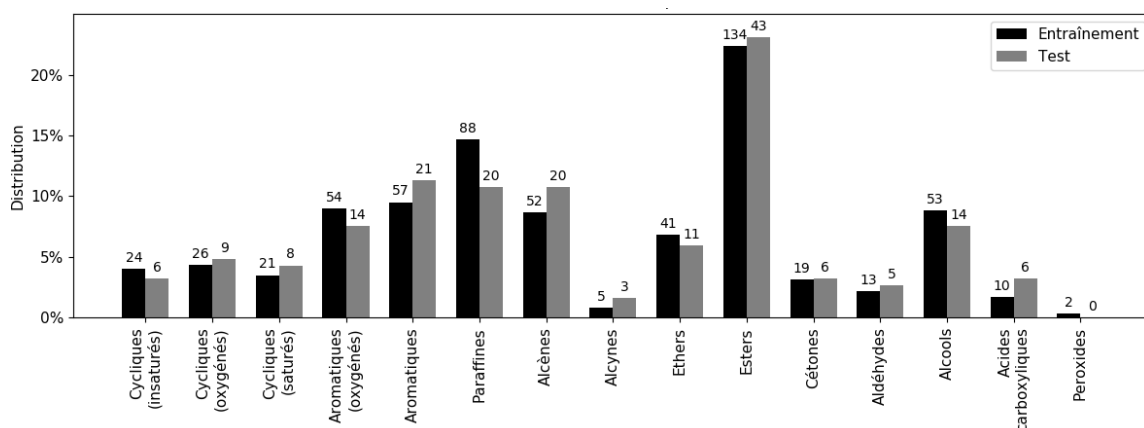


Figure 9 : Distribution des molécules du jeu de données (point d'éclair) en fonction de leur famille chimique. Adapté avec permission, issu de Gantzer *et coll.*<sup>94</sup> Copyright 2021 American Chemical Society.

## 2.2.2 Modélisation par SVR

Cette section est divisée en deux sous-sections. Tout d'abord, nous vérifions que la méthode de création des SVR est efficace. Ensuite, nous l'appliquons pour modéliser la propriété de point d'éclair avec le jeu de données présenté ci-dessus.

### 2.2.2.1 Validation de la méthode d'optimisation des SVR

Pour valider notre approche d'optimisation des modèles SVR décrite dans la partie 2.1.2.1 « Modèles SVR », nous avons comparé les performances de modèles construits suivant cette approche avec celles des modèles publiés par Saldana *et coll.*<sup>91</sup> Dans leur travail, les auteurs ont optimisé des modèles SVR pour prédire le point d'éclair et l'indice de cétane d'hydrocarbures avec des descripteurs FGCD. Pour permettre une comparaison rigoureuse des modèles, nous avons utilisé les mêmes molécules, valeurs de propriétés, descripteurs et répartitions des données entre jeu d'apprentissage et de test que ceux de Saldana *et coll.*<sup>91</sup> avec notre méthode. Nos optimisations ont été répétées 10 fois pour homogénéiser les résultats, dépendants de la répartition des molécules dans chaque paquet. Les résultats sont décrits dans le Tableau 3 pour l'indice de cétane et dans le Tableau 4 pour le point d'éclair. Les espaces de descripteurs SMF ont aussi été testés avec les mêmes molécules que celles utilisées dans les travaux de Saldana *et coll.*<sup>91</sup>, pour modéliser les deux propriétés. Les performances du meilleur modèle construit avec un espace de descripteurs SMF sont également présentées dans le Tableau 3 et dans le Tableau 4.

Indice de cétane		
Descripteurs	RMSE (K)	
	Apprentissage	Test
FGCD (Saldana et coll.) <sup>91</sup>		12,4
FGCD (ce travail)	10,5 ± 0,2	12,6 ± 1,2
SMF t1011u4 (ce travail)	8,8	12,2

Tableau 3 : Performances des modèles (RMSE) pour prédire l'indice de cétane.

<b>Point d'éclair</b>		
<i>Descripteurs</i>	<i>RMSE (K)</i>	
	<i>Apprentissage</i>	<i>Test</i>
<i>FGCD (Saldana et coll.)<sup>91</sup></i>		11,6
<i>FGCD (ce travail)</i>	15,0 ± 0,6	10,8 ± 0,8
<i>SMF t1l2u4 (ce travail)</i>	8,8	12,1

Tableau 4 : Performances des modèles (RMSE) pour prédire le point d'éclair.

Nos modèles créés à partir des descripteurs FGCD obtiennent des performances de prédiction similaires à celles des modèles publiés par Saldana *et coll.*<sup>91</sup> Les modèles utilisant les descripteurs SMF obtiennent des performances similaires à celles des modèles utilisant les descripteurs FGCD. Nous en déduisons que la méthodologie d'optimisation des SVR utilisée pour la thèse permet d'obtenir des modèles de qualité équivalente à celle des modèles obtenus avec la méthodologie d'optimisation précédemment utilisée à IFPEN.

### 2.2.2.2 Modélisation par SVR de la propriété de point d'éclair

Une fois validée, la méthode d'optimisation des SVR a été appliquée sur notre jeu de molécules présenté dans la partie 2.2.1 « Jeu de données » concernant le point d'éclair. Un des modèles les mieux classés a été obtenu avec les descripteurs SMF t3l2u4 (fragments de deux à quatre atomes avec leurs liaisons). Les valeurs des paramètres optimisés de ce modèle sont présentées dans le Tableau 5.

Paramètre	Valeur
Coût	1868,48
Gamma	0,16
Epsilon	6,98

Tableau 5 : Valeurs des hyperparamètres optimisés du modèle utilisant les descripteurs SMF t3l2u4.

L'ensemble des valeurs de PE prédites, par le modèle obtenu avec les descripteurs SMF t3l2u3, pour les molécules de notre base, est présenté dans l'Annexe A. La comparaison entre les valeurs de PE expérimentales et prédites est présentée graphiquement sur la Figure 10, pour les molécules des ensembles d'entraînement et de test. La très grande majorité des molécules possède une bonne corrélation entre valeur de PE prédite et expérimentale.

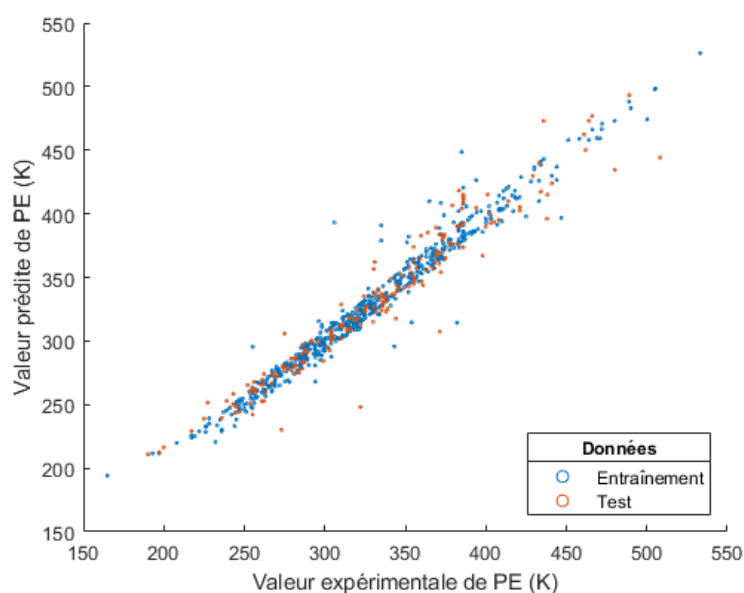


Figure 10 : Comparaison entre les valeurs de point d'éclair, expérimentales et prédites par le modèle obtenu avec les descripteurs SMF t3l2u3, pour les molécules des ensembles d'entraînement et de test.

Les performances de notre modèle sont présentées dans le Tableau 6, et comparées à celles du modèle dit « consensus » de Saldana *et coll.*<sup>91</sup> – utilisant un ensemble de plusieurs modèles individuels avec différentes MAA pour modéliser le point d’éclair. Notre modèle a obtenu des valeurs de RMSE en validation comprises entre 15,5 et 15,7 K ainsi que des valeurs de R<sup>2</sup> supérieures à 0,920. Ce modèle obtient des performances comparables à celles du modèle « consensus » de Saldana *et coll.*<sup>91</sup>

Jeu de données	Performances de notre modèle		Performances du modèle consensus de Saldana <i>et coll.</i> <sup>91</sup>	
	RMSE (K)	R <sup>2</sup>	RMSE (K)	R <sup>2</sup>
Jeu d’apprentissage (validation croisée)	15,7	0,920	---	---
Jeu de test (validation externe)	15,5	0,935	---	---
Jeu complet de Saldana <i>et coll.</i> <sup>91</sup>	12,7	0,948	10,9	0,959
Jeu de test de Saldana <i>et coll.</i>	9,6	0,967	13,2	0,944

Tableau 6 : Performances du modèle QSPR utilisé pour la suite de la thèse et du modèle de Saldana *et coll.*, sur plusieurs jeux de molécules.

### 2.2.3 Représentation des données par ACP

Les descripteurs SMF t3l2u4 ayant permis d’obtenir le meilleur modèle SVR, ils ont été utilisés pour encoder les molécules à représenter par ACP, selon la méthode présentée dans la partie 2.1.2.2 « Représentation des données par ACP ». Les trois premières composantes principales (PC, pour « Principal Component ») obtenues permettent d’expliquer seulement 37% de la variance du jeu de données, ce qui est relativement faible. Ces 3 PC ont été utilisées pour créer un espace 3D illustré sur la Figure 11, qui représente partiellement l’espace chimique du jeu initial, et dans lequel nous avons projeté les molécules initiales.

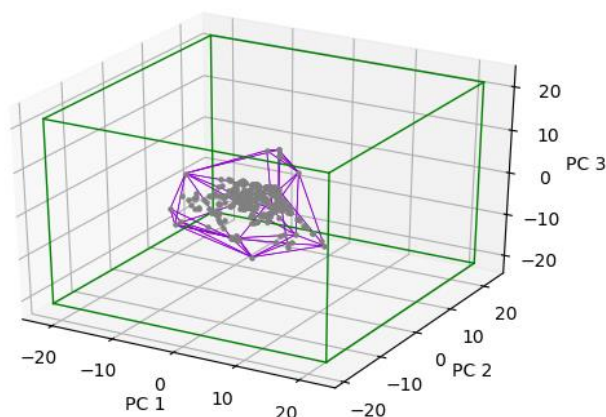


Figure 11 : Représentation de l'espace  $\mathbb{C}$  (en vert), issu de la PCA du jeu de données concernant le point d'éclair. Les molécules initiales  $y$  sont projetées en gris et sont englobées par une enveloppe convexe représentée en violet. Réimprimé avec permission, issu de Gantzer *et coll.*<sup>94</sup> Copyright 2021 American Chemical Society.

Sur cette figure, nous avons dessiné une enveloppe convexe autour des molécules initiales pour approximer graphiquement l'espace chimique couvert par nos modèles. Cependant, parce que la représentation par ACP est une approximation et ne représente pas toute l'information chimique de notre base, et que nous souhaitons générer des molécules possédant des combinaisons de valeurs de descripteurs qui ne sont peut-être pas représentées dans la base, certaines molécules générées pourraient être projetées en dehors de cette enveloppe. Nous avons alors défini un espace étendu nommé  $\mathbb{C}$ .  $\mathbb{C}$  est représenté par un parallélépipède rectangle dont les frontières sont définies comme les valeurs frontières de l'enveloppe convexe sur chaque PC, élargies de manière à y permettre la projection de toute nouvelle molécule appartenant à l'AD. Le Tableau 7 présente les valeurs limites de l'espace initial (représenté par l'enveloppe convexe) et étendu (représenté par  $\mathbb{C}$ ).

Axe	Espace initial		Espace étendu	
	Minimum	Maximum	Minimum	Maximum
<i>PC 1</i>	-11,8	12,4	-19,8	20,4
<i>PC 2</i>	-9,2	13,0	-19,2	23,0
<i>PC 3</i>	-11,2	11,7	-19,2	19,7

Tableau 7 : Valeurs limites des espaces initial et étendu sur chaque axe, dans l'espace formé par les trois premières PC de la PCA du jeu de données concernant le point d'éclair.

## 2.2.4 Représentation des données et modélisation par GTM

Des cartes GTM ont été optimisées pour modéliser notre jeu de données sur le point d'éclair, encodé par les descripteurs ISIDA. Les meilleures cartes GTM issues de cette optimisation ont été manuellement analysées pour sélectionner celle permettant la meilleure description du jeu de données initiales. La carte GTM utilisant les descripteurs t3l2u2 (fragments de deux atomes et leur liaison) a obtenu une des meilleures performances (valeur  $R^2$  en validation croisée de 0,81) et présente visuellement un bon compromis : une densité moyenne par nœud élevée – indiquant un bon agencement du *manifold* – et une gamme de valeurs moyennes de point d'éclair par nœud s'étendant de 225 à 505 K – indiquant une bonne séparation des molécules selon leur valeur de propriété. Les valeurs optimisées des paramètres de cette carte sont présentées dans le Tableau 8.

Paramètre	Valeur
# Centres RBF	5
# Nœuds dans l'espace latent	21
Coefficient de régulation	3,31
Largeur de chaque fonction RBF	0,7

Tableau 8 : Valeurs optimisées des paramètres de la carte GTM utilisant les descripteurs t3l2u2. # signifie « nombre ».



La Figure 12 présente la carte associée aux descripteurs t312u2 sur laquelle les molécules de la base de données ont été projetées et sont représentées en fonction de leur famille chimique : les hydrocarbures y sont représentés par des croix et les composés contenant un ou des atomes d'oxygène par des cercles. La couleur des symboles représente la sous-famille chimique des molécules projetées. Cette carte GTM discrimine correctement les molécules en fonction de leur famille et sous-famille chimique, et de plus, discrimine également les molécules en fonction de leur valeur de propriété.

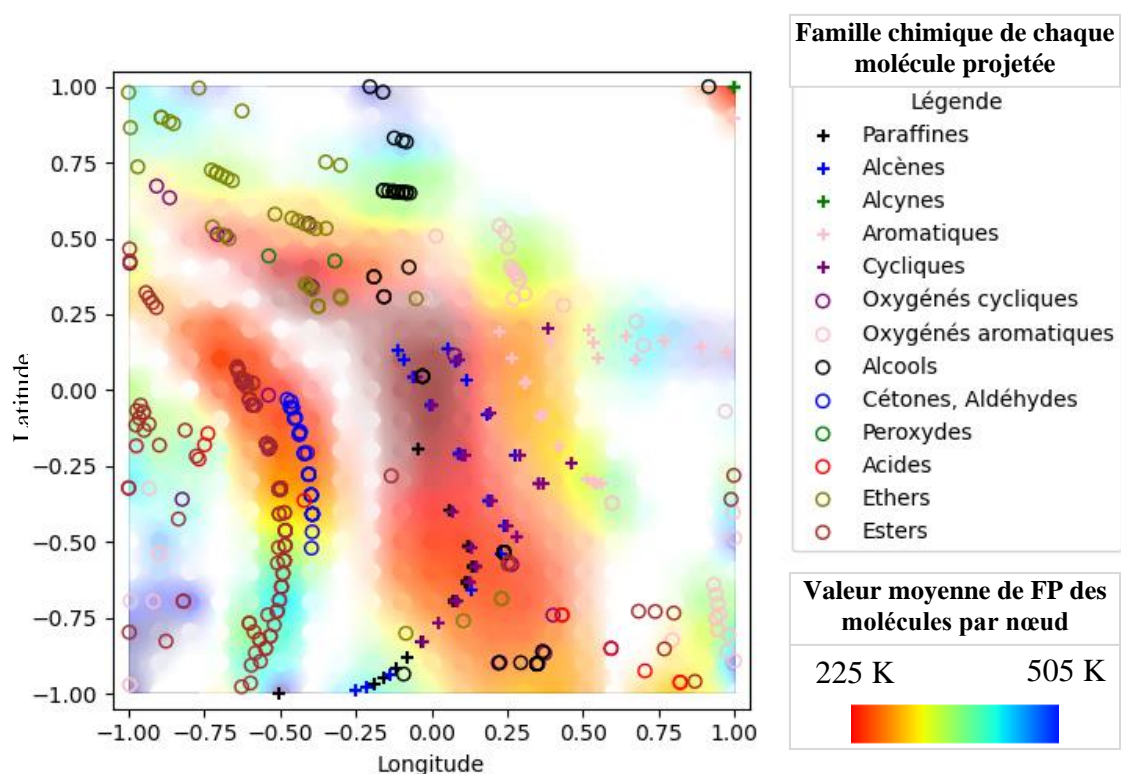


Figure 12 : Carte GTM construite à partir des descripteurs t212u2 pour la base de point d'éclair.

Nous avons ensuite utilisé cette carte comme outil de prédiction du point d'éclair. Les performances du modèle sont présentées dans le Tableau 9. Le modèle obtient des performances plus basses que celles du modèle SVR. Cette différence de performances pourrait s'expliquer par le fait que la carte GTM utilise moins de descripteurs, correspondant à des fragments plus petits (de deux atomes), que le modèle SVR (qui utilise des fragments de deux à quatre atomes). L'usage de fragments de deux atomes permettrait d'obtenir une carte GTM généralisant bien notre jeu de donnée, mais modélisant moins bien la propriété.

Jeu de données	RMSE (K)	R <sup>2</sup>
Jeu d'apprentissage	22,8	0,834
Jeu de test (validation externe)	29,0	0,774

Tableau 9 : Performances du modèle GTM pour la prédiction du point d'éclair.

## 2.3 Conclusions sur la partie modélisation QSPR

Dans ce chapitre, après avoir fait un rappel sur les méthodes QSPR utiles au travail de thèse, nous avons présenté le flux opérationnel utilisé pour obtenir des modèles QSPR et des représentations graphiques des jeux de molécules dans ce travail de thèse.

Cette méthodologie a été appliquée au jeu de données du point d'éclair (PE). Le modèle SVR construit avec les descripteurs ISIDA de type t3l2u4, regroupant des fragments de deux à quatre atomes, s'est avéré le plus performant parmi les modèles construits : il a obtenu un RMSE de 15,5 K et une valeur R<sup>2</sup> de 0,935 en validation externe. Les descripteurs utilisés par ce modèle ont ensuite été utilisés dans une ACP. Les trois premières composantes principales de cette ACP ont été utilisées pour approximer l'espace chimique étendu du jeu de données ; cet espace 3D est nommé  $\mathbb{C}$ . Parallèlement, une carte GTM a été construite avec les descripteurs t3l2u2 (fragments de deux atomes). Cette carte a été utilisée pour représenter les données et prédire la valeur de PE. Elle a obtenu une valeur RMSE de 22,8 K et une valeur R<sup>2</sup> égale à 0,774 en validation externe.

L'espace  $\mathbb{C}$  et la carte GTM représentent graphiquement l'espace chimique défini par notre jeu de données. Ces méthodes sont complémentaires.  $\mathbb{C}$  est un espace en 3D construit par ACP, une méthode non supervisée.  $\mathbb{C}$  est une approximation de l'espace chimique initial. Les molécules sont majoritairement projetées au milieu de cet espace. La carte GTM produit un espace 2D où les molécules sont réparties plus uniformément, et où des corrélations entre zones de la carte GTM et valeurs de PE sont observées.

Le modèle SVR et la carte GTM modélisent la propriété de PE. Les performances du modèle SVR sont supérieures à celles de la carte GTM, et comparables à celle du modèle précédemment publié par Saldana *et coll.*<sup>91</sup> Comparé à celui-ci, notre modèle utilise davantage de données, plus diverses, permettant de couvrir un plus grand espace chimique et augmenter le domaine d'applicabilité (AD). La valeur de RMSE de notre modèle SVR étant proche de l'erreur

expérimentale généralement observée,<sup>89</sup> il peut être utilisé comme alternative à des mesures au laboratoire.

Dans la suite de la thèse, les modèles SVR et GTM sont inversés pour contraindre les générations de molécules. Le modèle SVR est utilisé pour prédire la propriété de molécules générées. La représentation par ACP  $\mathbb{C}$  est exploitée comme outil pour représenter l'espace chimique des molécules générées et observer leur diversité.

## Chapitre 3. Méthodes pour la génération moléculaire et pour leur comparaison

---

### 3.1 Introduction sur les méthodes de génération et leur comparaison

Dans la littérature, des méthodes de génération virtuelle de molécules ont été mises en place afin d'identifier des structures pertinentes, voire de compléter des bases de données existantes. Comme discuté dans l'introduction et dans le chapitre précédent, les modèles QSPR peuvent être utilisés pour prédire les valeurs de propriétés des molécules générées. On notera que ces modèles peuvent aussi être utilisés pour identifier les motifs structuraux pertinents dans la valeur prédite d'une propriété. Une autre voie d'utilisation des modèles QSPR réside dans leur inversion : l'i-QSPR, pour identifier des structures moléculaires respectant des contraintes spécifiques en termes de propriétés.

Au cours de mon travail de thèse, j'ai écrit une revue bibliographique dont le but est de recenser les avantages et inconvénients des méthodes de génération existantes et d'en discuter, préambule nécessaire au travail de cette thèse. Cet article a été publié dans le journal « Molecular Informatics »<sup>34</sup> et est présenté dans l'0 de ce document. Contrairement aux autres revues disponibles dans la littérature sur ce sujet, nous avons également souhaité donner au lecteur des informations sur les principes de la Chémoinformatique associés à la génération de molécules. J'expose ainsi dans la première partie de la revue : la représentation des molécules en Chémoinformatique, le stockage de l'information chimique dans les bases de données, et la modélisation QSPR. La génération virtuelle de molécules est ensuite abordée. Trois ensembles de méthodes ont pu être identifiés et sont présentés. Les premières méthodes remontent aux années 80 et génèrent des molécules sans contrainte sur la propriété.<sup>95</sup> Ce type de construction, menant à la génération d'un grand nombre de molécules inutiles, a ensuite été amélioré pour guider la génération vers la production de molécules possédant des valeurs de propriétés désirées. Ce second type de méthodes utilise notamment des contraintes mathématiques sur les occurrences des fragments et sur les valeurs prises par les descripteurs, emploie des descripteurs particuliers, ou modifie les structures avec des méthodes s'inspirant des algorithmes génétiques.

Les méthodes du troisième type, regroupées sous le terme générique « Deep Learning », utilisent quant à elles des réseaux de neurones pour générer des molécules à partir de leur représentation sous forme de SMILES ou de graphes.

Dans ce chapitre du manuscrit de thèse, nous commençons par récapituler l'information donnée dans notre revue bibliographique sur les méthodes de génération. Nous illustrons chaque ensemble de méthodes avec des travaux publiés dans la littérature. L'information donnée par la revue est enrichie avec l'analyse d'autres travaux publiés et plus récents que la soumission de la revue. De plus, l'émergence des méthodes de génération a nécessité la mise en place d'outils pour permettre leur comparaison. C'est pourquoi, nous complétons notre étude de la littérature avec la présentation et l'analyse des outils de comparaison des méthodes de génération. Finalement, nous dressons les conclusions qui ont permis d'orienter le travail de recherche durant la thèse.

## 3.2 Génération non guidée de molécules

Les premières méthodes génèrent des molécules sans contrainte liée à leurs propriétés et remontent aux années 80.<sup>95</sup> Il s'agit par exemple d'énumérer les combinaisons d'atomes, de liaisons, ou de fragments, pouvant mener à des molécules réalistes. Gani *et coll.*, dans leur méthode, ont tout d'abord sélectionné des fragments pertinents parmi ceux de la base UNIFAC<sup>96</sup> puis les ont assemblés en suivant des règles définies en fonction de la complexité moléculaire souhaitée.<sup>95</sup> Plusieurs définitions de ces règles existent en fonction du problème posé.<sup>95,97,98</sup> Ces règles régissent par exemple : le nombre et la localisation des atomes pouvant être utilisés dans une liaison inter-fragments, et les types de liaisons possibles. Nilakantan *et coll.* ont travaillé avec une base de données privée de 200 000 molécules, à partir desquelles ils ont directement extrait des fragments assortis de leur probabilité d'occurrence.<sup>99</sup> Ils ont ensuite généré des molécules en assemblant successivement ces fragments, les sélectionnant selon leur probabilité d'occurrence dans la base de données. Les points d'accroche libres – préalablement définis – des entités en cours de génération sont utilisés pour y greffer les nouveaux fragments. Les étapes d'assemblage se répètent jusqu'à l'obtention d'une molécule possédant un poids moléculaire dans un intervalle donné. On notera que les fragments étant définis en amont, l'espace de génération l'est également. Cette restriction de l'espace de génération permet d'éviter de générer des molécules totalement différentes de celles initialement connues.

Peironcely et coll. ont développé quant à eux un outil qui permet de générer exhaustivement toutes les structures avec une formule brute donnée, en partant des atomes donnés dans la formule brute et en ajoutant des liaisons entre ces derniers de manière à obtenir une structure unique.<sup>100</sup>

Alternativement, Blum *et coll.* ont énuméré tous les graphes possédant un nombre maximal de nœuds puis ont substitué les nœuds par des atomes et les arêtes par des liaisons chimiques.<sup>101,102</sup> Des filtres ont permis la suppression des structures moléculaires incorrectes, telles que celles ne respectant pas les critères de valence. La base GDB-13 regroupe ainsi des molécules comportant jusqu'à 13 atomes non-hydrogène.<sup>101</sup> Enfin, on peut également s'inspirer des mécanismes réactionnels pour générer automatiquement de nouveaux produits à partir d'une liste de réactifs et/ou de mécanismes réactionnels.<sup>103-105</sup>

Les molécules générées sont ensuite testées, par exemple par QSPR, pour extraire celles qui possèdent les propriétés désirées. Les méthodes de génération libre permettent certes de produire facilement une grande quantité de structures, mais la génération est aléatoire, ces techniques peuvent ne pas trouver de molécule comme solution au problème posé. La restriction sur les choix des fragments, mentionnée précédemment, permet de limiter dans un premier temps le nombre de molécules pouvant être générées. Des améliorations plus poussées se sont avérées nécessaires pour diriger les générations vers des valeurs de propriété souhaitées.

### 3.3 Génération guidée des molécules

Nous présentons dans cette partie les contraintes qui peuvent être utilisées pour diminuer le nombre de molécules générées inutilement. Ces contraintes sont basées sur l'inversion des modèles QSPR (i-QSPR, pour « inverse-QSPR »). L'i-QSPR est principalement assurée (i) en construisant itérativement les molécules qui sont testées par QSPR après chaque modification et/ou (ii) en utilisant des valeurs de descripteurs issues des QSPR pour contraindre la génération.

Dans le premier cas (i), les structures sont modifiées itérativement et seules les molécules s'approchant au plus des valeurs de propriété souhaitées sont conservées entre les différentes

itérations. Les Algorithmes Evolutionnaires (EA, pour « Evolutionary Algorithm ») ont été adaptés pour de telles générations, considérant chaque molécule comme un objet sur lequel des opérations génétiques sont réalisables. Parmi les EA, les algorithmes génétiques ont été adaptés pour la génération de structures moléculaires possédant un nombre fixe et identique de sous-structures ou de fragments. Les structures y sont décrites par des chromosomes, encodant chaque sous-structure ou fragment moléculaire par un caractère. Sheridan et Kearsley<sup>106</sup> se sont intéressés à la génération de tripeptides, molécules composées de 3 peptides contenant chacun 2 fragments. Un chromosome de 6 caractères encode chaque structure grâce à un alphabet qui utilise une lettre pour désigner chaque type de fragment. Deux méthodes d'évolution ont été employées : la méthode élitiste et stochastique. La méthode élitiste sélectionne le meilleur tiers de la population. Les chromosomes de ce tiers sont alors triplés tandis que les autres chromosomes sont supprimés. La première copie reste inchangée, sur la seconde copie un fragment est muté et la troisième copie subit un croisement avec un autre chromosome. La méthode stochastique garde tous les chromosomes et opère aléatoirement une mutation sur le tiers d'entre eux, des croisements avec le second tiers et conserve les chromosomes du dernier tiers. Plus récemment, de tels algorithmes génétiques ont été employés pour proposer, par exemple : de nouveaux polymères respectant une température de transition vitreuse et des contraintes environnementales définies,<sup>107</sup> des nucléotides pour des applications en biologie,<sup>108</sup> ou encore des oligomères pour une utilisation dans les cellules photovoltaïques.<sup>109</sup> L'application des méthodes utilisant des algorithmes génétiques est limitée aux jeux de données pour lesquels les structures sont décomposables en un nombre défini de sous-structures, par conséquent aux problèmes pour lesquels le nombre de molécules réalisables est connu.

D'autres méthodes basées sur les EA permettent de s'affranchir de ce cadre. Elles modifient directement les constituants des molécules – atomes et liaisons – à partir de leur graphe. Nachbar a utilisé la programmation génétique pour transformer les graphes moléculaires en structures arborescentes.<sup>110</sup> Chaque nœud (représentant un atome) et arête (représentant une liaison) d'un graphe moléculaire peut être indépendamment modifié par une opération. La représentation sous forme d'arborescence n'est cependant pas idéale pour représenter les cycles au sein des molécules : il faut « lier » deux arêtes existantes pour fermer un cycle. Globus *et coll.* ont défini les Graphes Génétiques (GG) comme la méthode génétique faisant évoluer les graphes moléculaires.<sup>111</sup> En GG, les graphes moléculaires sont directement manipulés. Dans leur papier, les auteurs ont prouvé l'efficacité des GG pour obtenir sept molécules cibles (notamment butane et morphine) en réalisant des croisements successifs sur un ensemble de

molécules initiales. Dans d'autres travaux, le terme général d'évolution génétique est retenu pour définir la méthode de modification des graphes par les opérations génétiques.<sup>112,113</sup>

Dans le second cas (ii), les valeurs de descripteurs des nouvelles molécules sont imposées préalablement à leur construction. Ces valeurs sont par exemple issues de systèmes d'(in)équations établis à partir des QSPR. Les indices topologiques (TI) ont été les premiers descripteurs utilisés pour cette approche.<sup>114-116</sup> En effet, des modèles QSPR sont basés sur des corrélations entre les valeurs des TIs et les valeurs de propriétés comme le point d'ébullition ou le coefficient de partage n-octanol/eau. Baskin *et coll.* ont résolu (dit inversé) de telles équations QSPR pour obtenir les valeurs de TIs des molécules répondant à une valeur de propriété.<sup>116</sup> À partir des valeurs de TIs, les différentes distributions possibles des arêtes dans un graphe moléculaire sont obtenues, et les graphes correspondants sont modélisés. Les graphes correspondant à des molécules réalistes sont ensuite considérés comme solution.

L'inversion des QSPR telle que présentée ci-dessus n'est pas toujours évidente, car elle nécessite de résoudre des équations établies sur mesure en fonction de la propriété ciblée. Miyao *et coll.* ont proposé une alternative qui restreint les générations par une méthode de boîte englobante avec les valeurs de descripteurs MCD.<sup>117</sup> Dans cette méthode, les molécules de la base QSPR initiale qui possèdent la valeur de propriété souhaitée sont identifiées comme référence. Les valeurs de leurs MCDs sont ensuite calculées, puis des molécules sont construites itérativement par ajouts d'atomes ou de fragments jusqu'à ce que leurs MCDs atteignent des valeurs incluses parmi les valeurs des molécules de référence.

### **3.4 Génération de molécules par apprentissage profond**

Une autre thématique est abordée dans cette revue et fait l'objet de la partie intitulée « Deep Learning for Molecular Generation », traitant des algorithmes de génération par apprentissage profond. L'apprentissage profond se base sur l'utilisation des réseaux de neurones dits profonds, c'est-à-dire des réseaux utilisant plusieurs couches de neurones consécutives.<sup>118</sup> Contrairement aux générations guidées par les modèles QSPR, où la qualité des molécules générées est fonction de la qualité des QSPR, les méthodes d'apprentissage profond apprennent directement à reconnaître des molécules et à en générer des similaires. Les molécules sont pour cela très souvent encodées à partir de leur chaîne SMILES (sous forme de vecteurs « one-hot »,



comme présenté sur la Figure 13), ou à partir de leur graphe moléculaire (sous forme de matrices et ensembles de variables).

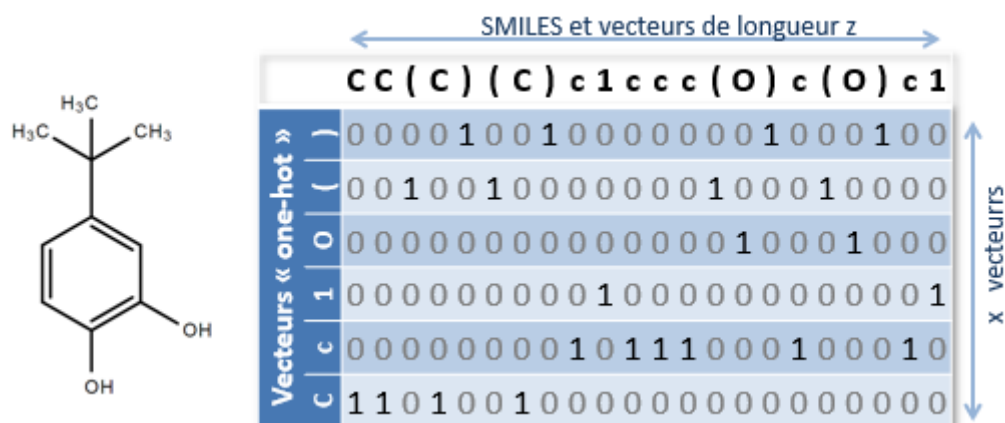


Figure 13 : Représentation du graphe moléculaire, notation SMILES et encodage sous forme de vecteurs « one-hot » de la molécule de 4-tert-Butylcatechol. Adapté avec permission de Gantzer et coll.<sup>34</sup>, Copyright Wiley 2019.

Parmi les outils d'apprentissage profond pour la génération, le réseau de neurones récurrent (RNN, pour « Recurrent Neural Network ») est un réseau qui apprend un langage à partir de la distribution des caractères dans les chaînes initiales.<sup>119</sup> La génération s'effectue ensuite en générant des chaînes caractère par caractère, en se basant sur la probabilité de chaque nouveau caractère en fonction des caractères de la chaîne déjà générée et en fonction des probabilités de chaque caractère dans les chaînes initiales. Des cellules LSTM<sup>120</sup> (pour « Long Short Term Memory »), qui comme leur nom l'indique possèdent une mémoire des caractères déjà générés, sont souvent utilisées dans ce but comme cellules constituant les RNN.

La génération de molécules peut aussi être réalisée avec un auto-encodeur variationnel (VAE, pour « Variational AutoEncoder »). Le VAE est un ensemble de deux RNN (comme présenté sur la Figure 14). Selon cette architecture, le premier réseau « encode » les molécules dans un espace latent, le second est chargé de retranscrire les vecteurs issus de cet espace en structures moléculaires.<sup>121</sup> Les vecteurs à décoder sont issus soit directement de l'encodeur (des molécules initiales) et sont bruités, ou soit d'une extraction manuelle de vecteurs dans l'espace latent, après avoir identifié les zones de l'espace latent dans lesquelles se trouvent des molécules intéressantes pour l'application souhaitée. Sattarov *et coll.* ont implémenté un VAE et utilisé la technique GTM à la fois pour représenter l'espace latent et pour y sélectionner les vecteurs à transcrire en molécules.<sup>122</sup>

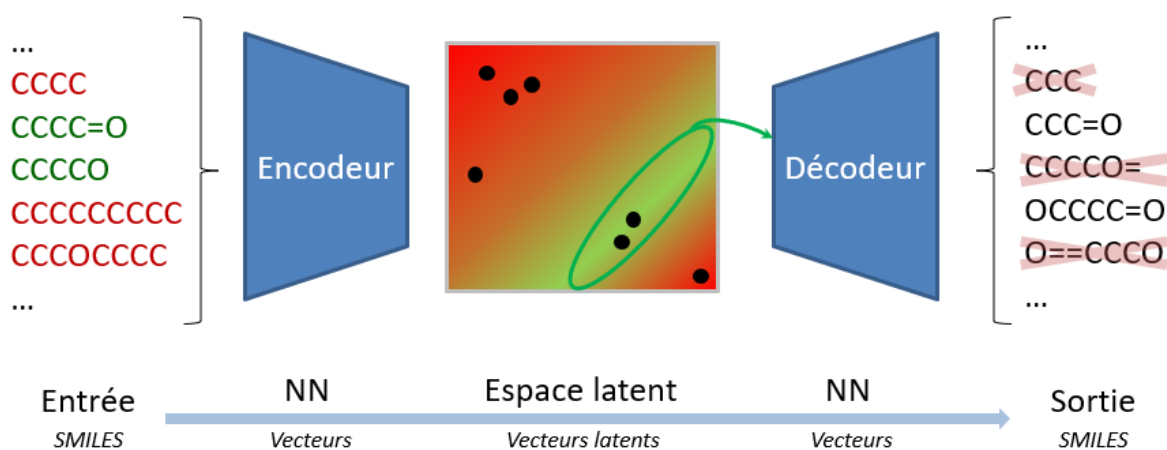


Figure 14 : Représentation d'un VAE et de son fonctionnement pour la génération moléculaire. Les molécules initiales, sous forme de SMILES, sont présentées avec une couleur fonction de leur propriété (rouge : indésirable, vert : désirable). Les chaînes SMILES sont ensuite encodées dans un espace latent. Une région de l'espace latent est sélectionnée (ellipse verte), et les vecteurs de cette région sont décodés en SMILES. Certains SMILES décodés sont écartés, soit car ils ne correspondent pas à des molécules valides, soit car les molécules résultantes ne possèdent pas la propriété désirée. Adapté avec permission de Gantzer et coll.<sup>34</sup>, Copyright Wiley 2019.

Le réseau antagoniste génératif est également un ensemble de deux réseaux de neurones. Le premier réseau, générateur, est chargé de reconstruire des molécules à partir de leur représentation bruitée. Le second réseau, évaluateur, note les aptitudes du générateur en fonction de la validité des molécules produites, et parfois aussi en fonction de leur valeur prédite de propriété. Itérativement, en suivant les indications de l'évaluateur, le générateur devient capable de proposer des molécules valides.<sup>123</sup> Il s'agit d'un apprentissage par renforcement.

Enfin, avec l'apprentissage par *transfert*, il est possible d'entraîner un réseau de neurones profond en deux étapes.<sup>124</sup> La première étape consiste à entraîner le réseau sur un maximum de structures afin d'apprendre la syntaxe moléculaire. Ensuite, un plus petit ensemble de molécules spécifique au problème est utilisé pour finaliser l'apprentissage et spécialiser la génération. Il est aussi possible de contraindre la génération par l'utilisation d'un motif moléculaire, Langevin *et coll.*<sup>125</sup> ont par exemple proposé d'induire en entrée des réseaux de neurones une chaîne SMILES incomplète, forçant le modèle à conserver dans les nouvelles molécules les fragments déjà présents en entrée.

Toutes ces techniques d'apprentissage profond sont indépendantes et peuvent être utilisées conjointement. Bung *et coll.*<sup>126</sup> ainsi que Santana *et coll.*<sup>127</sup> ont chacun proposé de nouvelles structures pour inhiber la protéase principale du virus du syndrome respiratoire aigu sévère (SARS-CoV-2) en utilisant plusieurs méthodes d'apprentissage profond. Bung *et coll.* ont entraîné un premier réseau (générateur) avec environ  $1,6 \times 10^6$  SMILES et parallèlement un autre réseau (utilisé comme modèle QSPR) avec 2 515 SMILES.<sup>126</sup> Par *transfert*, le générateur a ensuite été placé dans un réseau antagoniste génératif avec le modèle QSPR comme évaluateur. Santana *et coll.* ont également entraîné un premier réseau génératif avec environ  $1,9 \times 10^6$  SMILES et ont ensuite transféré les paramètres du premier réseau dans un nouveau réseau, affiné en le réentraînant avec plus de  $2,8 \times 10^5$  SMILES.<sup>127</sup> Le réseau affiné a ensuite servi pour générer de nouvelles molécules. Parallèlement, par transfert et transformation du réseau affiné, un modèle QSPR de classification a été obtenu. Ce modèle a permis d'évaluer les molécules générées.

En apprentissage *profond* et contrairement aux méthodes de la partie précédente, les molécules sont construites par mimétisme sans utiliser de descripteurs ou de fragments prédéfinis. Cette caractéristique permet de moins restreindre l'espace chimique disponible pour la génération qu'en utilisant les méthodes présentées dans la partie précédente. Cependant, un grand nombre de molécules est nécessaire pour entraîner les réseaux de neurones, même si l'apprentissage par *transfert* permet d'atténuer cette contrainte (voir le Tableau 1, 0, qui présente le nombre de molécules utilisées pour entraîner les méthodes d'apprentissage *profond*).

### 3.5 Evaluation des méthodes de génération

Face à l'émergence des méthodes de génération virtuelle, le choix d'un algorithme pour une génération spécifique n'est pas évident. Le besoin d'outils pour comparer et évaluer les algorithmes existants a alors été soulevé. Nous n'avons pas abordé l'évaluation des méthodes de génération dans notre article de revue, car ce sujet, plus récent encore que la génération par réseaux de neurones, n'était pas encore mentionné dans la littérature. Selon nos recherches postérieures, l'évaluation des algorithmes de génération est effectuée en analysant les molécules générées à l'aide d'indices.<sup>128</sup> Nous proposons de classer ces indices en deux ensembles : les indices *individuels* et les indices de *collectifs*.

Les indices *individuels* reflètent les capacités des modèles à produire des molécules valides et répondant à des contraintes. Ils sont uniquement calculés à partir des molécules générées. Il s'agit par exemple du pourcentage de molécules générées valides (indice de validité), valides et non générées précédemment (unicité), uniques et non présentes dans le jeu initial (nouveau) (Les indices d'unicité et de nouveauté sont définis dans la section 4.1 « Evaluation de la qualité des générations » pour plus de clarté). La valeur de certaines propriétés ou leur distance à des valeurs cibles de propriété (telles que l'accessibilité synthétique, la ressemblance à un médicament, le coefficient de partage n-octanol/eau...) sont également considérées comme des indices *individuels*. La majorité des publications utilise déjà ces indices pour évaluer les performances des méthodes proposées.<sup>121,123–127</sup>

Les indices de *collectifs* examinent les molécules générées face à des molécules de référence. Il peut s'agir de comparaisons « une-à-une » (la similarité de chaque molécule générée à une molécule de référence).<sup>121,124,126</sup> Des comparaisons entre deux ensembles de molécules peuvent aussi être effectuées à l'aide d'indices liés aux distributions de descripteurs ou de propriétés dans ces deux ensembles.<sup>123</sup> L'idée est alors d'évaluer la capacité des méthodes de génération à produire des molécules similaires, en termes de structures et de propriétés, aux molécules initiales.

Les plateformes d'analyse comparative (PAC) sont des outils pour comparer les méthodes de génération. Leur fonctionnement est similaire : elles utilisent une base de référence pour entraîner des méthodes de génération, des molécules sont ensuite générées, puis les molécules produites par chaque méthode sont comparées à l'aide d'indices *individuels* et *collectifs*. Les méthodes de génération les plus performantes sont celles qui obtiennent les meilleures valeurs d'indices. Trois PAC ont été proposés dans la littérature par Arús-Pous *et coll.*<sup>129</sup>, Polykovskiy *et coll.*<sup>130</sup> (MOSES) et Brown *et coll.*<sup>131</sup> (GuacaMol). Elles sont codées en Python et leur code est disponible sur la plateforme GitHub, rendant leur utilisation accessible à tous. Le Tableau 10 récapitule les caractéristiques de chacune de ces PAC.

<b>Auteurs (algorithme)</b>	Arús-Pous <i>et coll.</i> <sup>129</sup>	Polykovskiy <i>et coll.</i> <sup>130</sup> MOSES	Brown <i>et coll.</i> <sup>131</sup> GuacaMol
<b>Base de référence</b>	GDB-13*	ZINC Clean Leads*	ChEMBL*
<b>Validité</b>	X	X	X
<b>Unicité</b>	X	X	X
<b>Nouveauté</b>	X	X	X
<b>Comparaisons de distributions</b>	de propriétés	de propriétés et de fragments	de propriétés
<b>Comparaisons individuelles</b>		Similarités de structure	Capacité à générer des substances actives connues

\* base de données filtrée ou partielle

Tableau 10 : Caractéristiques des différentes plateformes d'analyses comparatives existantes.

L'évaluation des molécules générées s'effectue en deux étapes. Les trois PAC utilisent d'abord un ensemble d'indices *individuels*. Ils vérifient tous la validité des molécules, leur nouveauté et leur unicité. Les PAC comparent ensuite les molécules générées aux molécules de référence (ayant été utilisées pour entraîner les générations), se basant sur l'hypothèse que les méthodes de génération doivent produire des molécules similaires à celles de référence. Les valeurs de descripteurs et de propriétés moléculaires (par exemple, le poids moléculaire et le coefficient de partage n-octanol/eau dans les trois outils), ou les distributions de ces valeurs, sont alors comparées. Certains indices *individuels* supplémentaires comme la capacité à générer une molécule cible peuvent aussi être employés. Dans cette étape, la spécificité de chaque outil réside dans la base de référence et dans les indices *collectifs* utilisés :

- L'outil proposé par Arús-Pous *et coll.*<sup>129</sup> utilise comme référence la base virtuelle GDB-13, regroupant les molécules générées de façon exhaustive jusqu'à 13 atomes lourds (non-hydrogène) et mentionnée dans la partie 3.2 « Génération non guidée de molécules ». <sup>101</sup> Les auteurs se basent sur l'idée que les molécules issues de générations dites « partielles » (effectuées avec les méthodes de génération comparées, qui sont entraînées avec une partie

des molécules de la base GDB-13) doivent représenter un maximum de l'espace chimique occupable (représenté par la base GDB-13). Les distributions des propriétés des molécules générées ont été comparées visuellement à celles des molécules de GDB-13 en superposant les distributions. Les courbes des distributions, très similaires, ont démontré que les modèles génératifs comparés généralisent la base initiale.

- La plateforme Molecular Sets (MOSES) utilise comme référence la base de données ZINC<sup>9</sup> regroupant des molécules commercialisées, qui a été filtrée pour ne conserver que les molécules désirables en chimie médicinale (les atomes chargés, les métaux, les époxydes et aldéhydes ont par exemple été supprimés, car ils peuvent endommager les protéines). Les molécules générées sont triées avec ces mêmes filtres. Ensuite, les distributions des propriétés utiles à la chimie médicinale entre les molécules initiales et générées sont comparées à l'aide d'indices, tout comme les distributions des occurrences de fragments. L'outil calcule également la similarité de structures entre chaque molécule générée et la molécule initiale la plus proche, ainsi qu'entre les structures générées.
- La plateforme GuacaMol<sup>131</sup> utilise comme référence la base ChEMBL 24, regroupant des molécules possédant une activité biologique.<sup>10</sup> Les molécules générées sont également filtrées pour écarter celles ne convenant pas à une application en chimie médicinale. Ensuite, les distributions de propriétés entre les molécules initiales et générées sont comparées avec la divergence de Kullback-Leibner.<sup>132</sup> La capacité des méthodes à générer certaines molécules actives connues est également calculée.

Ces PAC permettent de vérifier dans un premier temps des paramètres importants pour toute génération avec des indices *individuels* : validité, unicité et nouveauté des molécules générées. Des indices *collectifs* sont ensuite employés pour évaluer la similarité des molécules générées aux molécules initiales. Cette approche est adaptée pour juger les capacités des méthodes à générer des molécules similaires aux molécules initiales. Cependant, les générations de molécules possédant une autre chimie que celle de la référence (par exemple, possédant d'autres types d'atomes, ou ne passant pas les filtres de chimie médicinale appliqués par MOSES et GuacaMol) ne peuvent pas être comparées avec ces outils si on ne change pas la base de référence et/ou les filtres. Dans le cas d'un changement de base initiale, les PAC ne semblent pas adaptées pour évaluer la qualité de génération dans le cas où la base initiale n'est pas représentative. Les bases de données industrielles, par exemple, ne sont souvent pas très grandes, et de ce fait peuvent ne pas posséder assez de molécules pour pleinement représenter,

de manière représentative, la diversité chimique de toutes les molécules pouvant être générées. Enfin, les PAC ne discutent pas de la variation de performance des générations en fonction de la base initiale. Par exemple, est-ce que les méthodes de générations par apprentissage profond sont aussi efficaces pour générer des complexes organométalliques qu'elles le sont pour générer des molécules liées à la chimie médicinale ?

### 3.6 Conclusions sur la partie méthodes pour la génération

Nous venons de présenter plusieurs méthodes pour générer de nouvelles molécules. Parmi elles, les méthodes sans contrainte sur la propriété sont les plus simples à mettre en place, mais elles peuvent générer un grand nombre de structures sans intérêt. Les méthodes i-QSPR utilisent quant à elles les modèles QSPR pour guider les générations, en imposant aux nouvelles molécules des valeurs de descripteurs et/ou en vérifiant la convergence de leur valeur de propriété lors de leur construction itérative. Les méthodes d'apprentissage *profond* construisent de nouvelles molécules par mimétisme des structures initiales. Nous avons mentionné tout au long du chapitre des travaux récents de génération utilisent aussi bien des méthodes i-QSPR que d'apprentissage profond, montrant que les deux approches sont actuellement prisées. Parallèlement au développement de méthodes génératives, des outils pour les comparer ont vu le jour. Ils utilisent des indicateurs regroupés en indices *individuels*, jugeant les performances de chaque molécule individuellement, et *collectifs*, comparant l'ensemble des molécules générées avec celles de référence. Des plateformes d'analyse comparative (PAC) permettent d'évaluer différentes méthodes de génération à l'aide de ces indices, de manière identique. Cependant, les PAC évaluent la similarité des molécules générées à celles répertoriées dans des bases connues. Comme nous l'avons vu, cela ne permet pas de comparer les méthodes de générations pour une application spécifique et/ou personnalisée.

Cette étude de la littérature nous a permis d'affiner nos choix de recherche pour le travail de thèse. Nous avons d'abord souhaité pouvoir implémenter et améliorer trois méthodes de génération issues de la littérature pour une utilisation à IFPEN, à savoir la génération par assemblage de fragments, par modifications successives (à l'aide des graphes génétiques) et par autoencodeur variationnel. La génération par assemblage de fragments est la méthode la plus simple à mettre en place pour obtenir de nouvelles molécules. La méthode par modifications

successives, quant à elle, permet d'orienter la génération vers des valeurs de propriétés sans mettre en place et résoudre des systèmes d'équations pour contraindre les valeurs de descripteurs. La génération par autoencodeur variationnel nécessite davantage de molécules que les deux méthodes précédentes, mais présente également la capacité de pouvoir générer des molécules dans un processus automatique. Ainsi, nous pourrions posséder trois outils complémentaires pour la génération et maximiser les chances d'obtenir de nouvelles molécules avec des valeurs de propriétés souhaitées.

Dans un second temps, nous avons souhaité pouvoir améliorer les outils d'évaluation des méthodes de génération pour permettre leur utilisation avec des jeux de molécules plus petits (contenant moins de  $10^3$  molécules) et contenant plusieurs types de chimie. Pour cela, nous présentons une nouvelle série d'indices *collectifs* utilisant une base de référence construite dynamiquement à partir des molécules générées, sans s'appuyer sur une base de référence fixe.



## Chapitre 4. Génération virtuelle de molécules

---

Nous décrivons dans ce chapitre la mise en place de processus automatiques, pour générer virtuellement de nouvelles structures moléculaires, parmi les outils d'IFPEN. Particulièrement, ces processus doivent être adaptés à la génération à partir de bases de données industrielles qui peuvent contenir peu de molécules (moins de 1 000 molécules). Trois types d'approches, présentés dans le Chapitre 3 « Méthodes pour la génération moléculaire et pour leur comparaison » sont considérées : la génération par assemblage de fragments (méthodes « F »), la génération par modification successive de structures (méthodes « G ») et la génération par autoencodeur variationnel (méthodes « E »). Pour chaque approche, plusieurs améliorations sont mises en place de manière à maximiser le nombre de nouvelles molécules appartenant à l'AD du modèle QSPR utilisé.

### 4.1 Evaluation de la qualité des générations

Nous évaluons les méthodes de génération et l'impact des améliorations implémentées sur celles-ci à partir des molécules produites. Chaque molécule générée est classée dans une ou plusieurs catégories présentées dans le Tableau 11.

Catégories de molécules	Conditions
Valides	Générées avec une structure réaliste.
Uniques	Valides, épurées des molécules doublons.
Nouvelles	Uniques et non présentes dans le jeu initial.
Prédictibles	Nouvelles et appartenant du modèle QSPR.
Redécouvertes	Présentes dans un jeu externe de molécules et générées.

Tableau 11 : Les différentes catégories de molécules générées et leurs conditions.

Des indices *individuels*, ou ratios sont calculés à partir du nombre de molécules dans chaque catégorie. Les indices d'unicité (Equation (6)) et de nouveauté (Equation (7)), sont ceux présentés dans la partie 3.5 « Evaluation des méthodes de génération ». L'indice de prédictibilité est défini comme le nombre de molécules prédictibles sur le nombre de molécules nouvelles (Equation (8)). Même si, en théorie, la prédiction par QSPR est possible pour toutes les molécules générées, la valeur prédite ne peut être considérée comme fiable que si la molécule est dite prédictible – appartient à l'AD du modèle QSPR. L'indice global regroupe l'information donnée par les indices précédents. À la manière de l'indice « Uniformity–Completeness–Closedness Ratio », défini dans le papier d'Arus *et coll.*<sup>129</sup> comme le produit de trois indices distincts, l'indice global est défini comme le rapport des indices de validité, unicité, nouveauté et prédictibilité, ou encore comme le nombre de molécules prédictibles sur le nombre de molécules générées (Equation (9)). L'indice de redécouverte est inspiré de l'outil GuacaMol, où la capacité des algorithmes à générer certaines molécules biologiquement actives est évaluée.<sup>131</sup> Dans notre implémentation, nous fournissons une liste de molécules, différentes des molécules initiales, que nous souhaitons régénérer. À la fin de la génération, nous recensons les molécules qui ont effectivement été régénérées et l'indice de redécouverte est défini comme le rapport entre le nombre de molécules qui ont pu être régénérées et le nombre de molécules constituant cette liste prédéfinie (Equation (10)).

$$\text{Unicité} = \#Uniques / \#Générées \quad (6)$$

$$\text{Nouveauté} = \#Nouvelles / \#Uniques \quad (7)$$

$$\text{Prédictibilité} = \#Prédictibles / \#Nouvelles \quad (8)$$

$$\begin{aligned} \text{Global} &= \text{Unicité} * \text{Nouveauté} * \text{Prédictibilité} \\ &= \#Prédictibles / \#Générées \end{aligned} \quad (9)$$

$$\text{Redécouverte} = \#Regénérées / \#A \text{ régénérer} \quad (10)$$

Où # signifie nombre de molécules.

Les molécules présentes dans la base de données sur le point d'éclair ont été utilisées dans nos travaux comme ensemble de molécules initiales, et le modèle QSPR dérivé de cette base, construit dans la partie 2.2.2.2 « Modélisation par SVR de la propriété de point d'éclair », a servi comme outil de prédiction. Nous avons utilisé la base test de ce modèle comme liste de molécules prédéfinies.

## 4.2 Génération par assemblage de fragments

En nous inspirant des travaux présentés dans le Chapitre 3 « Méthodes pour la génération moléculaire et pour leur comparaison », et plus particulièrement de la méthode proposée par Nilakantan *et coll.*<sup>99</sup>, nous avons implémenté une méthode pour générer des molécules sous forme de graphes à partir de fragments, nommée « F ».

Les fragments utilisés pour la construction sont définis en considérant les molécules initiales de deux manières différentes :

- Ils peuvent être fournis par l'utilisateur qui liste manuellement des fragments pertinents dans les molécules initiales.
- Ils peuvent être directement extraits des molécules initiales à partir de leur encodage sous forme de SMILES ou encore sous forme de descripteurs, dont les SMF font partie.

La construction d'une nouvelle structure moléculaire est un processus itératif qui se déroule selon les étapes suivantes :

1. Deux fragments sont sélectionnés, les atomes d'hydrogène présents sont supprimés, et la valence des autres atomes est calculée.
2. Les fragments sont assemblés. S'il existe un atome sur chaque fragment possédant une valence inférieure à leur valeur de valence maximale, une liaison est créée entre les deux fragments à partir de ces atomes. S'il existe plusieurs atomes disponibles pour créer une liaison sur chaque fragment, le choix des atomes se fait de façon aléatoire.
3. Des atomes d'hydrogène sont ajoutés pour saturer le graphe. Nous considérons la saturation d'un graphe comme la création de liaisons entre de nouveaux atomes d'hydrogène et les atomes du graphe dont la valence maximale n'est pas atteinte.
4. La structure est considérée comme un nouveau fragment, à partir duquel les étapes 1 à 3 peuvent être répétées jusqu'à atteindre l'un des deux critères de fin :
  - Le premier critère de fin est défini comme l'atteinte ou le dépassement du nombre d'atomes lourds (autre qu'hydrogène) maximum – aléatoirement défini pour chaque graphe dans l'intervalle [2 ; 30] (l'intervalle du nombre d'atomes lourds dans le jeu initial sur le point d'éclair).
  - Le second critère de terminaison est atteint si le graphe en cours de construction ne possède plus d'atome lourd capable de participer à une nouvelle liaison avec un

nouveau fragment (par exemple : la molécule d'éthènedione, de notation SMILES O=C=C=O).

Le SMILES canonique du graphe obtenu est ensuite établi. Ce SMILES canonique est comparé à ceux des molécules du jeu initial puis à ceux des molécules précédemment générées. Tout doublon pénalise les scores de nouveauté ou d'unicité. Pour cette nouvelle structure, les valeurs des descripteurs intervenant dans le modèle QSPR sont calculées et on s'assure que le modèle est applicable à cette molécule, selon la procédure décrite dans la partie 2.1.1.5 « Domaine d'applicabilité ». Si un seul des critères de l'AD n'est pas respecté, la molécule est écartée et pénalise le score de prédictibilité.

Cette méthode a été implémentée dans un script codé en Python, utilisant la librairie « RDKit »<sup>45</sup> pour manipuler les atomes et liaisons au sein des graphes, et pour convertir les nouveaux graphes en SMILES.

Nous présentons dans les parties suivantes l'application et l'évaluation de la méthode de génération « F ». Comme nous le verrons, plusieurs contraintes ont été mises en place pour améliorer la qualité des générations. Ci-après, chaque génération produit 10 000 structures et est répétée 10 fois pour évaluer sa qualité rigoureusement. Dix répétitions sont suffisantes pour obtenir des valeurs d'indices *individuels* dont l'écart moyen ne dépasse pas 0,01.

#### 4.2.1 Assemblage de fragments simples

Dans un premier temps, les fragments « simples » suivants ont été employés pour la génération : C, O, C-C, C=C, C#C, C-O et C=O (où - représente une liaison simple, = une liaison double, # une liaison triple). Nous supposons que ces fragments « simples » permettent du fait de leur taille de construire des molécules variées et d'explorer tout l'AD du modèle QSPR.

La méthodologie de génération par assemblage de fragments simples a été employée dans un premier temps sans contrainte, laissant l'algorithme libre du choix des fragments à ajouter et des liaisons à créer. Le Tableau 12 présente la valeur moyenne de chaque indice individuel sur les dix générations. Toutes les molécules générées sont valides, démontrant que les étapes de construction se déroulent correctement tout comme la conversion de graphe moléculaire à notation SMILES. Les indices d'unicité et de nouveauté indiquent que plus de 80% des molécules générées sont uniques et nouvelles. Cependant, moins d'un dixième des molécules nouvelles est prédictible avec le modèle QSPR (valeur de l'indice de prédictibilité égale à 0,09),

les autres molécules contenant des fragments inconnus et/ou un nombre trop élevé de certains fragments vis-à-vis du jeu initial du modèle.

<b>Indice</b>	<b>Valeur</b>	<b>Nombre</b>
Unicité	0,81	8 123 ± 24
Nouveauté	0,99	8 064 ± 24
Prédictibilité	0,09	723 ± 17
Global	0,07	723 ± 17

Tableau 12 : Valeurs moyennes des indices *individuels* pour dix générations par assemblage libre de fragments simples. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Après l'ajout d'un fragment à la structure en cours de construction, la décomposition de la nouvelle structure en fragments SMF contient : les précédents fragments, le nouveau fragment, mais également d'autres fragments. Ces autres fragments comprennent (i) des atomes et liaisons appartenant à la structure en cours de construction, (ii) des atomes et liaisons appartenant au fragment ajouté et (iii) la liaison nouvellement créée entre le fragment ajouté et la structure en cours de construction. Le choix du degré de la liaison (iii) semble alors important pour obtenir des molécules ne possédant pas de fragments inconnus du jeu initial – et donc des molécules prédictibles. Nous avons testé dans cette optique une variation de génération par assemblage de fragments qui contraint le choix des liaisons créées entre les fragments ajoutés. Pour cela, nous avons dénombré et exprimé en pourcentage de présence les trois types de liaisons au sein du jeu initial de structures dans notre base de point d'éclair (Tableau 13). Les liaisons des noyaux aromatiques ont été aléatoirement considérées dans l'une de leurs formes de résonance. Le choix du degré de la liaison à créer lors de la combinaison des fragments suit la distribution présentée dans le Tableau 13. Cette contrainte est souple. Elle ne force pas la génération de molécules dans l'AD. Elle l'induit en s'inspirant de la diversité des liaisons du jeu initial.

Type de liaison	Distribution
Simple	77%
Double	22%
Triple	1%

Tableau 13 : Distribution du type de liaisons dans les molécules du jeu initial.

Indice	Valeur	Nombre
Unicité	0,83	8 342 ± 22
Nouveauté	0,99	8 259 ± 23
Prédictibilité	0,18	1 447 ± 16
Global	0,14	1 447 ± 16

Tableau 14 : Valeurs moyennes des indices *individuels* pour dix générations par assemblage de fragments simples avec contraintes sur le choix des liaisons entre fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Le Tableau 14 présente les résultats moyens de l'application de cette contrainte sur dix générations. Les valeurs des indices d'unicité et de nouveauté sont similaires à celles obtenues précédemment, la contrainte sur le choix du type de liaisons influence donc peu le nombre molécules nouvelles. Cependant, la valeur de l'indice de prédictibilité augmente par rapport à la génération sans contrainte sur les liaisons, ce qui indique que moins de molécules sont générées en dehors de l'AD du modèle QSPR. De ce fait, la valeur de l'indice global est doublée en comparaison de celle des générations précédentes.

## 4.2.2 Assemblage de fragments SMF

Dans un second temps, nous avons inversé le modèle QSPR décrit dans la partie 2.2.2.2 « Modélisation par SVR de la propriété de point d'éclair ». Pour rappel, ce modèle QSPR utilise les descripteurs SMF de type t3l2u4, correspondant à des fragments de deux à quatre atomes. Ces fragments ont été extraits et employés pour le travail de génération de structures présenté dans cette sous-section.

Comme avec les fragments simples (sous-section 4.2.1), de nouvelles structures moléculaires ont d'abord été générées sans utiliser de contraintes. Le Tableau 15 présente la valeur moyenne de chaque indice individuel sur les dix générations utilisant ces fragments.

Indice	Valeur	Nombre
Unicité	0,90	8 995 ± 26
Nouveauté	0,99	8 934 ± 28
Prédictibilité	0,22	1 986 ± 28
Global	0,20	1 986 ± 28

Tableau 15 : Valeurs moyennes des indices *individuels* pour dix générations par assemblage libre de fragments SMF. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

90% des molécules générées le sont de manière unique et 99% d'entre elles sont nouvelles. Seuls 22% des molécules nouvelles sont prédictibles par le modèle QSPR, les autres (78%) possédant soit de nouveaux fragments soit leurs valeurs de descripteurs hors des bornes du jeu initial. Un cinquième des molécules générées (valeur d'indice global égale à 0,20) correspond à des molécules prédictibles. La génération à partir de fragments SMF surperforme les générations précédentes utilisant des fragments simples (partie 4.2.1), qui avaient obtenu des valeurs d'indice global égales à 0,07 (sans contrainte sur le choix des liaisons entre fragments) et 0,14 (avec contrainte sur le choix des liaisons entre fragments).

Une quantité importante de molécules générées est toujours considérée hors du domaine d'applicabilité, car la génération n'est pas supervisée. En effet, même si la bibliothèque de fragments à disposition du modèle est obtenue à partir des molécules initiales, ni la manière

d'assembler les fragments ni leur choix au sein de la bibliothèque n'est restreint. Comme précédemment, des contraintes ont ensuite été mises en place pour inciter le modèle à générer davantage de molécules dans le domaine d'applicabilité. Ces contraintes concernent le choix du degré de la liaison chimique reliant deux fragments et/ou le choix des fragments. Comme précédemment ces contraintes ne forcent pas la génération de molécules dans l'AD. Elles incitent cependant les molécules à respecter l'AD pendant leur construction, en pondérant les choix des fragments et des degrés de liaisons, selon la diversité des molécules du jeu initial.

Le premier type de contraintes a déjà été utilisé pour la génération avec des fragments simples et concerne le choix du degré de la liaison à créer. Le choix du degré de la liaison à créer lors de la combinaison des fragments SMF suit la distribution précédemment présentée dans le Tableau 13. Le Tableau 16 présente les résultats moyens de l'application de cette contrainte sur dix générations.

<b>Indice</b>	<b>Valeur</b>	<b>Nombre</b>
Unicité	0,90	8 985 ± 21
Nouveauté	0,99	8 908 ± 22
Prédictibilité	0,30	2 658 ± 39
Global	0,27	2 658 ± 39

Tableau 16 : Valeurs moyennes des indices *individuels* pour dix générations par assemblage de fragments SMF avec contraintes sur le choix des liaisons entre fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

L'application de cette contrainte permet de générer 8% de molécules valides supplémentaires dans le domaine d'applicabilité (l'indice de prédictibilité augmente de 0,08) comparé à la méthode sans contrainte. Nous avons manuellement analysé les molécules générées hors de l'AD et avons remarqué que l'application de cette contrainte permet principalement de faire diminuer le pourcentage de molécules faisant apparaître de nouveaux fragments par rapport aux générations n'utilisant pas cette contrainte. La valeur de l'indice global augmente également ; plus du quart des molécules générées sont à présent prédictibles.



Le deuxième type de contraintes concerne le choix des fragments à ajouter. Chaque fragment possède une probabilité différente d'être choisi pendant la génération. Cette probabilité est définie pour chaque fragment comme suit (Equation (11)) :

$$P_{\text{fragment}} = \frac{\text{Occurrence}_{\text{fragment}}^{\text{maximale}}}{\sum_{\text{fragments}} \text{Occurrence}_{\text{fragment}}^{\text{maximale}}} \quad (11)$$

Où  $\text{Occurrence}_{\text{fragment}}^{\text{maximale}}$  est le nombre maximal d'occurrences du fragment dans une molécule du jeu initial. Techniquement, comme les fragments sont repris des descripteurs ISIDA,  $\text{Occurrence}_{\text{fragment}}^{\text{maximale}}$  est égal à la valeur maximale sur le jeu initial du descripteur encodant le fragment.

Le Tableau 17 présente les performances moyennes des générations combinées à une contrainte sur les fragments, uniquement. On peut relever une diminution de l'indice d'unicité (valeur de 0,90 sans contrainte et de 0,83 avec contrainte sur le choix des fragments), démontrant que davantage de molécules ont été générées plusieurs fois. Le nombre de molécules nouvelles reste globalement constant comparé aux générations précédentes. L'indice de prédictibilité passe de 0,22 (sans contrainte) à 0,57, l'indice global passe de 0,27 à 0,47. La génération avec contraintes sur le choix des fragments obtient les meilleures performances sur toutes les générations étudiées jusqu'à présent, quasiment la moitié des molécules générées sont prédictibles.

Indice	Valeur	Nombre
Unicité	0,83	8 353 ± 21
Nouveauté	0,98	8 243 ± 22
Prédictibilité	0,57	4 717 ± 39
Global	0,47	4 717 ± 39

Tableau 17 : Valeurs moyennes des indices *individuels* pour dix générations par assemblage de fragments SMF avec contraintes sur le choix des fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Les deux contraintes précédentes, sur le choix du type de liaison entre les fragments et sur le choix des fragments, ont permis d'améliorer les performances de génération. Elles ont été utilisées simultanément dans la troisième variation de génération. Le Tableau 18 présente les performances moyennes des générations à travers les valeurs d'indices *individuels*.

<b>Indice</b>	<b>Valeur</b>	<b>Nombre</b>
Unicité	0,81	8 143 ± 19
Nouveauté	0,98	7 997 ± 27
Prédictibilité	0,78	6 237 ± 17
Global	0,62	6 237 ± 17

Tableau 18 : Valeurs moyennes des indices *individuels* pour dix générations par assemblage de fragments SMF avec contraintes sur le choix des fragments et sur des liaisons entre fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Les valeurs d'unicité et de nouveauté sont globalement stables vis-à-vis des générations précédentes. La valeur de prédictibilité augmente encore avec l'usage des contraintes sur les choix du type de liaisons et de fragments, pour atteindre une valeur de 0,78. L'indice global possède une valeur de 0,62. L'association des contraintes sur le choix du type de liaisons et de fragments permet de maximiser le nombre de molécules prédictibles, en diminuant le nombre de molécules pour lesquelles un nouveau fragment est présent ou pour lesquelles les intervalles des valeurs des descripteurs ne sont pas respectés.

### **4.2.3 Conclusions sur la génération par assemblage de fragments**

Nous venons de décrire et mettre en place une méthode de génération de structures moléculaires basée sur un assemblage de fragments. L'utilisation de contraintes sur le type de liaisons à former et sur la sélection des fragments à ajouter ont permis de rendre cette méthode plus efficace. Les contraintes sur le type de liaisons à former semblent indispensables. De ce fait,

pour la suite de nos travaux, nous conservons les variations de génération présentées dans le Tableau 19.

Méthode	Fragments	Pondération du choix	
		Liaisons	Fragments
F0	Simple	X	
F1a	SMF	X	
F1b	SMF	X	X

Tableau 19 : Les différentes variations de génération par assemblage de fragments conservées.

Dans nos approches, des fragments sont ajoutés aux structures en cours de génération jusqu'à atteindre l'un des deux critères de fin – atteindre un nombre prédéfini d'atomes ou obtenir un graphe incapable de créer une nouvelle liaison. Cette opération d'ajout se déroule, quelle que soit la qualité de la structure, y compris si elle ne respecte plus le domaine d'applicabilité du modèle QSPR. Nous avons implémenté un troisième critère de fin de génération pour les méthodes présentées dans le Tableau 19. Le respect du domaine d'applicabilité est dynamiquement vérifié à chaque étape de la construction. Dès qu'il n'est plus respecté, le fragment fraîchement ajouté est supprimé de la structure, la molécule est alors considérée comme finalisée.

### 4.3 Génération par modifications successives des structures

Dans cette section, nous nous intéressons à une autre méthode de génération, nommée G, inspirée des algorithmes génétiques, les graphes génétiques (GG, présentées dans la partie 3.3 « Génération guidée des molécules »).<sup>111</sup> L'objectif est de générer des molécules similaires à celles du jeu initial, en les modifiant par différentes opérations issues des approches génétiques : sélection, croisement, mutation...

Tout comme pour la méthode F, les graphes moléculaires ont été employés pour manipuler les structures et l'encodage SMILES pour stocker les molécules. En effet, les graphes moléculaires

sont des entités de taille variable décrivant explicitement atomes et liaisons. Ces deux derniers éléments peuvent facilement être modifiés par la méthode GG. L'encodage SMILES transmet et stocke l'information moléculaire plus aisément que des successions de chromosomes ou qu'un encodage sous forme de graphes. La méthode G a été implémentée dans un script écrit en Python, utilisant la librairie « RDKit »<sup>45</sup> pour manipuler atomes et liaisons au sein des graphes, et enfin, pour convertir les nouveaux graphes en SMILES.

Sept opérateurs ont été utilisés pour apporter des modifications aux graphes moléculaires. La Figure 15 illustre ces différents opérateurs : l'ajout et la suppression de fragments, le croisement, la mutation de liaison et d'atome, la cyclisation et l'ouverture de cycles.

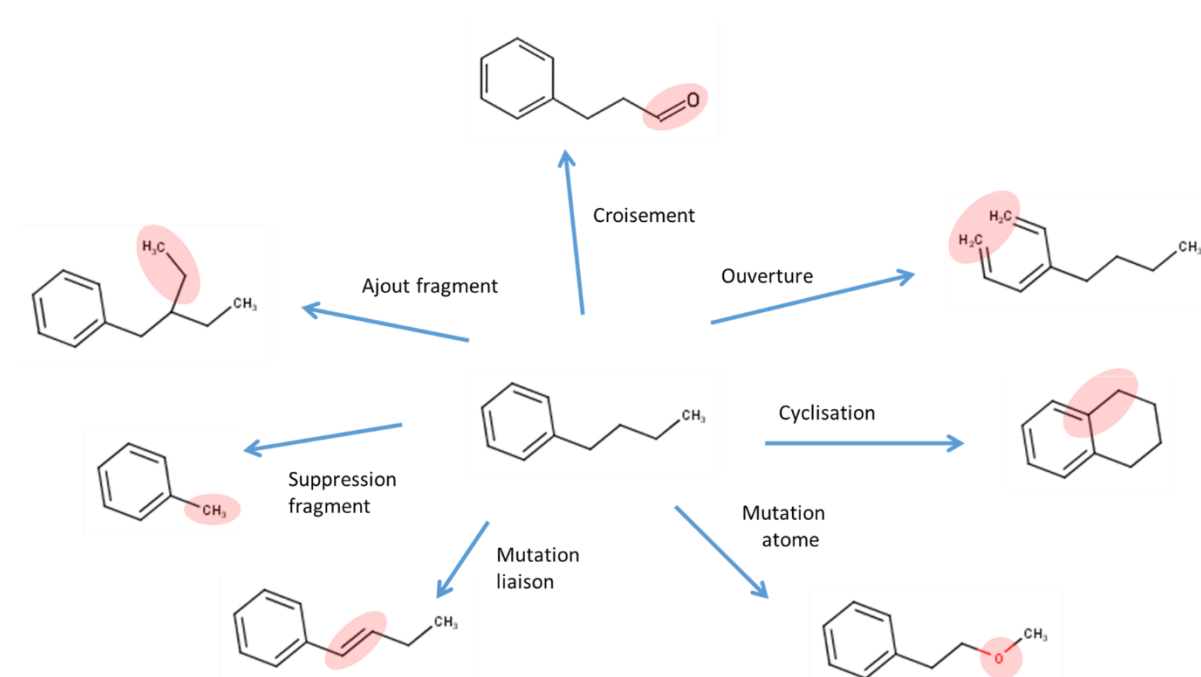


Figure 15 : Opérateurs de modifications implémentés pour application sur les graphes moléculaires. Un exemple est donné pour chaque opérateur, et les modifications effectuées sont mises en valeur par des patches rouges.

Le processus de génération fonctionne de façon itérative. À chaque itération un ensemble de molécules, dites molécules parentes, est sujet à modifications. Dans notre implémentation, les molécules parentes sont les molécules initiales ainsi que celles générées lors des précédentes itérations. Si une opération ne réussit pas à modifier une molécule, la molécule est écartée de l'itération en cours, mais n'est pas retirée des molécules parentes pour les itérations suivantes. À la fin d'une génération, les solutions sont l'ensemble des molécules générées, qu'elles aient été modifiées ou non. Le nombre hypothétique de structures pouvant être générées en  $i$

itérations,  $N_{s,h}$ , est donné par le produit entre le nombre de molécules initiales  $n$  et  $2^i$  (Equation (12)).

$$N_{s,h} = N_{s,i} * 2^i \quad (12)$$

où  $N_{s,i}$  est le nombre de structures initiales et  $i$  le nombre d'itérations.

Nous étudions ci-dessous, indépendamment, l'impact des différents opérateurs appliqués dans notre méthode GG. L'implémentation de chaque opérateur est ajustée pour obtenir un maximum de molécules prédictibles. Les 599 molécules composant le jeu d'apprentissage des modèles QSPR sur le point d'éclair ont été utilisées comme jeu initial. Les molécules composant le jeu de test de ces modèles QSPR ont été utilisées pour calculer l'indice de redécouverte, décrit dans le Tableau 11.

Dans chaque cas, les générations s'effectuent sur cinq itérations. Le nombre hypothétique de structures générées est alors égal à 19 168, il est du même ordre de grandeur que le nombre de molécules obtenues avec la génération par assemblage de fragments ( $10^4$  molécules). Comme pour les générations F, chaque génération G a été lancée dix fois pour s'assurer de sa répétabilité.

### 4.3.1 Ajout et suppression de fragments

La méthode d'ajout et de suppression de fragments, comme l'indique son nom, permet d'ajouter et de supprimer des fragments aux molécules. Nous avons fait le choix de donner à une molécule la même probabilité de subir un ajout qu'une suppression de fragment. Une liste de fragments disponibles à l'ajout est mise en place à partir des fragments précédemment supprimés. Initialement, cette liste contient le fragment méthyl ( $-CH_3$ ) et éthyl ( $-CH_2CH_3$ ).

La méthode d'ajout identifie dans le graphe moléculaire, sur lequel les atomes d'hydrogène ont été supprimés, un atome sur lequel on peut créer une liaison. Il s'agit d'un atome dont la valence est inférieure à sa valence maximale. Un fragment est choisi au hasard dans la liste des fragments disponibles et un atome  $y$  est identifié de la même manière. Une liaison simple est créée entre les atomes identifiés. Des atomes d'hydrogène sont ensuite ajoutés pour saturer le graphe.

La méthode de suppression consiste quant à elle à diviser une molécule en deux, ou encore à supprimer aléatoirement une liaison ne participant pas à un cycle. Un des fragments obtenus est utilisé comme solution, en rajoutant un atome d'hydrogène sur l'atome dont on a supprimé la liaison. L'autre fragment vient compléter la bibliothèque des fragments disponibles pour l'ajout. Dans notre approche, seuls des fragments terminaux peuvent être supprimés (rupture d'une seule liaison).

Une méthode de base ainsi que deux variations ont été testées pour l'ajout et la suppression de fragments. La méthode de base modifie à chaque itération toutes les molécules du jeu initial ainsi que celles précédemment générées, qu'elles soient prédictibles par le modèle QSPR – incluses dans le domaine d'applicabilité – ou non. Après 10 générations de 5 itérations, les valeurs moyennes des indices *individuels* sont présentées avec le nombre moyen de structures associées dans le Tableau 20. La méthode a généré en moyenne  $10\,784 \pm 88$  molécules, ce qui est presque deux fois moins que le nombre hypothétique de structures générées. Nous supposons que cette différence est due au fait que certaines molécules parentes sélectionnées pour un ajout de fragment ne possèdent pas d'atome disponible pour y greffer un fragment. 6 458 molécules sont prédictibles par le modèle QSPR, soit 60% de l'ensemble des molécules générées. Par ailleurs, 43% des molécules du jeu initial de test ont été régénérées.

<b>Indice</b>	<b>Valeur</b>	<b>Nombre</b>
Unicité	0,72	7 816 $\pm$ 84
Nouveauté	0,96	7 501 $\pm$ 83
Prédictibilité	0,86	6 458 $\pm$ 82
Global	0,60	6 458 $\pm$ 82
Redécouverte	0,43	80 $\pm$ 3

Tableau 20 : Valeurs moyennes des indices *individuels* pour dix générations par la méthode de base d'ajout et suppression de fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Une première variation de l'ajout/suppression de fragments a été mise en place. Dans celle-ci, les molécules en dehors du domaine d'applicabilité (contenant de nouveaux fragments ou ne respectant pas les bornes d'au moins un des descripteurs) ne peuvent subir que des suppressions

de fragments. Cette variation se veut générer des molécules respectant « à nouveau » l'AD à partir de celles ne le respectant pas. Après 5 itérations,  $10\,179 \pm 113$  structures sont obtenues en moyenne. Les valeurs moyennes d'indices *individuels* ont été calculées sur les dix générations à partir des molécules générées et sont présentées avec le nombre moyen de structures associées dans le Tableau 21. La première variation a permis de générer en moyenne 6 071 molécules prédictibles, soit également 60% de l'ensemble des molécules générées. La légère augmentation de la prédictibilité est contrastée par la diminution de l'unicité et de la nouveauté. Par ailleurs, 44% des molécules du jeu externes ont été générées. Ces résultats sont très similaires à ceux obtenus avec la méthode de base.

<b>Indice</b>	<b>Valeur</b>	<b>Nombre</b>
Unicité	0,70	$7\,129 \pm 93$
Nouveauté	0,95	$6\,804 \pm 90$
Prédictibilité	0,89	$6\,071 \pm 96$
Global	0,60	$6\,071 \pm 96$
Redécouverte	0,45	$84 \pm 5$

Tableau 21 : Valeurs moyennes des indices *individuels* pour dix générations par la première variation d'ajout et suppression de fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

La deuxième variation de la méthode d'ajout/suppression de fragments ne considère que les molécules respectant l'AD comme molécules à modifier à chaque itération. Les valeurs d'indices *individuels* ont été calculées sur les dix générations de cinq itérations à partir des molécules générées et sont présentées avec le nombre moyen de structures associées dans le Tableau 22. Cette variation a permis de générer en moyenne  $9\,369 \pm 143$  molécules, ce qui est moins que la méthode de base, car le choix des parents est limité dans la variation.  $5\,742 \pm 113$  molécules sont en moyenne prédictibles, soit 61% de l'ensemble des molécules générées. 44% des molécules du jeu externes ont été générées. Même si ces résultats sont similaires à ceux obtenus précédemment, on remarque une augmentation de la valeur de prédictibilité (0,92) vis-à-vis de la méthode de base (0,86) et également de la première variation (0,89).

Indice	Valeur	Nombre
Unicité	0,69	6 546 ± 120
Nouveauté	0,95	6 238 ± 122
Prédicibilité	0,92	5 742 ± 113
Global	0,61	5 742 ± 113
Redécouverte	0,43	81 ± 5

Tableau 22 : Valeurs moyennes des indices *individuels* pour dix générations par la deuxième variation d'ajout et suppression de fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Nous avons implémenté une méthode permettant de générer des molécules par ajout et suppression de fragments définis aléatoirement sur les molécules. Parmi les différentes variations testées, la deuxième, celle ne modifiant que les molécules respectant l'AD, semble être très légèrement plus efficace sur cette base de structures moléculaires : elle permet d'obtenir les meilleures valeurs de prédictibilité et globales. La méthode d'ajout et de suppression de fragments nous permet de générer en 5 itérations à partir des 599 molécules du jeu initial sur le point d'éclair plus de 5 700 nouvelles molécules dans l'espace chimique défini.

### 4.3.2 Croisement de graphes

Le croisement de graphes appliqué dans ce travail est similaire aux croisements opérés par les algorithmes génétiques. Dans notre implémentation deux molécules échangent un fragment. Les fragments sont extraits aléatoirement par la technique de suppression de fragments définie dans la partie 4.3.1 « Ajout et suppression de fragments », puis combinés par la technique d'ajout de fragments. Une méthode de base et deux variations sont étudiées pour le croisement de graphes.

La méthode de base utilise comme précédemment toutes les molécules du jeu initial ainsi que celles précédemment générées comme molécules parentes. Les molécules ainsi générées ont été analysées par les indices *individuels* dont la valeur moyenne sur dix générations de cinq



itérations chacune est présentée dans le Tableau 23. Après 5 itérations,  $9\,337 \pm 86$  structures ont été générées en moyenne. Parmi ces structures, 59% sont prédictibles par le modèle QSPR, ce qui est similaire aux résultats obtenus par ajout et suppression de fragments.

Indice	Valeur	Nombre
Unicité	0,70	$6\,525 \pm 86$
Nouveauté	0,94	$6\,158 \pm 90$
Prédictibilité	0,90	$5\,547 \pm 62$
Global	0,59	$5\,547 \pm 62$
Redécouverte	0,48	$89 \pm 5$

Tableau 23 : Valeurs moyennes des indices *individuels* pour dix générations par la méthode de base de croisement. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

On souhaite, comme précédemment dans la méthode d'ajout et de suppression de fragments, éviter de générer des molécules en dehors du domaine d'applicabilité. La variation de cette méthode consiste à ne modifier que des molécules respectant l'AD pour permettre une génération plus efficace. Les molécules générées avec cette variation ont été analysées par les indices *individuels* dont les valeurs moyennes sur les dix générations de cinq itérations sont présentées dans le Tableau 24. Après cinq itérations,  $8\,628 \pm 100$  molécules ont été générées en moyenne, ce qui est inférieur à la méthode originale car seules les molécules appartenant à l'AD sont modifiées. Les scores d'unicité et de nouveauté sont similaires à ceux de la méthode originale. L'indice de prédictibilité augmente de 0,90 à 0,98, comparé à la méthode originale, du fait que les molécules hors de l'AD ne sont pas modifiées. L'indice global de cette variation est égal à 0,62, il est similaire à celui de la méthode originale (0,59).

Comme pour la génération de molécules par ajout et suppression de fragments, les meilleures performances sont obtenues quand les molécules sélectionnées pour être modifiées appartiennent au domaine d'applicabilité.

Indice	Valeur	Nombre
Unicité	0,68	5 887 ± 79
Nouveauté	0,94	5 521 ± 82
Prédictibilité	0,98	5 427 ± 80
Global	0,62	5 427 ± 80
Redécouverte	0,48	88 ± 4

Tableau 24 : Valeurs moyennes des indices *individuels* pour dix générations par la deuxième variation de croisement. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

### 4.3.3 Mutation de liaisons et d'atomes

La mutation d'une liaison ou d'un atome modifie le type d'une liaison ou d'un atome au sein d'une molécule. Il ne modifie pas la taille des molécules. Le choix du type de mutation appliqué – choix entre atome et liaison – n'est pas pondéré dans notre approche.

La mutation d'atomes ne s'applique que sur les atomes dits lourds, c'est-à-dire les atomes autres que celui d'hydrogène. Quand un atome est sélectionné, ses potentiels remplaçants sont listés en respectant la valence minimale nécessaire pour maintenir les liaisons auxquelles il participe avec des atomes lourds. Par exemple, un atome d'oxygène ne peut pas remplacer un atome de carbone impliqué dans 3 liaisons avec des atomes lourds. L'atome sélectionné est ensuite muté, remplacé par un des types d'atomes choisis au hasard parmi les remplaçants potentiels.

La mutation de liaisons est opérée de manière similaire. Une liaison entre deux atomes lourds est choisie au hasard dans la molécule. Pour cette liaison, on vérifie la faisabilité d'un changement du degré de la liaison sans dépasser la valence maximale des deux atomes participants à la liaison. Par exemple, une liaison carbone-carbone simple ne peut pas être remplacée par une liaison double ou triple si un des atomes de carbone est déjà impliqué dans trois autres liaisons covalentes avec d'autres atomes lourds. Les liaisons covalentes avec des atomes d'hydrogène ne sont pas prises en compte dans le calcul de la valence maximale, les

atomes d'hydrogène pouvant être supprimés. Si le changement du degré de la liaison est possible, il est réalisé.

Comme précédemment pour le croisement de graphes, la méthode de base utilise toutes les molécules générées à chaque itération tandis que sa variation n'utilise que les molécules générées appartenant au domaine d'applicabilité. Les mutations d'atomes et de liaisons permettent de générer plus de molécules que les opérations de croisement ou d'ajout et suppression de fragments ( $16\ 003 \pm 86$  molécules pour la méthode de base,  $11\ 045 \pm 86$  molécules pour la variation); et donc de se rapprocher davantage du nombre hypothétique de structures générées. Les molécules générées ont ensuite été analysées par les indices *individuels*, dont les valeurs moyennes sur les dix générations de cinq itérations sont présentées dans le Tableau 25. Nous observons que des fragments inconnus du jeu d'apprentissage du modèle QSPR ont tendance à apparaître plus aisément qu'avec les autres opérations, dus au fait que les mutations changent aléatoirement les caractéristiques des liaisons et des atomes. La prédictibilité s'en retrouve impactée comme le montre sa faible valeur (0,72) quand toutes les molécules sont sélectionnées comme parents. La prédictibilité est améliorée (valeur de 0,81) quand seules les molécules appartenant au domaine d'applicabilité sont sélectionnées comme parents. Les valeurs de l'indice global (0,57 et 0,62) restent similaires à celles obtenues avec les autres opérations de modification : ajout et suppression de fragments, croisement de graphes. Comme précédemment, la variation de méthode est choisie comme méthode de génération face à la méthode de base grâce à ses meilleures performances.

Indice	Base		Variation	
	V	#	V	#
Unicité	0,81	12 969 ± 93	0,79	8 745 ± 77
Nouveauté	0,97	12 575 ± 96	0,96	8 388 ± 79
Prédictibilité	0,72	9 088 ± 61	0,81	6 794 ± 72
Global	0,57	9 088 ± 61	0,62	6 794 ± 72
Redécouverte	0,38	71 ± 3	0,36	67 ± 4

Tableau 25 : Valeurs moyennes (V) des indices *individuels* pour dix générations par mutation d'atomes et de liaisons dans la méthode de base et sa variation. Le nombre moyen de molécules (#) associées à chaque indice est également indiqué avec son écart moyen.

#### 4.3.4 Cyclisation et ouverture de cycles

La cyclisation crée une liaison entre deux atomes de la molécule pour former un cycle. L'algorithme sélectionne aléatoirement un atome de départ lourd ne faisant pas partie d'un cycle. Tous les atomes lourds à une distance de 4 ou 5 liaisons de l'atome de départ sont listés, pour créer des cycles à 5 à 6 atomes. La possibilité de créer une liaison entre ces atomes en respectant leur valence est testée. Nous avons dans notre implémentation limité la cyclisation de manière à obtenir des cycles de cinq à six atomes, car ce sont les cycles les plus communément observés dans les molécules. Cette limite peut être modifiée dans l'implémentation de la méthode. Par ailleurs, il est important de préciser que si le jeu initial possède des molécules avec des cycles d'autres tailles – ce qui est le cas du jeu sur le point d'éclair –, ces derniers peuvent être modifiés par d'autres opérations comme la mutation de liaisons et d'atomes pour obtenir de nouveaux cycles. Enfin, le nombre de cycles n'est pas limité dans les molécules.

L'ouverture de cycle peut être réalisée sur les molécules contenant un ou plusieurs cycles. Cette méthode supprime l'une des liaisons intramoléculaires appartenant à un cycle en s'assurant de

ne pas générer une molécule fragmentée. Pour les molécules aromatiques, une de leurs formes mésomères est utilisée pour permettre d'ouvrir les cycles aromatiques.

Dans le jeu initial de données sur le point d'éclair, 182 des 599 molécules possèdent un ou plusieurs cycles. Au cours de la première itération de cyclisation et ouverture de cycle, une opération de cyclisation ou d'ouverture de cycle est aléatoirement appliquée à chaque graphe de ces 182 molécules. Une cyclisation est appliquée aux graphes des 417 autres molécules.

Comme précédemment, deux méthodes sont testées pour la sélection des molécules parentes. La méthode de base utilise toutes les molécules générées à chaque itération tandis que la variation n'utilise que les molécules générées appartenant au domaine d'applicabilité. En moyenne,  $10\,408 \pm 80$  molécules sont générées par l'opération de base et  $9\,797 \pm 66$  molécules par sa variation. Les molécules générées avec l'opération de base et sa variation ont été analysées par les indices *individuels*, dont les valeurs moyennes sur les dix générations de cinq itérations sont présentées dans le Tableau 26. Avec ou sans sélection préalable des molécules parentes, les valeurs des indices sont similaires, ce qui traduit que le taux de molécules générées appartenant au domaine d'applicabilité est similaire.

Indice	Base		Variation	
	V	#	V	#
Unicité	0,72	7 456 ± 88	0,71	6 916 ± 63
Nouveauté	0,95	7 116 ± 83	0,95	6 587 ± 67
Prédictibilité	0,95	6 743 ± 75	0,97	6 410 ± 64
Global	0,64	6 743 ± 75	0,65	6 410 ± 64
Redécouverte	0,23	43 ± 2	0,23	43 ± 3

Tableau 26 : Valeurs moyennes (V) des indices *individuels* pour dix générations par cyclisations et ouvertures de cycle. La méthode de base et sa variation sont étudiées. Le nombre moyen de molécules (#) associées à chaque indice est également indiqué avec son écart moyen.

### 4.3.5 Conclusions sur la génération par modification successives des structures

Nous avons étudié indépendamment l'impact de chacun des opérateurs que nous avons implémentés pour modifier des graphes moléculaires. L'utilisation de chaque opérateur sur cinq itérations permet d'obtenir entre  $8.10^3$  et  $16.10^3$  molécules alors que  $19.10^3$  molécules peuvent théoriquement être générées. Nous avons expliqué ce comportement par le fait que tous les opérateurs implémentés ne peuvent pas modifier toutes les molécules parentes. Une molécule de moins de cinq atomes ne peut par exemple pas subir de cyclisation. Parmi les molécules générées, environ 60% d'entre elles sont prédictibles. La majorité des autres molécules sont rejetées, car elles ont déjà été générées (indice d'unicité compris entre 0,68 et 0,81). La contrainte ne sélectionnant que les molécules appartenant à l'AD du modèle QSPR comme molécules parentes a permis d'augmenter légèrement les valeurs des indices de prédictibilité et global lors des générations. L'indice de redécouverte varie entre 0,23 et 0,48 avec l'usage des différents opérateurs. Les opérateurs permettant de redécouvrir le plus grand nombre de molécules sont ceux de croisement et d'ajout/suppression de fragments ; ce sont les opérateurs qui peuvent modifier le nombre d'atomes dans les molécules.

Ces opérations ont ainsi été regroupées pour fonctionner ensemble dans un script unique, illustré par la Figure 16. Ce script reprend le fonctionnement itératif des générations précédentes : tour à tour, une molécule parente est sélectionnée et modifiée. Plusieurs opérateurs peuvent maintenant être appliqués et leur choix est aléatoire. Si la génération échoue (structure incorrecte, déjà générée, ou génération d'une molécule hors de l'AD), un autre opérateur est sélectionné, et ce jusqu'à épuisement des opérateurs disponibles pour la molécule. Ce processus permet de remédier aux problèmes précédemment rencontrés, à savoir l'usage d'opérateurs qui ne sont pas adaptés à toutes les molécules, à la génération de molécules en doublon, et à la génération de molécules hors du domaine d'applicabilité. Nous nous assurons alors de n'obtenir que des molécules exploitables (uniques, nouvelles, et prédictibles par le modèle QSPR).

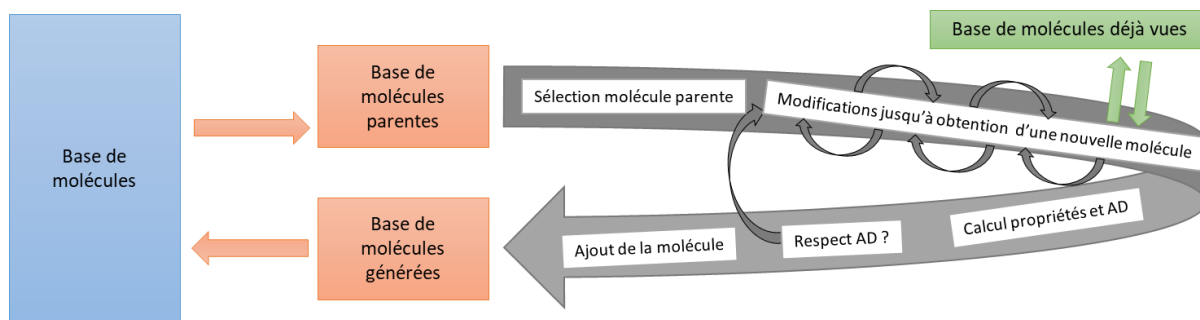


Figure 16 : Schéma de la méthode de génération par modifications successives de structures regroupant tous les opérateurs.

Deux méthodes ont été implémentées avec l'approche décrite par la Figure 16. La méthode « G1a » génère uniquement des molécules non cycliques. Les opérateurs d'ajout/suppression de fragments, de croisement et de mutation sont utilisés par G1a. La base de molécules parentes est définie comme l'ensemble des molécules non cycliques de notre jeu de données sur le point d'éclair (545 molécules parmi les 785 molécules du jeu). G1a a été implémentée comme méthode alternative aux méthodes F, qui ne considèrent pas non plus les molécules cycliques. La méthode « G1b » génère des molécules cycliques et non cycliques. Tous les opérateurs de modification sont utilisés. La base de molécules parentes y est définie comme l'ensemble des molécules de notre jeu de données sur le point d'éclair (785 molécules).

Une contrainte supplémentaire a été implémentée pour générer avec G1b des molécules possédant une propriété ciblée. La Figure 17 propose une schématisation de la procédure de génération avec la contrainte suivie sur la propriété. La procédure diffère uniquement de celle présentée sur la Figure 16 dans la constitution de la base de molécules parentes. Cette base était précédemment constituée de toutes les molécules du jeu initial ainsi que des molécules générées. Dorénavant, seules les X molécules générées ou initiales ayant une valeur de propriété la plus proche de la valeur cible peuvent être utilisées. Pour évaluer la proximité des valeurs de propriété à la valeur cible, les carrés des écarts entre la valeur de propriété ciblée et celle prédite pour chacune des molécules générées sont calculés et les molécules sont triées de façon croissante sur ce critère. Les X premières molécules (Top X) – proches de la valeur ciblée – sont utilisées comme « parentes » pour générer de nouvelles molécules à l'itération suivante. Le classement est rafraîchi après chaque génération. Cette génération est nommée « G2-X », où X est le nombre de molécules sélectionnées.

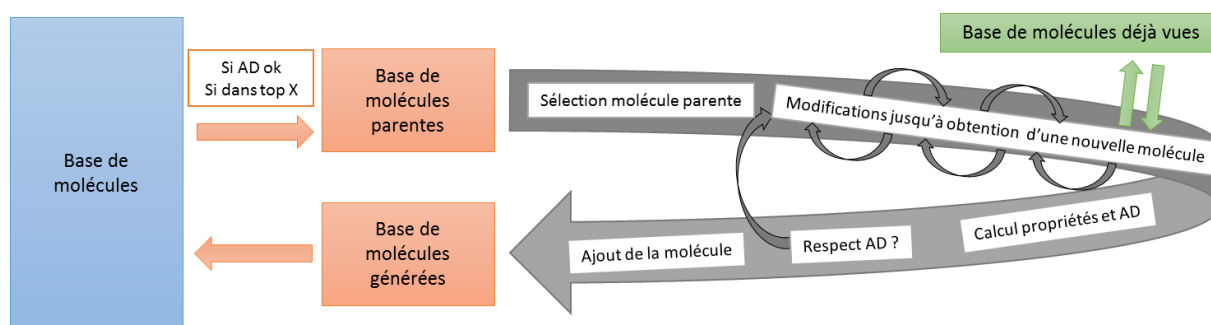


Figure 17 : Schéma de la méthode de génération par modifications successives de structures, avec contraintes sur la propriété.

## 4.4 Génération par apprentissage profond

Nous avons également souhaité expérimenter une méthode à base de réseaux de neurones, présentées dans la partie 3.4 « Génération de molécules par apprentissage profond », comme troisième méthode de génération. Nous avons dans cette partie tenté d'utiliser l'outil développé par Sattarov *et coll.*, à savoir un autoencodeur variationnel (VAE, pour « Variational Autoencoder ») utilisant la technique GTM, pour à la fois représenter l'espace latent et pour sélectionner les vecteurs à transcrire en molécules.<sup>122</sup>

La mise en œuvre d'un VAE nécessite un nombre de molécules conséquent, de l'ordre de  $10^4$  à  $10^6$ , particulièrement pour entraîner les réseaux de neurones au langage SMILES.<sup>34</sup> Or notre base de données expérimentales sur le point d'éclair ne contient pas assez de molécules pour entraîner de tels réseaux. Nous avons alors utilisé les  $5.10^6$  structures précédemment générées par G1b. Ces structures ont été préférées à celles générées par assemblage de fragments, pour entraîner le modèle avec des molécules moins ramifiées, davantage réalistes et pouvant comporter des cycles. De plus, afin de ne pas entraîner le VAE à générer des molécules difficilement synthétisables, les structures générées par G1b ont été filtrées selon leur indice d'accessibilité synthétique.<sup>23</sup> En ne conservant que les structures ayant un indice inférieur à 3, un nombre suffisant de structures pour entraîner une VAE et construire l'espace latent est obtenu (1 847 995 structures).

Nous avons testé deux approches pour extraire des vecteurs de cet espace latent : la première utilise une carte GTM,<sup>122</sup> la seconde reprend manuellement les vecteurs des molécules initiales



dans l'espace latent et les bruite. Les vecteurs extraits sont ensuite transformés en molécules par le décodeur.

#### **4.4.1 Génération de molécules à partir d'une carte GTM construite avec les vecteurs latents**

Dans cette sous-section, nous exploitons l'idée que l'espace latent, créé par le VAE pour représenter l'espace chimique des molécules d'entraînement, pourrait permettre de modéliser la propriété. Pour ce faire, les coordonnées des molécules dans l'espace latent en X dimensions sont utilisées comme valeurs de X descripteurs différents. Une carte GTM est construite à partir de ces descripteurs. Elle est ensuite utilisée comme une représentation 2D de l'espace latent et permet de sélectionner les zones de l'espace latent (à partir des combinaisons de descripteurs) dans lesquelles se trouvent des molécules initiales. Enfin, de nouvelles molécules sont générées par le décodeur du VAE qui échantillonne les zones de l'espace latent sélectionnées.

Après entraînement de notre VAE, l'espace latent comporte 18 dimensions. Les molécules du jeu initial d'apprentissage (599 molécules) ont été projetées dans l'espace latent. L'optimisation d'une carte GTM (voir section 2.1.2.3 « Représentation des données et modèles GTM ») a été lancée à partir des molécules du jeu initial encodées par les vecteurs de l'espace latent, associées à leur valeur de propriété expérimentale. La meilleure carte possède une valeur de  $R^2$  en validation interne égale à 0,57, alors que les mêmes molécules, encodées par les descripteurs ISIDA, avaient permis d'obtenir la carte GTM avec des valeurs de  $R^2$  égales à 0,834 (validation interne) ou 0,774 (validation externe) et le modèle QSPR avec des valeurs de  $R^2$  égale à 0,935 (validation interne) et 0,935 (validation externe)(voir partie 2.2 « Modèles construits au cours de la thèse »). Nous en avons déduit que l'espace latent ne permet pas de décrire aussi efficacement la propriété du point d'éclair par QSPR que les descripteurs initiaux, selon notre jeu initial de données. Ce comportement peut être dû au fait que l'espace latent a été construit avec un nombre plus important de structures que celles de notre base initiale, résultant à une projection très éparpillée des molécules initiales dans l'espace latent, et par conséquent à un espace ne pouvant pas être facilement représenté par GTM.

Nous n'avons alors pas pu échantillonner l'espace latent comme Sattarov *et coll.* pour générer de nouvelles molécules<sup>122</sup>, c'est-à-dire en sélectionnant des combinaisons de coordonnées dans les zones de l'espace latent prédites d'intérêt par GTM.

#### 4.4.2 Génération de molécules à partir des vecteurs latents bruités

Une méthode de génération alternative consiste à directement explorer les zones voisines de celles où sont localisées les molécules initiales, dans l'espace latent construit.<sup>121</sup> Itérativement, une molécule initiale est sélectionnée, ses coordonnées dans l'espace latent sont extraites, puis légèrement modifiées. Les coordonnées modifiées sont utilisées par le décodeur du VAE pour tenter de générer une nouvelle molécule. La modification des coordonnées s'effectue en leur appliquant un bruit ; la quantité de bruit influençant la dissimilarité des nouvelles molécules avec celles de départ.

Nous avons suivi cette méthode et défini le bruit comme un échantillon provenant d'une distribution normale gaussienne, centrée sur la valeur 1, avec un écart-type  $\sigma_G$ . Il est appliqué à un vecteur en multipliant la valeur de chacune de ses coordonnées par celle du bruit. Une valeur de bruit égale à 1 signifie un bruit nul, ne modifiant pas le vecteur, tandis que plus la valeur de bruit s'éloigne de 1 plus le vecteur est « bruité ». Les vecteurs bruités sont décodés, les chaînes de caractères produites sont vérifiées, et si elles correspondent à une chaîne SMILES valide, la molécule résultante est conservée comme résultat.

Nous avons étudié la relation entre la quantité de bruit appliqué et le nombre de molécules générées. Pour cela, nous avons sélectionné au hasard 10 molécules dans la base de données initiales, et ensuite bruité et décodé leurs vecteurs ( $v_1$  à  $v_{10}$ ) en faisant varier  $\sigma_G$ . 2 000 tentatives de transcription des vecteurs en molécules ont été effectuées pour chaque molécule et chaque valeur  $\sigma_G$ . La Figure 18 présente le nombre total de vecteurs décodés ayant mené à des molécules valides, nouvelles et uniques en fonction de la valeur  $\sigma_G$  pour chaque molécule initiale. Le nombre moyen de molécules nouvelles obtenues, quelle que soit la molécule initiale, est également présenté. Une valeur  $\sigma_G$  comprise entre 0,03 et 0,06 permet de générer en moyenne plus de 300 molécules nouvelles en 2000 tentatives. Jusqu'à 360 molécules nouvelles ont pu être générées en moyenne quand  $\sigma_G$  prend la valeur 0,04. Nous avons alors fixé la valeur  $\sigma_G$  à 0,04. Les vecteurs latents  $v_8$  et  $v_9$  permettent d'obtenir un nombre de molécules supérieur à la moyenne. Ils correspondent à des molécules linéaires oxygénées comportant des doubles liaisons (de notation SMILES CCOC(=O)CCCC=C et CCCCCCCCCCCCCCCC(=O)OC), par conséquent le nombre de modifications réalisables pour obtenir une nouvelle structure, par application d'un bruit, y est élevé. Au contraire et par exemple, les vecteurs  $v_3$  et  $v_{10}$  correspondent à de petites molécules (de notation SMILES COC(C)=O et OCCCCO), pour

lesquelles le nombre de modifications réalisables est moins élevé, notamment à cause de leur taille.

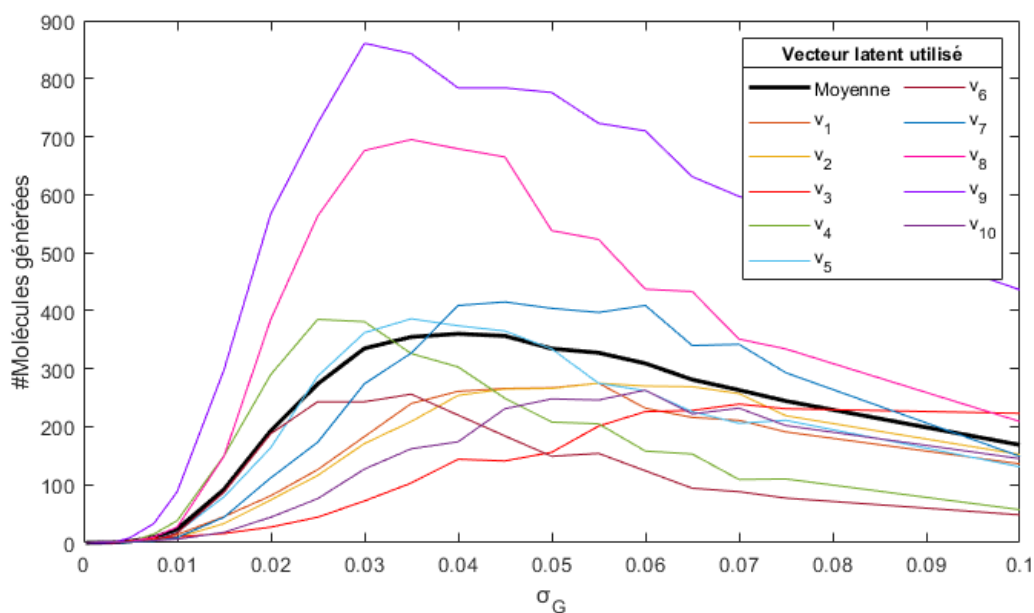


Figure 18 : Nombre de molécules valides, nouvelles et uniques générées après 2 000 tentatives pour chacun des 10 vecteurs latents, ainsi qu'en moyenne, en fonction de la valeur de  $\sigma_G$ .

Le VAE est à présent paramétré pour générer de nouvelles molécules à partir des 785 molécules de la base sur le point d'éclair. Pour augmenter encore plus la diversité des molécules générées, nous avons souhaité utiliser non seulement les vecteurs des molécules de la base, mais également ceux des molécules déjà générées comme vecteurs à bruitez et décoder. Un processus itératif, inspiré de notre méthode de génération par modifications successives, a été mis en place. À chaque itération, les vecteurs des molécules initiales et précédemment générées sont bruités pour générer de nouvelles structures. Jusqu'à 2 000 tentatives de bruitage, puis décodage sont effectuées par molécule jusqu'à obtenir une structure valide. Les générations sont effectuées dix fois sur quatre itérations. Le Tableau 27 présente les performances de génération moyennes sur les dix générations à travers les valeurs d'indices *individuels*.

<b>Indice</b>	<b>Nombre</b>	<b>Valeur</b>
Unicité	8 021 ± 47	0,80
Nouveauté	3 054 ± 97	0,38
Prédictibilité	886 ± 23	0,29
Global	886 ± 23	0,09

Tableau 27 : Valeurs moyennes des indices *individuels* pour dix générations de quatre itérations par la méthode VAE. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.

Le VAE produit à partir des molécules initiales environ  $10\,004 \pm 34$  structures en quatre itérations. La différence entre le nombre hypothétique de structures générées et le nombre de structures effectivement générées est due à certaines molécules qui ne permettent pas de produire une structure à partir de leur vecteur bruité, malgré les 2 000 tentatives effectuées par molécule. Quatre-vingts pour cent des molécules générées sont uniques – produites une seule fois. La très grande partie de ces structures sont déjà connues du VAE, pénalisant le score de nouveauté – les molécules connues du VAE sont non seulement les molécules initiales de notre jeu de données sur le point d'éclair, mais également les 1 847 995 molécules générées par G1. Près des deux tiers des molécules générées ne sont pas prédictibles par notre modèle QSPR, car elles contiennent des fragments ISIDA non présents dans le jeu d'apprentissage ou dont le nombre d'occurrences dépasse les valeurs observées dans le jeu d'apprentissage. Par conséquent, le score global de notre VAE est faible (0,09).

#### **4.4.3 Conclusion sur la génération par apprentissage profond**

Dans cette section, nous avons considéré la génération moléculaire à l'aide d'un VAE pour compléter notre BDD sur le PE, en utilisant l'outil développé par Sattarov et coll.<sup>122</sup> Nous avons utilisé un ensemble de 1 847 995 molécules générées par la méthode G1b et filtrées pour entraîner l'encodeur, le décodeur et créer l'espace latent. La quantité de bruit appliquée aux

vecteurs initiaux a été optimisée pour augmenter le nombre de molécules générées. La génération de nouvelles molécules à partir des molécules du jeu initial et des molécules précédemment générées a ensuite été étudiée dans un processus itératif. Le VAE a généré une majorité de molécules non exploitables, car déjà connues ou non prédictibles par le modèle QSPR. Nous manquons encore de recul pour pouvoir maîtriser correctement l’outil développé par Sattarov et coll.<sup>122</sup> et générer davantage de structures satisfaisantes; le temps limité de la thèse ne nous ayant pas permis de travailler davantage avec cet outil.

Toutefois, nous avons alors démontré qu’il est possible d’utiliser un VAE avec de petites bases de données en s’aidant préalablement d’une autre méthode de génération. D’autres approches d’augmentation de données existent dans la littérature, telles que l’énumération des SMILES non canoniques.<sup>133</sup> Nous considérons notre approche – utiliser des structures préalablement générées – comme une alternative aux techniques d’énumération.

## **4.5 Conclusions sur la génération par apprentissage profond**

Dans ce chapitre, nous avons étudié trois approches de génération moléculaire pour compléter une base de données pouvant contenir peu de molécules. La base sur le point d’éclair a servi de base d’exemple pour l’application des méthodes.

Les méthodes F concatènent des fragments en suivant les règles de chimie. Deux types de fragments, issus de considérations sur les molécules de notre base de données, ont été utilisés : des fragments « simples » d’un à deux atomes et les fragments de deux à quatre atomes encodés par les descripteurs du modèle QSPR. La mise en place de contraintes sur le choix du degré des liaisons à créer entre fragments, et sur le choix des fragments à assembler, a permis de rendre cette méthode plus efficace. Les méthodes G effectuent itérativement des opérations sur les graphes des molécules afin de créer de nouvelles structures. Nous avons implémenté chaque opération de modification et observé que si nous autorisons la possibilité d’être modifiées qu’aux molécules prédictibles, nous améliorons légèrement le score de prédictibilité. Nous avons également mis en place une contrainte sur la construction de la base de molécules parentes quand une valeur de propriété est ciblée. La troisième méthode de génération E est basée sur un autoencodeur variationnel. Il génère des molécules à partir de vecteurs, représentant les molécules dans un espace à haute dimension, qui sont bruités et décodés. Les molécules générées par modifications successives de structures ont été utilisées pour entraîner

la méthode E. Même si le nombre de molécules générées en doublons ou hors de l'AD de notre modèle QSPR est plus important qu'avec les autres méthodes, des molécules prédictibles ont été produites. Nous avons alors démontré qu'il est possible d'adapter de manière originale un VAE pour générer des molécules à partir de petites bases de données.

## Chapitre 5. Comparaison des générations moléculaires

---

Dans le chapitre précédent, trois méthodes de génération moléculaire ont été décrites et améliorées. Nous proposons dans ce chapitre de comparer les méthodes de génération par assemblages de fragments (F) et par modifications successives (G), car ce sont des méthodes qui peuvent être utilisées directement avec les molécules initiales – sans générer préalablement d'autres structures. Le chapitre est structuré en quatre parties. La première partie compare les méthodes de génération à la manière des Plateformes d'Analyses Comparatives (PAC, qui sont détaillées dans la section 3.5 « Evaluation des méthodes de génération »). La seconde partie détaille le développement d'un nouvel outil pour la comparaison des méthodes de génération. Dans la troisième partie, ce nouvel outil est appliqué pour comparer les méthodes de génération. La dernière partie expose nos conclusions sur la comparaison des méthodes F et G. Les travaux présentés dans ce chapitre ont fait l'objet d'une publication.<sup>94</sup>

### 5.1 Données et comparaisons préliminaires

#### 5.1.1 Méthodes de génération comparées

Dans ce chapitre, nous comparons deux des méthodes de génération présentées précédemment, ainsi que leurs variations, à savoir : la génération par assemblages de fragments (F et ses variations F0, F1a, F1b) et la génération par modifications successives de structures (G et ses variations G1a, G1b, G2-X). Nous avons généré jusqu'à cinquante mille (50k) puis cinq millions (5M) de structures par chacune de ces deux méthodes.

Avec la génération par assemblages de fragments, la méthode produit des molécules séquentiellement, sans considérer les molécules précédentes autrement que pour vérifier les éventuels doublons. De ce fait, 5M de molécules peuvent être générées en une seule exécution.

Avec la génération par modifications successives de structures, une seule exécution ne permet pas de produire 5M de molécules. En effet, avec G1b, des ressources en mémoire vive trop importantes sont nécessaires à partir d'une dizaine d'itérations pour manipuler la base de

parents (qui comporte alors près de  $10^6$  molécules). D'autre part, avec G2-X, particulièrement quand le nombre de parents X est faible, les générations ne produisent plus de nouvelles molécules après quelques milliers d'itérations (après environ  $10^4$  à  $10^5$  molécules générées quand X=50). Nous expliquons ce comportement du fait que la base de molécules parentes est trop restreinte et donc toutes les modifications pouvant être réalisées sur les molécules parentes sont épuisées. Nous avons alors lancé chaque variation de G plusieurs fois pour à la fois ne pas dépasser les capacités de mémoire et pour rafraîchir manuellement la base de molécules parentes. Les molécules générées dans chaque lancement sont concaténées jusqu'à obtenir 5M de molécules uniques.

### 5.1.2 Comparaisons des distributions de propriétés

Comme mentionné dans le Chapitre 3 « Méthodes pour la génération moléculaire et pour leur comparaison », il est d'usage d'examiner les molécules générées à l'aide d'indices *collectifs*. Les indices *collectifs* évaluent différents ensembles de molécules en comparant leurs caractéristiques collectives, comme leurs distributions de descripteurs ou de propriétés à celles de molécules de référence – ayant servi à l'entraînement des méthodes de générations. Ces indices vérifient que la diversité des molécules générées est équivalente à celle des molécules de référence. De ce fait, plus les distributions des propriétés des molécules générées sont similaires à celles des molécules de référence, plus la génération est considérée comme efficace. Dans cette sous-section, nous avons testé cette approche avec les générations F1b et G1a, produisant toutes deux des molécules sans cycles et sans contrainte sur la propriété. Nous avons considéré les propriétés de poids moléculaire, de point d'éclair (PE, prédit par le modèle QSPR construit dans la partie 2.2.2.2 « Modélisation par SVR de la propriété de point d'éclair ») et d'accessibilité synthétique (AS, prédite selon l'outil développé par Ertl *et coll.*<sup>23</sup>). Le poids moléculaire informe sur la taille de la molécule, des corrélations entre le poids moléculaire et le PE ont démontré que des molécules lourdes (de grande taille) ont tendance à posséder une valeur de PE plus haute que celle des molécules plus légères (de petite taille)<sup>91,134,135</sup>. L'AS note les composés de 1 (facile à synthétiser) à 10 (très compliqué à synthétiser). Dans notre cas, l'AS représente une grandeur intéressante puisque nous visons à proposer des composés pouvant être synthétisés.

La Figure 19 présente la distribution de ces trois caractéristiques sur les différents jeux de molécules, initiales et générées, pour les 50k et les 5M premières molécules. Les distributions



de propriétés des molécules générées sont décalées à droite (composés plus lourds, complexes et par conséquent, plus difficilement synthétisables) comparées aux distributions au sein du jeu initial. Ce décalage s'accroît avec le nombre de molécules générées, comme observé pour 50k et 5M de molécules générées. Nous expliquons ce phénomène par le fait que l'espace chimique des molécules de petite taille est plus restreint (moins de combinaisons d'atomes possibles) que celui de molécules de plus grande taille. Au fur et à mesure de l'avancement des générations, il devient difficile d'explorer davantage l'espace chimique des molécules de petite taille. À 5M de molécules, la méthode F1b génère plutôt des molécules avec un poids moléculaire supérieur à 100g/mol et une AS supérieure à 2,5. La méthode G1a produit en général des molécules avec un poids moléculaire compris dans le même intervalle que celui des molécules produites par F1b, mais avec une AS moins élevée. La distribution des valeurs de point d'éclair des molécules générées par G1a est légèrement plus étendue que celle des molécules générées par F1b.

L'emploi d'indices *collectifs* comparant les distributions de propriétés entre molécules initiales et générées ne semble alors pas adéquat pour évaluer les performances des générations, notre base de données initiales n'étant pas représentative de toute la diversité chimique des molécules pouvant être générées et appartenir à l'AD du modèle QSPR. La diversité des molécules augmente au fur et à mesure de l'avancement de nos générations. De plus, dans une optique d'automatisation des comparaisons, le choix des propriétés à comparer en fonction de l'application visée n'est pas évident, et il peut difficilement être automatisé.

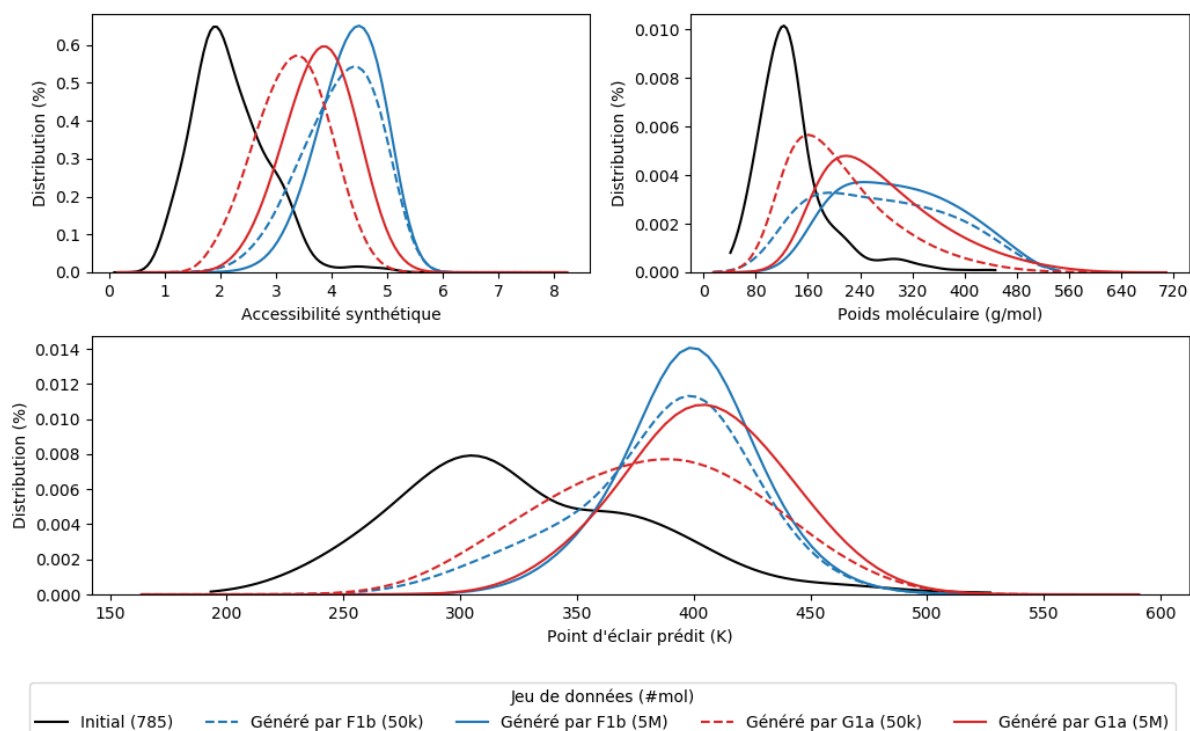


Figure 19 : Distribution des valeurs d'Accessibilité Synthétique (AS), de poids moléculaire, et de point d'éclair prédit dans les jeux de molécules initiales (du modèle QSPR) ainsi que générées par F1b et G1a, exprimé en pourcentage de molécules de chaque jeu.

## 5.2 Indices pour la comparaison des générations

Dans cette sous-section, nous investiguons la comparaison des générations de manière alternative à la comparaison des distributions de propriétés. Nous avons repris de la section 2.2.3 « Représentation des données par ACP » la représentation du jeu de données initial, construite sur la base des trois premières composantes principales issues d'une ACP, où les valeurs de descripteurs des molécules sont standardisées. L'espace étendu  $\mathbb{C}$ , dérivé de cette représentation, a été utilisé pour essayer d'évaluer les molécules générées.

### 5.2.1 Utilisation de l'espace $\mathbb{C}$ pour la comparaison moléculaire

Les projections des molécules générées par G1a puis par F1b dans l'espace  $\mathbb{C}$  sont présentées sur la Figure 20, pour évaluer la faisabilité de comparaisons à partir de  $\mathbb{C}$ . Nous remarquons que

les projections des molécules produites par G1a et F1b occupent différemment  $\mathbb{C}$ . La Figure 20 ne permet pas d'autres analyses.

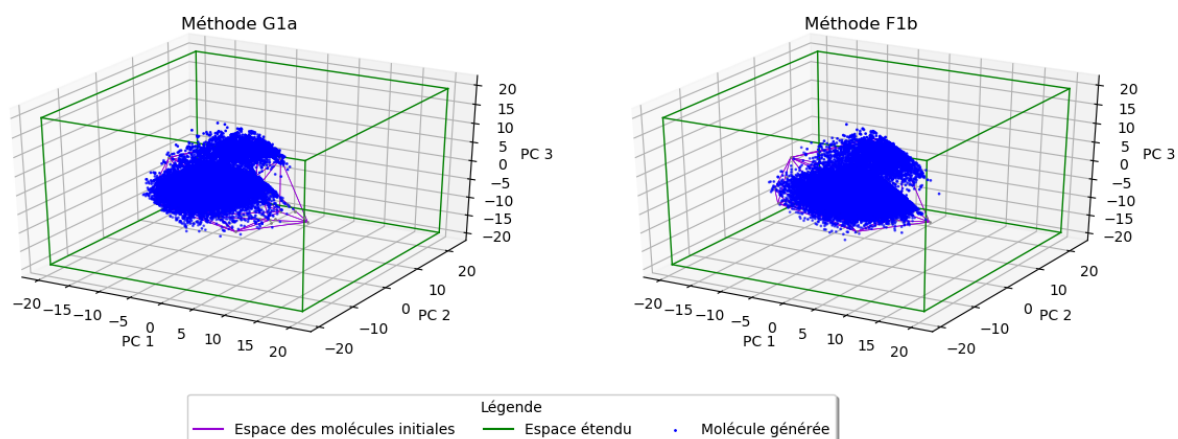


Figure 20 : Projection des 50k premières molécules générées dans l'espace  $\mathbb{C}$ , pour la méthode G1a (à gauche) et F1b (à droite).

Nous avons alors discrétisé  $\mathbb{C}$  en cubes de taille fixe le long des trois composantes principales, pour analyser la densité d'occupation de  $\mathbb{C}$  à travers les cubes. Chaque molécule générée est alors projetée dans un cube distinct de  $\mathbb{C}$ . Plusieurs valeurs ont été testées pour la taille des cubes et les résultats obtenus sont présentés dans l'Annexe D. Le meilleur compromis a été obtenu avec une taille cubique de 1 unité.  $\mathbb{C}$  est alors discrétisé à l'aide de 68 757 cubes. La Figure 21 présente l'espace  $\mathbb{C}$  et les cubes dans lesquels est projetée au moins une molécule générée, par la méthode G1a (à gauche) et F1b (à droite). Chaque cube est représenté en son centre de gravité par une sphère dont la taille et la couleur sont fonction du nombre de molécules projetées dans le cube. Avec G1a, le nombre maximal de molécules par cube avoisine les 250 tandis que ce nombre est proche de 350 pour F1b. Le nombre de cubes possédant plus de 150 molécules semble plus élevé pour G1a que pour F1b. Une analyse des cubes a été menée ensuite, montrant que les 50k molécules générées par G1a sont projetées dans 3 884 cubes (sur 68 757) tandis que celles générées par F1b le sont dans 4 211 cubes. G1a semble alors produire des molécules projetées dans moins de cubes que F1b, mais de manière plus uniforme.

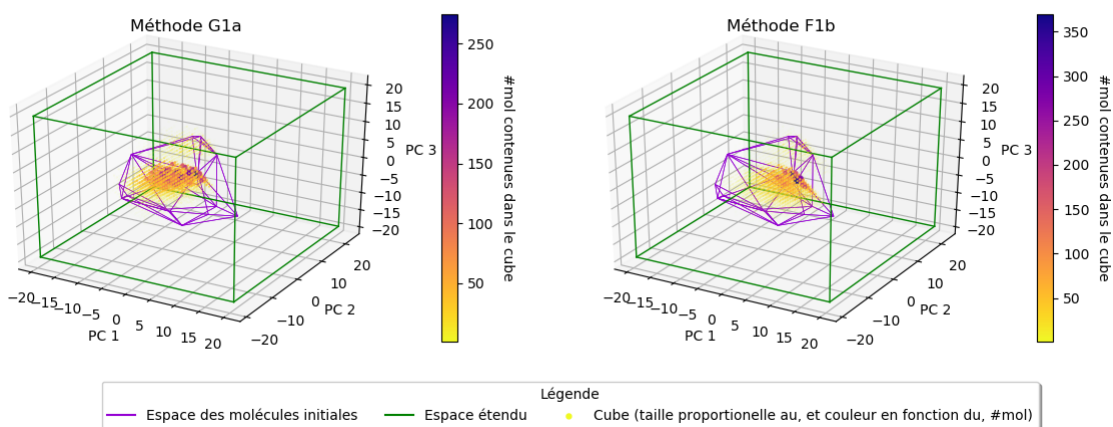


Figure 21 : Représentation de la densité spatiale des 10 000 premières molécules générées par la méthode G1a (à gauche) et F1b (à droite) dans l'espace  $\mathbb{C}$ , à l'aide de cubes de taille unitaire.

Dans un troisième temps, nous avons projeté simultanément les molécules générées par G1a et F1b dans l'espace  $\mathbb{C}$  discrétisé. Le ratio (R) entre les molécules projetées issues de G1a et de F1b a été calculé dans chaque cube. Nous avons ensuite étudié les projections des molécules générées sur trois plans de cet espace. Chaque représentation présente tour à tour deux des trois dimensions de  $\mathbb{C}$ , faisant la moyenne des ratios R dans les cubes le long de la troisième dimension. La Figure 22 présente ces trois représentations, sur lesquelles la moyenne des ratios R le long de la troisième dimension est présentée par un code couleur. Sur la Figure 22, nous observons que G1a produit plutôt des molécules projetées à des valeurs négatives de l'axe PC 1 tandis que F1b génère des molécules projetées davantage à des valeurs positives sur le même axe (représentations 18a et 18b). La représentation 18c ne permet pas d'observer d'autres différences.

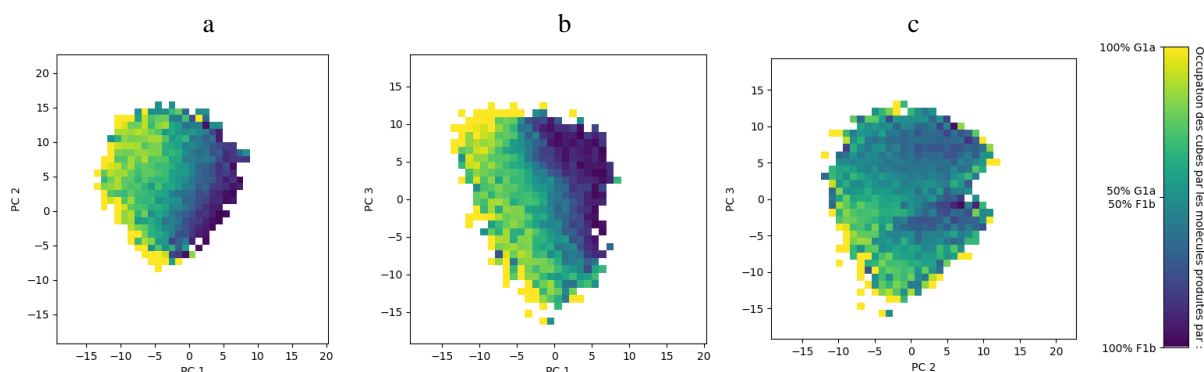


Figure 22 : Moyennes des ratios des molécules générées par G1a et par F1b dans chaque ensemble de cubes le long de la composante principale (PC) 3 (diagramme a), PC 2 (b) et PC 1 (c).

La représentation  $\mathbb{C}$  semble adéquate pour comparer les méthodes de génération. De nouveaux indices *collectifs* sont développés dans la suite de cette section pour quantifier les différences de peuplement de  $\mathbb{C}$  par les molécules générées. Ci-dessous, l'ensemble des méthodes comparées est noté  $M$ , et chaque méthode de  $M$  est notée  $m$ . Comme observé dans la partie 5.1.2 « Comparaisons des distributions de propriétés », notre base initiale – contenant les molécules utilisées pour construire le modèle QSPR – ne reproduit pas toute la diversité chimique des molécules pouvant être générées et appartenir à l'AD du modèle QSPR. Par conséquent, il est nécessaire d'utiliser une autre référence pour les comparaisons. La référence est définie dans notre travail comme l'ensemble des molécules générées par toutes les méthodes de  $M$ . Cette référence est dynamiquement définie au fur et à mesure des générations, ce qui permet de la personnaliser selon l'avancement des générations : il n'est pas nécessaire de générer toutes les molécules pouvant appartenir à l'AD préalablement à la comparaison des générations. Il est cependant nécessaire d'inclure le même nombre de molécules produites par chaque méthode de génération dans la référence pour ne pas biaiser la diversité chimique de cette dernière. Plus le nombre de molécules générées est important, plus nous pouvons considérer la référence comme représentative. L'usage de cette référence permet de faire abstraction des zones de l'espace chimique non explorables ou non explorées (c.-à-d. des cubes vides) lors des générations, que ce soit du fait de la définition même de l'AD (par contrôle des fragments) ou de la violation de règles chimiques (comme la règle de l'octet).

### 5.2.2 Couverture de l'espace

La couverture de l'espace représente le pourcentage de l'espace occupé par les molécules générées par une méthode  $m$ , par rapport à l'occupation de l'espace par les molécules générées avec toutes les méthodes comparées ( $M$ ). L'indice  $I_1$  est défini comme le ratio entre  $N_{cubes}^{\mathbb{C},m}$ , le nombre de cubes de  $\mathbb{C}$  occupés par les molécules générées par  $m$ , et  $N_{cubes}^{\mathbb{C},M}$ , le nombre de cubes occupés par toutes les molécules générées (Equation (13)). La valeur de  $I_1$  est comprise dans l'intervalle ]0 ; 1]. Plus sa valeur est élevée, plus la couverture de l'espace par la méthode considérée est importante.

$$I_1 = \frac{N_{cubes}^{\mathbb{C},m}}{N_{cubes}^{\mathbb{C},M}}, m \in M \quad (13)$$

Où  $M$  est l'ensemble des méthodes comparées.

### 5.2.3 Représentativité de la couverture de l'espace

La représentativité de l'espace définit pour une méthode  $m$  la similarité de projection dans  $\mathbb{C}$  des molécules produites par  $m$  par rapport à toutes les molécules générées. Deux taux d'occupation sont définis et utilisés ensuite par les indices de représentativité.  $P_x^m$  représente le taux d'occupation du cube  $x$  pour la méthode  $m$ , il est défini comme le ratio entre  $N_{structures}^{x,m}$ , le nombre de structures générées par  $m$  dans le cube  $x$ , et  $N_{structures}^{\mathbb{C},m}$ , le nombre total de structures générées par  $m$  dans  $\mathbb{C}$  (Equation (14)).  $P_x^M$  désigne le taux d'occupation global du cube  $x$ , il est défini comme la moyenne des taux d'occupation du cube  $x$  par les différentes méthodes  $m \in M$  (Equation (15)).

$$P_x^m = \frac{N_{structures}^{x,m}}{N_{structures}^{\mathbb{C},m}}, x \in X \quad (14)$$

$$P_x^M = \frac{\sum_{m \in M} P_x^m}{M}, m \in M \quad (15)$$

Où  $X$  est l'ensemble des cubes dans  $\mathbb{C}$ .

Plusieurs indices ont été établis pour évaluer la représentativité de l'espace à partir des taux d'occupation individuels et globaux.  $I_{2a}$  est défini comme la différence entre le chiffre 1 et la distance de variation totale (Equation (16)). La distance de variation totale est la plus grande différence observée entre  $P_x^m$  et  $P_x^M$  pour un cube  $x$  parmi  $X$ , les cubes de  $\mathbb{C}$ .

$$I_{2a} = 1 - \max_{x \in X} |P_x^m - P_x^M| \quad (16)$$

$I_{2b}$  est défini comme la différence entre le chiffre 1 et la distance euclidienne au carré des taux d'occupation (Equation (17)). La distance euclidienne au carré des taux d'occupation est la somme sur tous les cubes  $x$  de  $X$  des différences au carré entre  $P_x^m$  et  $P_x^M$ .

$$I_{2b} = 1 - \sum_{x \in X} (P_x^m - P_x^M)^2 \quad (17)$$

$I_{2c}$  est défini comme la différence entre le chiffre 1 et la  $N_m^{\text{ième}}$  racine du rapport entre la distance euclidienne au carré des taux d'occupation et la somme des valeurs  $P_x^M$  au carré, pondéré par l'inverse du nombre de cubes occupés par des molécules générées par  $m$ ,  $N_{cubes}^{\mathbb{C},m}$ , où  $N_m$  est le nombre de méthodes comparées (Equation (18)). La pondération par  $1/N_{cubes}^{\mathbb{C},m}$

permet de normaliser les valeurs  $I_{2c}$  dans l'intervalle  $[0 ; 1]$ . L'usage de la  $N_m$ <sup>ième</sup> racine évite que les valeurs de l'indice convergent trop rapidement vers 1.

$$I_{2c} = 1 - \sqrt[N_m]{\frac{\sum_{x \in X} (P_x^m - P_x^M)^2}{\sum_{x \in X} (P_x^T)^2} * \frac{1}{N_{cubes}^{C,m}}} \quad (18)$$

Où  $N_m$  le nombre de méthodes comparées.

$I_{2d}$  est défini comme la différence entre le chiffre 1 et la distance d'Hellinger<sup>136,137</sup> entre  $P_x^m$  et  $P_x^M$  (Equation (19)). La distance d'Hellinger correspond à la racine carrée de la somme sur tous les cubes des carrés de la différence entre les racines carrées de  $P_x^m$  et de  $P_x^M$  ; pondérée par le coefficient  $1/\sqrt{2}$  pour normaliser la distance dans l'intervalle  $[0 ; 1]$ . Cette distance utilise des différences de racines carrées, ce qui la rend moins sensible à des grands écarts que d'autres mesures telles que  $I_{2a}$ ,  $I_{2b}$  ou  $I_{2c}$ .

$$I_{2d} = 1 - \left( \frac{1}{\sqrt{2}} * \sqrt{\sum_{x \in X} (\sqrt{P_x^m} - \sqrt{P_x^M})^2} \right) \quad (19)$$

Les valeurs de ces indices sont également incluses dans l'intervalle  $[0 ; 1]$ . Des valeurs de représentativité proches de 1 indiquent une similarité importante entre l'occupation de  $\mathbb{C}$  par les molécules de la méthode comparée et celle par les molécules des autres méthodes. Au début des générations, ces indices sont supposés prendre des valeurs faibles, chaque méthode générant probablement des molécules projetées dans des cubes distincts : si le nombre de cubes dans  $\mathbb{C}$  est élevé, la probabilité de générer une molécule projetée dans le même cube est faible. Quand des molécules générées par des méthodes différentes sont projetées dans les mêmes cubes, les valeurs des indices commencent à augmenter.

## 5.2.4 Uniformité de la couverture de l'espace

L'uniformité de la couverture de l'espace quantifie la similarité entre la projection de molécules générées avec une méthode  $m$  et la projection de molécules générées avec une méthode hypothétique  $h$ . La méthode  $h$  produit des molécules projetées équitablement dans tous les cubes utilisés par  $m$ . Le taux d'occupation associé à la méthode  $h$ ,  $P_x^{h,m}$  est défini comme le

ratio entre  $N_{structures}^{\mathbb{C},m}$ , le nombre total de structures générées par  $m$  dans  $\mathbb{C}$ , et  $N_{cubes}^{\mathbb{C},m}$ , le nombre de cubes de  $\mathbb{C}$  occupés par les molécules générées par  $m$  (Equation (20)).

$$P_x^{h,m} = \frac{N_{structures}^{\mathbb{C},m}}{N_{cubes}^{\mathbb{C},m}}, m \in M \quad (20)$$

L'indice  $I_3$  est défini comme la différence entre le chiffre 1 et la distance d'Hellinger entre  $P_x^m$  et  $P_x^{h,m}$  (Equation (21)). Les valeurs de  $I_3$  sont comprises dans l'intervalle  $[0 ; 1]$ . Plus la génération  $m$  est similaire à la génération  $h$ , plus les valeurs  $I_3$  sont proches de 1. Au début des générations,  $I_3$  est supposé prendre des valeurs élevées : les générations commencent par produire des molécules dans des cubes inoccupés, avec un taux d'occupation alors uniforme et égal à une molécule par cube. Au fur et à mesure de l'avancement des générations, selon les restrictions chimiques imposées par l'AD et selon la manière des méthodes de générer des molécules, ces valeurs vont avoir tendance à décroître plus ou moins rapidement.

$$I_3 = 1 - \left( \frac{1}{\sqrt{2}} * \sqrt{\sum_{x=0}^n (\sqrt{P_x^m} - \sqrt{P_x^{h,m}})^2} \right) \quad (21)$$

### 5.2.5 Spécificité de la génération

La spécificité de la génération évalue le pourcentage de molécules générées respectant une contrainte sur la valeur de leur propriété. La contrainte est définie par l'équation  $|p - T| \leq t$ , où  $p$  est la valeur de propriété de la molécule,  $T$  la valeur cible, et  $t$  la tolérance admise. L'indice individuel  $I_4$  évalue la spécificité d'une génération et est défini comme le rapport entre  $N_{structures}^m, |p - T| \leq t$ , le nombre de molécules générées par  $m$  respectant la contrainte, et  $N_{structures}^m$ , le nombre total de structures générées par  $m$  (Equation (22)). Les valeurs de l'indice  $I_4$  sont incluses dans l'intervalle  $[0 ; 1]$ , des valeurs proches de 1 indiquent que la méthode génère une majorité de molécules respectant la contrainte sur la valeur de propriété.

$$I_4 = \frac{N_{structures}^m, |p - T| \leq t}{N_{structures}^m} \quad (22)$$



## 5.3 Comparaison des méthodes de génération à l'aide des nouveaux indices

Dans cette partie, nous analysons les performances de génération des méthodes à l'aide des indices *collectifs* définis dans la section précédente. Les méthodes F0, F1a, F1b, G1a et G1b évoquées dans la section 5.1.1 « Méthodes de génération comparées » sont d'abord utilisées et comparées pour générer des molécules sans contrainte sur la propriété. Nous étudions ensuite la génération de molécules possédant leur point d'éclair dans des intervalles définis.

### 5.3.1 Génération de molécules diverses

Cette section est consacrée à la comparaison des méthodes de génération, sans considérer de contrainte sur une valeur de propriété. Nous avons généré jusqu'à 5M molécules avec chacune des méthodes F0, F1a, F1b, G1a et G1b. La Figure 23 présente les évolutions dynamiques (avec le nombre de molécules générées) des indices  $I_1$  à  $I_3$  et le Tableau 28 présente leur valeur à 5M molécules générées.

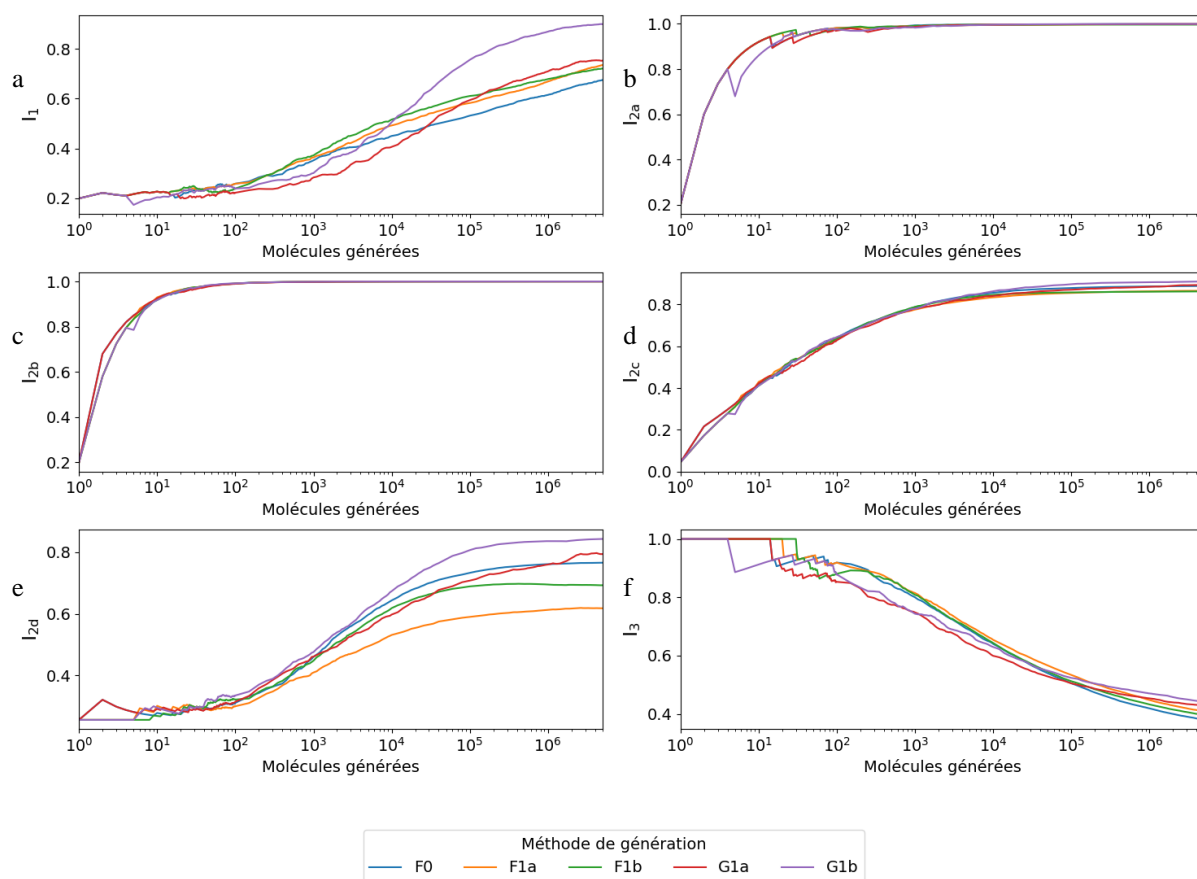


Figure 23 : Évolutions des indices (a)  $I_1$ , (b)  $I_{2a}$ , (c)  $I_{2b}$ , (d)  $I_{2c}$ , (e)  $I_{2d}$  et (f)  $I_3$  avec le nombre de molécules générées par les méthodes F0, F1a, F1b, G1a et G1b. Adapté avec permission, issu de Gantzer *et coll.*<sup>94</sup> Copyright 2021 American Chemical Society.

Indice	Valeur d'indice pour la génération par la méthode				
	F0	F1a	F1b	G1a	G1b
$I_1$	0,68	0,74	0,72	0,75	0,90
$I_{2a}$	0,99	0,99	0,99	0,99	0,99
$I_{2b}$	0,99	0,99	0,99	0,99	0,99
$I_{2c}$	0,89	0,87	0,86	0,89	0,91
$I_{2d}$	0,77	0,62	0,69	0,79	0,84
$I_3$	0,38	0,41	0,40	0,43	0,44

Tableau 28 : Valeurs des indices  $I_1$  à  $I_3$  à 5M molécules générées par les méthodes F0, F1a, F1b, G1a et G1b.

La couverture de l'espace est d'abord analysée (indice  $I_1$ ) sur la Figure 23, diagramme (a). Au début des générations et jusqu'à environ une centaine de molécules générées, l'indice  $I_1$  possède une valeur proche de 0,2 : chacune des cinq méthodes génère des molécules projetées dans des cubes différents, menant à une valeur d'indice proche de  $1/N_m$  où  $N_m=5$  est le nombre de méthodes comparées. Les valeurs augmentent à partir d'une centaine de molécules générées, les différentes méthodes commençant à produire des molécules localisées dans les mêmes cubes. Jusqu'à  $3 \cdot 10^3$  molécules générées, les méthodes F démontrent les meilleures couvertures. Parmi ces méthodes, F1a et F1b explorent davantage  $\mathbb{C}$  que F0 du fait qu'elles ont davantage de fragments à disposition pour construire des structures. Entre environ  $3 \cdot 10^3$  et  $2 \cdot 10^5$  molécules générées, les couvertures de l'espace des méthodes G dépassent celles des méthodes F. Nous faisons l'hypothèse que ce comportement est dû au fait qu'au début des générations, les méthodes G sont restreintes par la diversité présente dans le jeu initial. Ces méthodes auraient alors besoin d'effectuer plusieurs itérations pour posséder des molécules davantage diverses et dépasser la couverture des méthodes F. À la fin des générations, l'indice  $I_1$  classe les méthodes comme suit :  $G1b > G1a \approx F1a \approx F1b > F0$ . Les méthodes G sont aussi (pour G1a) ou plus (pour G1b) efficaces que les méthodes F à partir de  $3 \cdot 10^3$  molécules générées pour obtenir des molécules chimiquement diverses, selon la couverture de l'espace procurée.

Les évolutions des indices de représentativité de l'espace (indices  $I_{2a}$  à  $I_{2d}$ ) en fonction du nombre de structures générées sont représentées sur les Figure 23, diagrammes (b) à (e). Au début des générations, les valeurs de tous ces indices sont faibles. Cela est dû au fait que chaque méthode produit d'abord des molécules dans des cubes différents (voir la discussion sur les variations de  $I_1$ ). Par conséquent, un écart important existe entre les distributions dans  $\mathbb{C}$  des molécules générées par chaque méthode. Ensuite, les valeurs des indices augmentent, car les méthodes commencent à générer des molécules dans les mêmes cubes de  $\mathbb{C}$ . Les valeurs des indices  $I_{2a}$  et  $I_{2b}$  sont stables à partir de  $10^2$  molécules générées, et sont très similaires. Elles ne permettent pas de comparer la représentativité des molécules produites par les différentes générations. De même, les valeurs de l'indice  $I_{2c}$  sont stables à partir de  $10^5$  molécules générées, et sont similaires. Elles ne permettent pas non plus d'évaluer la représentativité de nos générations. Les valeurs de l'indice  $I_{2d}$  sont stables à partir de  $10^5$  molécules et les valeurs observées à la fin des générations sont plus dispersées. Un classement des méthodes selon cet indice est alors possible,  $I_{2d}$  classe les méthodes comme suit :  $G1b > G1a \approx F0 > F1b > F1a$ . On remarque que la représentativité de F0 surpasse celles de F1b et F1a : F0 semble générer des molécules de manière plus similaire à la référence que F1b et F1a, malgré le fait que les molécules produites par F0 couvrent le moins l'espace chimique. Les méthodes G sont aussi efficaces (pour G1a) ou plus efficaces (pour G1b) que les méthodes F pour produire des molécules occupant l'espace chimique de manière similaire à la référence.

L'uniformité de la couverture de l'espace (indice  $I_3$ ) des structures générées en fonction du nombre de structures générée est représentée sur la Figure 23, diagramme (f). Au début des générations, les valeurs d' $I_3$  sont à leur maximum : les générations ont, à ce moment, toutes une distribution uniforme des molécules générées dans les cubes occupés – soit une molécule projetée dans chaque cube. Les valeurs de l'indice  $I_3$  commencent ensuite à diminuer avec le nombre de molécules générées, signifiant que la distribution des molécules générées dans les cubes s'éloigne d'une distribution uniforme. À la fin des générations, les valeurs de l'indice  $I_3$  sont toutefois faibles et similaires, démontrant que toutes les méthodes produisent des molécules dont la distribution s'éloigne d'une distribution uniforme. Ce comportement n'est pas problématique, il s'explique par le fait que les molécules pouvant être projetées aux extrémités de  $\mathbb{C}$  ont davantage tendance à ne pas respecter les contraintes (sur le respect de l'AD du modèle QSPR ou sur les règles de chimie), et sont écartées. Les molécules non écartées sont projetées majoritairement au centre de  $\mathbb{C}$ .

En conclusion, nous venons d'analyser les méthodes de générations F (F0, F1a, F1b) et G (G1a, G1b) à l'aide des indices *collectifs* développés. Toutes les molécules générées ont été utilisées pour les comparaisons. La comparaison de 5M molécules est suffisante pour obtenir des valeurs d'indices stables et donc pour permettre des comparaisons. Parmi les méthodes F, la méthode F0 produit des molécules qui occupent le moins l'espace  $\mathbb{C}$ , mais dont la distribution dans  $\mathbb{C}$  s'approche le plus de la distribution de référence, ce qui signifie que les molécules produites par F0 occupent la majorité des cubes les plus peuplés par la référence. Au contraire, les méthodes F1a et F1b, qui utilisent des fragments plus grands et plus divers que F0, produisent des molécules occupant une plus grande partie de  $\mathbb{C}$  par rapport à celles de F0. Toutefois, leurs distributions dans  $\mathbb{C}$  s'éloignent plus de la distribution de référence par rapport à la distribution de F0. La méthode G1a obtient des résultats similaires à ceux des méthodes F1b (pour la couverture) et F0 (pour la représentativité). La méthode G1b est la mieux classée par tous les indices. La seule différence entre cette méthode et G1a étant la capacité de G1b à produire des molécules cycliques, nous attribuons les meilleures performances de G1b à cette capacité. Pour comprendre ce phénomène, nous avons étudié les variations des valeurs de descripteurs entre la molécule d'hexane et ses deux possibilités de cyclisation en cyclohexane et en méthylcyclopentane. Le Tableau 29 présente la valeur des descripteurs SMF de type t3l2u4 de ces trois molécules. La cyclisation de l'hexane, en cyclohexane comme en méthylcyclopentane, engendre la modification de la valeur de sept de ses huit descripteurs. Par conséquent, la génération de molécules cycliques permet une diversification importante des molécules.




Molécule	Valeur des descripteurs-fragments							
	CH	CC	CCH	CCC	HCH	HCCH	CCCH	CCCC
Hexane 	14	5	22	4	10	24	18	3
Cyclohexane 	12	6	24	6	6	24	24	6
Méthylcyclopentane 	12	6	22	7	7	19	28	2

Tableau 29 : Valeurs de huit descripteurs (parmi les descripteurs ISIDA encodant les fragments de 2 à 4 atomes reliés par des liaisons simples) des molécules d'hexane, de cyclohexane et de méthylcyclopentane.

### 5.3.2 Génération de molécules diverses possédant leur point d'éclair dans un intervalle défini

Nous nous intéressons dans cette partie à la comparaison des méthodes de génération pour obtenir des molécules possédant des valeurs de point d'éclair dans des intervalles définis. Les molécules présentes dans la base de données sur le PE ont été utilisées dans nos travaux comme ensemble de molécules initiales, et le modèle QSPR construit dans la partie 2.2.2.2 « Modélisation par SVR de la propriété de point d'éclair » a servi comme outil de prédiction. La distribution des valeurs de point d'éclair dans la base de données initiales (entre 165 et 533 K) est présentée sur la Figure 24.

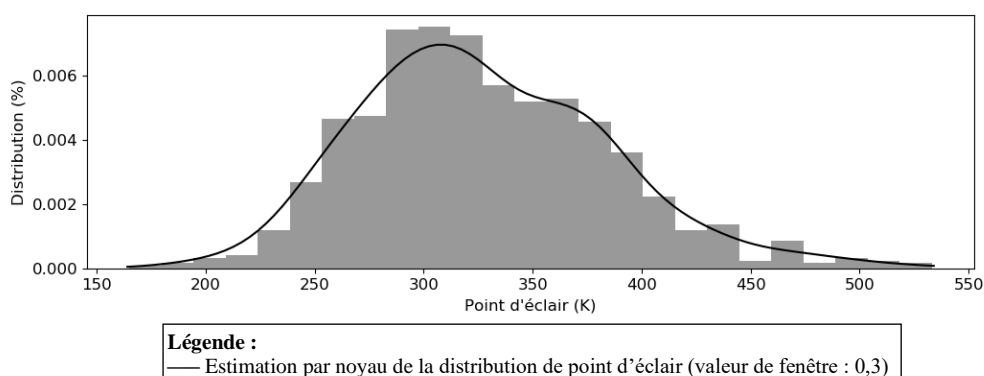


Figure 24 : Distribution des valeurs de point d'éclair dans l'ensemble d'apprentissage du modèle QSPR

Nous avons sélectionné dans la base initiale trois sous-ensembles de molécules en fonction de leur valeur de point d'éclair (PE), de manière à observer dans chaque sous-ensemble une variation du point d'éclair égale à 100 K. Cette amplitude a été définie de façon à couvrir le maximum de données initiales. Les sous-ensembles sont présentés dans le Tableau 30 avec leur nombre de molécules associées. Malgré notre volonté de couvrir un maximum du jeu initial avec les sous-ensembles, onze molécules possèdent leur valeur de propriété en dehors de ces intervalles (cinq molécules avec une valeur de PE inférieure à 200 K et six molécules avec une valeur de PE supérieure à 500 K).

Sous-ensemble	Bornes		Nombre de molécules
	Inférieure (K)	Supérieure (K)	
1	200	300	261
2	300	400	436
3	400	500	77

Tableau 30 : Les différents intervalles de point d'éclair utilisés comme cibles de génération.

Dans ce paragraphe, nous employons les méthodes de génération G1b et G2-X pour générer des molécules diverses possédant leur valeur de PE dans l'un des trois intervalles du Tableau 30. La méthode G1b a été sélectionnée, car elle a obtenu les meilleurs résultats dans la section précédente. La méthode G2-X a aussi été sélectionnée, car cette méthode, similaire à G1b, pourrait permettre d'obtenir plus facilement des molécules avec la propriété désirée. Les méthodes F sont quant à elle étudiées pour générer des molécules diverses dans l'Annexe C.

Dans un premier temps, nous avons généré 50k structures avec G1b et avec G2-X, qui restreint à chaque itération le nombre de parents à X molécules. Nous avons fait varier le nombre de parents X de 800 (soit légèrement supérieur à la taille du jeu initial de 785 molécules) à 50, pour étudier l'impact du nombre X sur la génération de molécules possédant tour à tour leur PE dans chaque intervalle. Avec G2-X, nous avons pris comme PE cible la valeur médiane de chaque intervalle : 250 K, 350 K et 450 K. Nous attirons l'attention sur le fait que chaque variation de G2-X, pour un même nombre de parents X, a été lancée indépendamment pour générer des molécules possédant une valeur de PE dans chaque intervalle, car le choix des molécules parentes est établi en fonction de la propriété ciblée. Les générations sont finalement analysées à l'aide de l'indice de spécificité  $I_4$ . Le Tableau 31 présente la spécificité moyenne sur dix générations des molécules obtenues par les méthodes comparées. Nous observons que les générations de molécules possédant une valeur de PE inférieure à 300 K obtiennent les plus basses valeurs de spécificité. Comme la valeur de PE est corrélée à la taille moléculaire<sup>91,134,135</sup>, nous expliquons ce phénomène par le fait que l'espace chimique des molécules respectant cette contrainte est plus restreint (moins de combinaisons d'atomes possibles) que celui des molécules dont la valeur de PE est plus élevée. Cela démontre que la génération de molécules avec cette contrainte sur la propriété est plus difficile que la génération sans cette contrainte. À

50k molécules générées, la méthode G1b et les variations de G2-X avec un nombre de parents X supérieur à 200 permettent de générer un maximum de structures possédant leur PE dans les deux premiers intervalles. Pour générer des molécules avec un PE compris dans le troisième intervalle, les méthodes G2-X avec un faible nombre de parents X, G2-100 et G2-50, sont les plus efficaces selon la spécificité observée.

Méthode	#Parents	Spécificité pour chaque intervalle de propriété ( $I_4$ )		
		[200 K ; 300 K[	[300 K ; 400 K[	[400 K ; 500 K[
G1b	Tous	0,04 ± 0,001	0,57 ± 0,01	0,39 ± 0,01
G2-800	800	0,03 ± 0,002	0,56 ± 0,02	0,40 ± 0,01
G2-600	600	0,03 ± 0,002	0,57 ± 0,01	0,42 ± 0,02
G2-400	400	0,03 ± 0,002	0,55 ± 0,01	0,42 ± 0,02
G2-200	200	0,02 ± 0,002	0,55 ± 0,03	0,44 ± 0,04
G2-100	100	0,02 ± 0,003	0,51 ± 0,05	0,49 ± 0,04
G2-50	50	0,01 ± 0,002	0,43 ± 0,06	0,49 ± 0,07

Tableau 31 : Spécificité moyenne et écart moyen des méthodes employées pour générer des molécules dans les différents intervalles de point d'éclair, à 50k molécules générées, sur dix générations.

Dans un second temps et suivant ces observations, nous avons conservé les méthodes suivantes pour une analyse plus approfondie : G1b, G2-800 et G2-50. G1b et G2-800 sont choisies car elles produisent des résultats de spécificité similaires pour chaque intervalle de PE, nous souhaitons observer si nous pouvons comparer autrement que par l'indice de spécificité ces deux méthodes. G2-50 a été choisie car elle obtient une des meilleures spécificités pour générer des molécules dans le troisième intervalle de PE. Comme pour la génération sans contrainte sur la propriété, 5M structures prédictibles sont maintenant générées avec chaque méthode. A la fin des générations, seules les y premières molécules générées par chaque méthode répondant à la contrainte sur la propriété sont conservées pour être analysées, où y est le plus petit nombre commun de molécules obtenues répondant à cette contrainte par une méthode. En effet et pour



rappel, notre méthode comparative nécessite d'utiliser un nombre identique de structures produites par chaque méthode, car la distribution de référence utilisée est construite avec le même nombre de molécules générées par chaque méthode comparée. Dès lors,  $I_4$  indique, pour un même nombre de molécules générées respectant la contrainte, la facilité de génération de celles-ci. Les indices évaluant la couverture, la représentativité et l'uniformité de la couverture de l'espace sont également calculés. Parmi les indices évaluant la représentativité de l'espace, nous avons choisi d'utiliser uniquement  $I_{2d}$ , car c'est celui qui a permis la meilleure comparaison des méthodes comparé à  $I_{2a}$ ,  $I_{2b}$  et  $I_{2c}$ .

Chaque méthode a permis d'obtenir, parmi les 5M molécules prédictibles générées, un minimum de  $3,53.10^3$  molécules avec une valeur de point d'éclair prédite dans l'intervalle [200 K ; 300 K]. Les valeurs des indices à  $3,53.10^3$  molécules générées répondant à la contrainte sur la propriété sont présentées dans le Tableau 32. Les évolutions de chaque indice en fonction du nombre de molécules générées répondant à la contrainte sur la propriété sont présentées sur la Figure 25. A la fin des générations, chaque méthode a produit des molécules projetées dans environ 80% des cubes utilisés par la référence (selon l'indice  $I_1$ ), avec une distribution assez semblable à celle de la référence (selon l'indice  $I_{2d}$ ). L'uniformité de la couverture de l'espace (indice  $I_3$ ) diminue avec le nombre de molécules générées de manière similaire pour les trois méthodes. Les valeurs de l'indice  $I_4$  diminuent rapidement pour atteindre une valeur proche de zéro, ce qui montre que les générations produisent de moins en moins de molécules possédant leur point d'éclair dans l'intervalle [200 K ; 300 K] avec le nombre total de molécules générées. G1b est la méthode permettant d'obtenir le plus facilement des molécules respectant la contrainte, la valeur de son indice  $I_4$  est plus de six fois supérieure à celle de G2-800 et plus de trente fois supérieure à celle de G2-50.

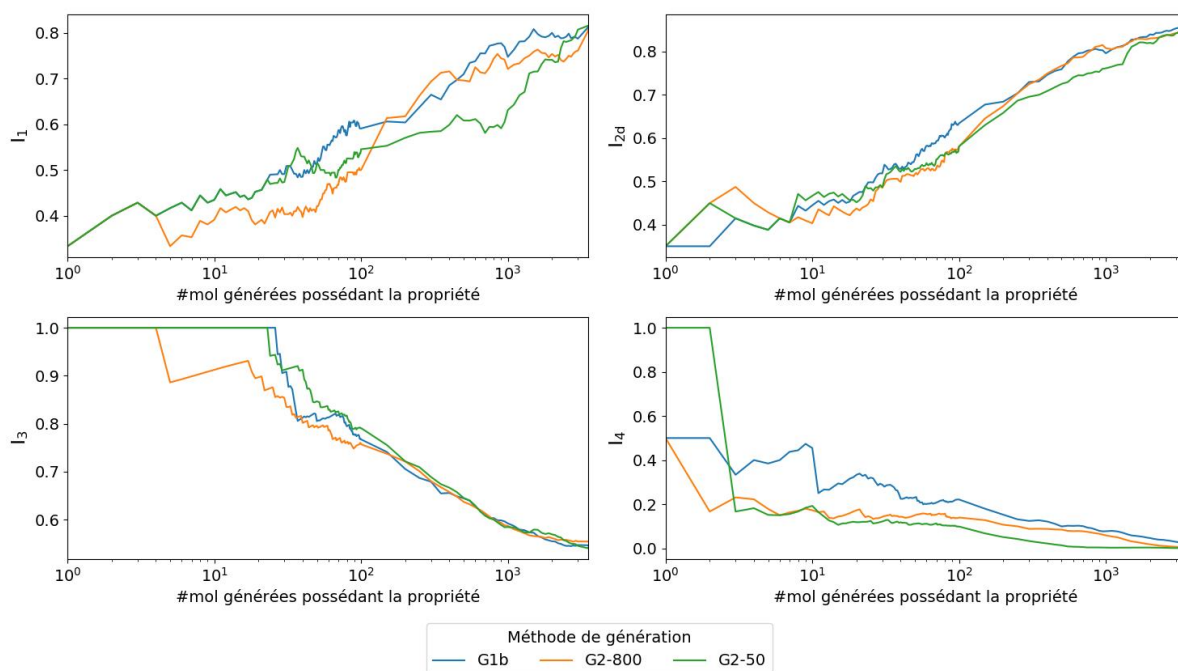


Figure 25 : Variation des indices  $I_1$  à  $I_4$  en fonction du nombre de molécules générées dans l'intervalle [200 K ; 300 K]. Adapté avec permission, issu de Gantzer *et coll.*<sup>94</sup> Copyright 2021 American Chemical Society.

Indice	Valeur d'indice pour la génération par		
	G1b	G2-800	G2-50
$I_1$	0,82	0,80	0,82
$I_{2d}$	0,86	0,85	0,85
$I_3$	0,55	0,55	0,54
$I_4$ ( $\times 10^{-2}$ )	2,35	0,37	0,07

Tableau 32 : Valeurs des indices  $I_1$  à  $I_4$  pour  $3,53 \cdot 10^3$  molécules générées possédant leur point d'éclair dans l'intervalle [200 K ; 300 K].

Chaque méthode a permis d'obtenir, parmi les 5M molécules prédictibles générées, un minimum de  $1,48 \cdot 10^6$  molécules avec une valeur de point d'éclair prédite dans l'intervalle [300

K ; 400 K]. Les valeurs des indices à  $1,48 \cdot 10^6$  molécules générées répondant à la contrainte sur la propriété sont présentées dans le Tableau 33. Les évolutions de chaque indice en fonction du nombre de molécules générées répondant à la contrainte sur la propriété de propriété sont observées sur la Figure 26. Nous observons une diminution puis une augmentation des valeurs d'indices  $I_1$ ,  $I_{2d}$  et  $I_4$  pour la méthode G2-50, vers  $4 \cdot 10^4$  molécules générées. Cette observation illustre la transition entre deux générations concaténées de la méthode G2-50 ; le rafraîchissement de la base de parents permet d'obtenir des molécules dont la projection se rapproche de celle des molécules générées par les autres méthodes. À la fin des générations, la méthode G1b obtient la meilleure couverture de l'espace (indice  $I_1$ ), suivie par G2-800 et G2-50. G2-800 obtient la meilleure représentativité de l'espace (indice  $I_{2d}$ ), suivie par G2-50 et G2-50. L'uniformité de la couverture de l'espace (indice  $I_3$ ) est similaire pour les trois méthodes. La spécificité (indice  $I_4$ ) diminue avec le nombre de molécules générées, comme pour les générations de molécules dans l'intervalle de FP précédent, mais uniquement à partir de  $10^3$  molécules générées. Selon les valeurs finales de l'indice  $I_4$ , G1b est la méthode la plus rapide pour obtenir des molécules avec un point d'éclair dans cet intervalle, suivie par G2-800 puis par G2-50.

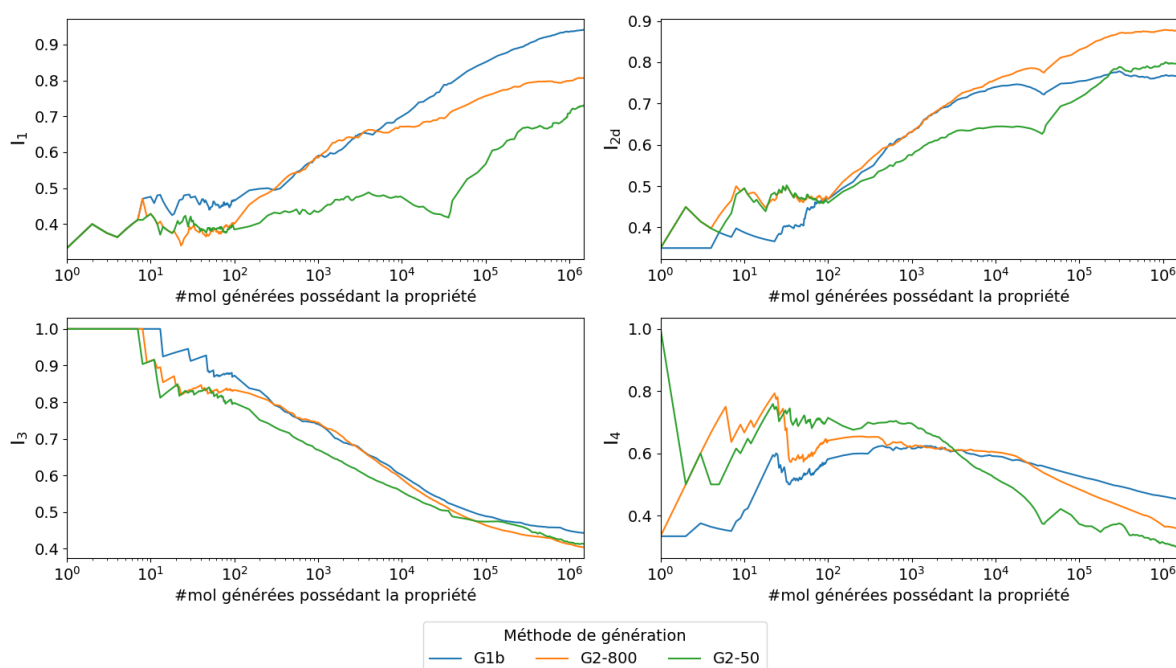


Figure 26 : Variation des indices  $I_1$  à  $I_4$  en fonction du nombre de molécules générées dans l'intervalle [300 K ; 400 K]. Adapté avec permission, issu de Gantzer *et coll.*<sup>94</sup> Copyright 2021 American Chemical Society.

Indice	Valeur d'indice pour la génération par		
	G1b	G2-800	G2-50
$I_1$	0,94	0,81	0,73
$I_{2d}$	0,77	0,88	0,80
$I_3$	0,44	0,40	0,41
$I_4$	0,45	0,36	0,30

Tableau 33 : Valeurs des indices  $I_1$  à  $I_4$  pour  $1,48.10^6$  molécules générées possédant leur point d'éclair dans l'intervalle [300 K ; 400 K], en fonction de la méthode de génération.

Chaque méthode a permis d'obtenir, parmi les 5M molécules prédictibles générées, un minimum de  $2,75.10^6$  molécules avec une valeur de point d'éclair prédite dans l'intervalle [400 K ; 500 K]. Les valeurs des indices à  $2,75.10^6$  molécules générées répondant à la contrainte sur la propriété sont présentées dans le Tableau 34. Les évolutions de chaque indice en fonction du nombre de molécules générées répondant à la contrainte sur la propriété sont présentées sur la Figure 27. À la fin des générations, comme pour les générations de molécules avec une valeur de FP comprise dans les intervalles précédents, G1b obtient la meilleure couverture de l'espace (indice  $I_1$ ), suivie par G2-800 et G2-50. La plus haute représentativité de l'espace (indice  $I_{2d}$ ) est obtenue par G2-800, suivie par G2-50 et G1b. L'uniformité de la couverture de l'espace (indice  $I_3$ ) est similaire pour les trois méthodes. Les valeurs de spécificité (indice  $I_4$ ) augmentent avec le nombre de molécules générées possédant la propriété, faisant voir l'augmentation des capacités des méthodes à générer de telles molécules. G2-50 obtient la meilleure spécificité, suivie par G2-800 et G1b.

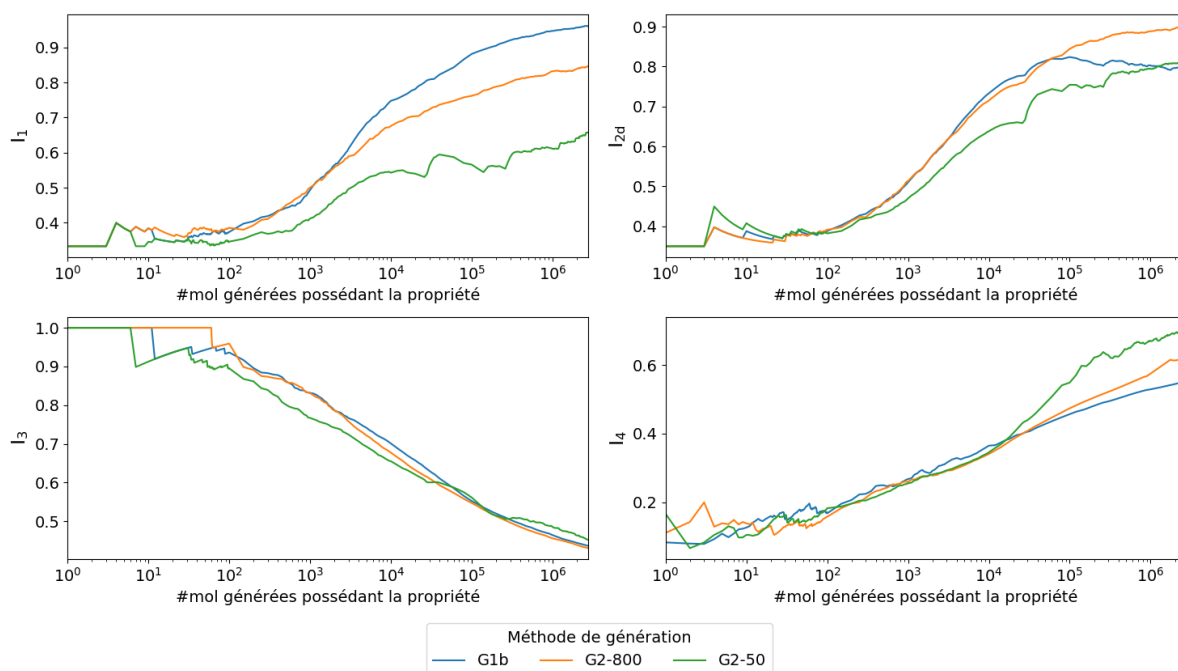


Figure 27 : Variation des indices  $I_1$  à  $I_4$  en fonction du nombre de molécules générées dans l'intervalle [400 K ; 500 K]. Adapté avec permission, issu de Gantzer *et coll.*<sup>94</sup> Copyright 2021 American Chemical Society.

Indice	Valeur d'indice pour la génération par		
	G1b	G2-800	G2-50
$I_1$	0,96	0,85	0,66
$I_{2d}$	0,80	0,90	0,81
$I_3$	0,44	0,43	0,45
$I_4$	0,55	0,62	0,71

Tableau 34 : Valeurs des indices  $I_1$  à  $I_4$  pour  $2,75 \cdot 10^6$  molécules générées possédant leur point d'éclair dans l'intervalle [400 K ; 500 K], en fonction de la méthode de génération.

Pour chaque méthode, et particulièrement pour G2-50, la concaténation de plusieurs générations a permis de contourner les contraintes liées à la diversité du jeu de molécules à modifier par les méthodes G. L'étude des variations des indices avec le nombre de molécules générées a mis en avant l'impact de ces concaténations. Les tendances de variation des indices  $I_1$ ,  $I_{2a}$  et  $I_3$  sont similaires à celles observées lors des générations sans contrainte sur la propriété. Parmi les méthodes étudiées dans cette sous-section, G1b est la méthode qui permet d'obtenir la meilleure couverture de l'espace, quelle que soit la contrainte sur la propriété utilisée. En effet, la diversité chimique contenue dans sa base de molécules parentes – non restreinte – est plus importante que celle des bases de G2-X – pour lesquelles le nombre de molécules est restreint à X molécules. Cette caractéristique de G1b est la raison de ses meilleures performances pour générer des molécules avec un point d'éclair inférieur ou égal à 400 K. G2, quant à elle, est davantage adaptée à la génération de molécules avec un point d'éclair supérieur à 400 K, pour lesquelles le nombre de combinaisons d'atomes est plus élevée, et par conséquent pour lesquelles le nombre d'opérations de modification réalisables est plus élevé. Cette caractéristique est la raison des bonnes performances de G2-50 pour générer des molécules de grande taille – avec un point d'éclair supérieur à 400 K.

## 5.4 Conclusions sur la comparaison des générations moléculaires

Dans ce chapitre, nous nous sommes intéressés à la comparaison de méthodes de génération. Nous avons pour cela utilisé l'espace étendu  $\mathbb{C}$ , qui est construit sur la base des trois premières composantes principales issues d'une ACP.  $\mathbb{C}$  est discrétisé à l'aide de cubes de taille identique, puis chaque molécule générée est projetée dans un cube de  $\mathbb{C}$ . L'occupation de ces cubes par les molécules générées est analysée avec un ensemble de nouveaux indices *collectifs* définis dans ce chapitre. La couverture (avec l'indice  $I_1$ ), la représentativité (avec les indices  $I_{2a}$ ,  $I_{2b}$ ,  $I_{2c}$  et  $I_{2d}$ ) et l'uniformité de la couverture (avec l'indice  $I_3$ ) de l'espace sont étudiées. L'indice de spécificité  $I_4$  est également utilisé et permet d'apprécier la capacité des méthodes à générer des molécules respectant une contrainte sur la propriété.

Ces indices ont été appliqués pour comparer des générations à partir du jeu initial sur le point d'éclair (décrites dans le Chapitre 4 « Génération virtuelle de molécules ») et son modèle QSPR associé. Les méthodes de génération par assemblages de fragments (F, avec les méthodes F0,

F1a, F1b) et par modifications successives (G, avec les méthodes G1a, G1b, G2-X) ont été examinées. Pour générer des molécules sans contrainte sur la propriété, les méthodes modifiant itérativement les molécules (G) ont obtenu d'aussi bons (pour G1a) voire meilleurs (pour G1b) résultats que celles assemblant des fragments (F). Nous avons attribué ce comportement au fait que les méthodes G utilisent plusieurs opérateurs de modification et une bibliothèque dynamique de fragments pour générer des molécules, alors que les méthodes F ne peuvent générer des molécules que par concaténation de fragments issus d'une bibliothèque statique. La capacité de manipuler et générer des molécules cycliques, implémentée dans G1b, permet d'améliorer encore plus les performances de génération des méthodes G face à celles des méthodes F, pour lesquelles nous n'avons pas implémenté l'ajout de fragments cycliques. Pour générer des molécules possédant leur point d'éclair dans un des trois intervalles parmi : [200 K ; 300 K[, [300 K ; 400 K[ et [400 K ; 500 K[, nous avons sélectionné la méthode G1b ainsi que sa variante G2-X, qui restreint le nombre de parents X à 800 (G2-800) et 50 (G2-50) molécules. G1b permet d'obtenir des molécules davantage clairsemées dans C et par conséquent des molécules plus diverses que celles obtenues avec G2-50 ou G2-800. Cette particularité facilite la génération de molécules petites à moyennes, pour lesquelles le nombre de combinaisons d'atomes est modéré. Au contraire, la génération de molécules plus grandes est favorisée avec G2-X, qui favorise les motifs structuraux ayant mené à la propriété souhaitée.

## Chapitre 6. Conclusions et Perspectives

---

### 6.1 Conclusions

Dans ce manuscrit de thèse, nous avons adressé les problèmes de la génération moléculaire et de la comparaison des générations, pour proposer de nouvelles structures dites « corps pur ». La première partie du travail de cette thèse a été consacrée à la mise en place d'outils – modèles QSPR – pour la prédiction des propriétés de telles molécules. La méthodologie utilisée à IFPEN, pour le développement et l'utilisation de modèles QSPR à partir de « Machine à Vecteurs de Support » (SVR, pour « Support Vector Regression »)<sup>63</sup> a été reprise et améliorée. Particulièrement, nous y avons intégré un processus automatique pour la vérification du domaine d'applicabilité (AD) des molécules à prédire.

Ces outils ont été utilisés pour modéliser le point d'éclair (PE), à partir d'un jeu de 785 molécules avec pour chacune une valeur expérimentale de référence. Tout d'abord, un modèle SVR a été construit à partir de ces molécules, encodées par des descripteurs ISIDA de type séquences de deux à quatre atomes et leurs liaisons (t3l2u4). Le modèle ainsi obtenu a démontré des performances en validation externe (valeurs de RMSE de 15,5 K et  $R^2$  de 0,935) similaires à celles reportées dans les travaux de Saldana *et coll.*<sup>91</sup> Nous avons utilisé une plus grande diversité de molécules lors de la construction du modèle que Saldana *et coll.*<sup>91</sup>, ce qui permet à l'AD de notre modèle de couvrir une plus grande variété de molécules que celui du modèle de Saldana *et coll.*<sup>91</sup> C, la représentation graphique de notre jeu de données sur le PE, a été construite à partir des trois premières composantes principales (3 PC) issues d'une « Analyse en Composantes Principales » (ACP)<sup>66</sup>. C approxime sur trois dimensions l'espace chimique du modèle QSPR. Parallèlement, la méthode de « Cartographie Topographique Générative » (GTM, pour « Generative Topographic Mapping »)<sup>68</sup> a été utilisée avec notre jeu de données sur le PE pour apporter une autre représentation de l'espace chimique. La carte GTM obtenue présente l'espace chimique sur deux dimensions. Elle distingue efficacement les hydrocarbures des composés hydrogénés, et également les molécules de différentes sous-familles chimiques entre elles (par exemple, les alcools des esters).



Nous avons ensuite effectué un travail bibliographique sur la génération de structures moléculaires, que nous avons valorisé à travers la publication d'une revue.<sup>34</sup> Cette étude bibliographique nous a permis d'identifier trois types de méthodes pour générer des structures moléculaires : la génération par assemblage de fragments (F), par modifications successives de graphes génétiques (G) et par autoencodeur variationnel (E). Nous avons implémenté, testé et amélioré ces méthodes pour générer des molécules à partir de notre base initiale (données de PE). Les méthodes F consistent à concaténer des fragments moléculaires entre eux.<sup>99</sup> Deux types de fragments ont été utilisés : des fragments dits « simples » de deux atomes, et les fragments de deux à quatre atomes encodés par les descripteurs t3l2u4. Les pondérations du choix du type de liaison à créer entre deux fragments, et du choix des fragments à concaténer, ont été étudiées. Elles ont permis de diminuer le pourcentage de structures incorrectes générées (doublons et/ou n'appartenant pas à l'AD du modèle QSPR). Les méthodes G s'inspirent des graphes génétiques<sup>111</sup> et modifient itérativement les structures de la base parente par un ensemble d'opérations sur leur graphe. Selon notre implémentation, la base parente est composée des structures de notre base initiale ainsi que de celles générées. En ne modifiant que les molécules de la base parente respectant l'AD, nous avons légèrement amélioré la génération et notamment, diminué le pourcentage de structures incorrectes générées. Nous avons également implémenté la possibilité de ne sélectionner comme parentes que les molécules dont la valeur de propriété est la plus proche de la valeur cible. Les méthodes E utilisent un autoencodeur variationnel (VAE).<sup>121,122</sup> Le VAE commence par projeter les molécules initiales dans un espace latent – à haute dimension – sous forme de vecteurs. L'application d'un bruit sur ces vecteurs latents puis leur décodage permet de générer de nouvelles structures. Nous avons démontré qu'il était possible d'entraîner un tel VAE à partir de structures précédemment générées.

Les outils actuellement disponibles pour comparer des méthodes de génération<sup>129-131</sup> entre elles ne sont pas adaptés à notre cas d'étude sur le PE. En effet, ils ont été conçus pour évaluer la capacité des méthodes à générer des molécules similaires à celles d'un jeu initial, représentatif de la diversité chimique souhaitée. Dans notre travail, nous avons souhaité évaluer la capacité des méthodes à compléter des bases de données moins représentatives – par exemple sur le PE. C'est pourquoi, nous avons proposé une nouvelle approche de comparaison des méthodes, basée sur une série d'indices *collectifs*. Dans notre approche, les molécules générées par chaque méthode sont d'abord projetées dans l'espace  $\mathbb{C}$ . La répartition de leurs projections est ensuite

analysée en termes de couverture de l'espace (indice  $I_1$ ), représentativité de la couverture (indice  $I_{2a}$  à  $I_{2d}$ ) et uniformité de la couverture (indice  $I_3$ ). L'analyse s'effectue à l'aide d'une référence, celle-ci est construite dynamiquement en considérant l'ensemble des molécules générées par toutes les méthodes à comparer. C'est à notre connaissance le premier outil qui permet de comparer des méthodes de génération, dynamiquement, sans base de référence fixe. Notre outil permet la comparaison de générations pour des applications diverses, et est complémentaire aux outils existants<sup>129-131</sup> qui, eux, comparent la capacité des méthodes de génération à reproduire la diversité chimique des bases initiales.

Notre approche a été utilisée dans un premier temps pour comparer les méthodes sélectionnées F (F0, F1a et F1b) et G (G1a et G1b), à partir des 5 millions de molécules générées par chacune d'entre elles. Ces comparaisons conduisent à la supériorité des méthodes G pour générer des molécules à partir de notre jeu de données sur le PE. En effet, les méthodes G utilisent des opérations variées pour générer des molécules, alors que les méthodes F n'ajoutent que des fragments. La méthode G1b est la mieux classée et cela est attribuable au fait qu'elle est la seule à considérer les molécules cycliques. Dans un second temps, G1b et ses variantes G2-800 et G2-50, qui restreignent le nombre de parents à 800 (G2-800) et 50 (G2-50) molécules, respectivement, ont été utilisées pour générer des molécules possédant une valeur de PE dans un intervalle défini parmi : [200 K ; 300 K[, [300 K ; 400 K[ et [400 K ; 500 K[. Nous en avons conclu que G1b permet d'obtenir des molécules davantage éparpillées dans  $\mathbb{C}$  que celles produites par G2-800 et G-50, et par conséquent permet la génération de molécules plus diverses. Cette caractéristique permet à G1b de surpasser G2-800 et G2-50 pour la génération de molécules avec une valeur de PE inférieure à 400 K. Au contraire, la contrainte sur le nombre de molécules parentes permet à G2-50 d'obtenir les meilleurs résultats pour la génération de molécules avec une valeur de PE supérieure ou égale à 400 K. La restriction sur le nombre de molécules parentes est d'autant plus efficace que la taille des molécules générées est importante.

## 6.2 Perspectives

Nous avons mis en évidence dans la section 5.3 « Comparaison des méthodes de génération à l'aide des nouveaux indices » que la génération de molécules cycliques était importante pour augmenter la diversité des molécules obtenues, ce qui rejoint des observations déjà émises par Inoue *et coll.*<sup>138</sup> Par conséquent, l'intégration de fragments contenant des cycles est une amélioration à considérer pour les méthodes F. Dans les méthodes G, la génération de cycles peut être davantage diversifiée en implémentant des opérations de croisement entre deux cycles, et de fusion de cycles.<sup>138</sup> La permutation de liaisons est une autre opération de modification des molécules qui pourrait également être implémentée dans les méthodes G.<sup>139</sup> Pour la méthode E, d'autres moyens existent pour contourner le manque de données comme l'énumération de tous les SMILES non canoniques de chaque molécule.<sup>133</sup> Combiner notre approche (utiliser les structures précédemment générées pour entraîner le VAE) à cette technique pourrait permettre d'augmenter les capacités de la génération par autoencodeur variationnel, car l'espace latent serait construit à partir d'encore davantage de structures.

La méthode que nous avons mise en place pour comparer les méthodes de génération examine les projections des molécules générées dans l'espace  $\mathbb{C}$ , un espace à 3D. On retiendra que  $\mathbb{C}$  approxime l'espace chimique des molécules initiales car les 3 PC constituant cet espace ne permettent d'expliquer que 37% de la variance du jeu initial de données. Notre méthode peut être adaptée pour analyser les projections des molécules générées dans d'autres espaces, par exemple celui de la carte GTM créée. La méthode GTM permet de représenter davantage de caractéristiques du jeu de données par rapport à la méthode ACP, et ainsi augmenter la description du jeu initial.<sup>140</sup> Un autre désavantage d'utiliser  $\mathbb{C}$  réside dans le fait que cet espace est construit à partir des données initiales, et donc n'est pas identique d'un jeu de données à l'autre. Alternativement, la cartographie de l'ensemble de l'espace chimique existant par un ensemble de cartes GTM « universelles » est actuellement étudiée et pourrait être utilisée dans notre approche.<sup>141,142</sup>

Enfin, bien que nous ayons uniquement considéré la génération de molécules répondant à une contrainte sur une seule propriété, il est possible d'étendre facilement ce travail pour la génération de molécules possédant des contraintes sur plusieurs propriétés. Pour cela, un

ensemble de modèles QSPR doit d'abord être construit pour prédire chaque propriété. L'AD « global » est ensuite défini comme l'espace chimique couvert par tous les modèles QSPR individuels, c'est-à-dire l'intersection des espaces chimiques occupés par les molécules de chaque modèle QSPR. Pour la génération, la méthode G2-X doit trier les molécules parentes avec une fonction de score basée sur l'écart entre leurs valeurs prédites et désirées de toutes les propriétés ciblées. La fonction de score actuellement utilisée par G2-X doit être adaptée, et peut par exemple être de la forme  $\sum_{x \in X} \left( \frac{|x_{pred} - x_{desiree}|}{x_{desiree}} \right) / n$ , où  $x$  représente une propriété parmi l'ensemble  $X$  des propriétés évaluées,  $x_{desiree}$  sa valeur souhaitée,  $x_{predite}$  sa valeur prédite par QSPR pour la molécule évaluée, et  $n$  le nombre de propriétés dans  $X$ . De même, les méthodes de génération peuvent être étendues à la génération de mélanges ; deux approches sont envisageables. Avec la première approche, toutes les combinaisons possibles de structures générées sont effectuées, pour obtenir un maximum de mélanges. Avec la seconde approche, la génération est séquentielle, c'est-à-dire que les structures de chaque mélange sont générées une à une, en prenant en compte les structures déjà générées pour le mélange et leurs impacts sur les contraintes sur les propriétés à respecter.

## Annexe A. Molécules de la base de données sur le PE

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
methyl propanoate	<chem>CCC(=O)OC</chem>	Ester	apprentissage	276,8	5,8
cyclopentene	<chem>C1CC=CC1</chem>	Cyclique	apprentissage	233,1	10,9
1,4-butanediol	<chem>OCCCCO</chem>	Alcool	apprentissage	391,3	15,8
2,2,3,3-tetramethylhexane	<chem>CCCC(C)(C)C(C)(C)C</chem>	Paraffine	apprentissage	300,4	3,6
4-ethyl-1,2-dimethylbenzene	<chem>CCc1ccc(C)c(C)c1</chem>	Aromatique	apprentissage	327,7	3,3
4-tert-butylpyrocatechol	<chem>CC(C)(C)c1ccc(O)c(O)c1</chem>	Aromatique (Oxygéné)	apprentissage	398,8	4,2
3-methyl-1-butanol	<chem>CC(C)CCO</chem>	Alcool	apprentissage	310,6	5,4
ethyl hept-6-enoate	<chem>CCOC(=O)CCCCC=C</chem>	Ester	apprentissage	338,5	13,5
methyl heptadecanoate	<chem>CCCCCCCCCCCCCCCC(=O)OC</chem>	Ester	apprentissage	402,2	19,2
methyl ethanoate	<chem>COC(C)=O</chem>	Ester	apprentissage	266,3	3,1
butylcyclohexane	<chem>CCCCC1CCCCC1</chem>	Cyclique	apprentissage	320,3	0,7
methyl (2E)-non-2-enoate	<chem>CCCCCC=CC(=O)OC</chem>	Ester	apprentissage	361,2	2,8
2,6-dimethylheptane	<chem>CC(C)CCCC(C)C</chem>	Paraffine	apprentissage	292,4	6,6
hexane-1,2,6-triol	<chem>OCCCC(O)CO</chem>	Alcool	apprentissage	459,3	11,7
2-methylheptane	<chem>CCCCC(C)C</chem>	Paraffine	apprentissage	279,9	2,8
pentyl methanoate	<chem>CCCCCOC=O</chem>	Ester	apprentissage	308,7	10,3
p-ethyltoluene	<chem>CCc1ccc(C)cc1</chem>	Aromatique	apprentissage	311,6	2,4
1,4-dioxane	<chem>C1COCCO1</chem>	Cyclique (Oxygéné)	apprentissage	278,2	6,8
2,2,4,4-tetramethylhexane	<chem>CCC(C)(C)CC(C)(C)C</chem>	Paraffine	apprentissage	298,0	6,1
4-ethylheptane	<chem>CCCC(CC)CCC</chem>	Paraffine	apprentissage	298,0	10,0
2-(2-ethylhexyloxy)-ethanol	<chem>CCCC(CC)COCCO</chem>	Ether	apprentissage	378,0	5,1
6,6-dimethyl-4-methylidenebicyclo[3.1.1]heptane	<chem>CC1(C)C2CC1C(=C)CC2</chem>	Cyclique	apprentissage	308,0	4,0
naphthalene	<chem>c1ccc2ccccc2c1</chem>	Aromatique	apprentissage	350,1	1,9
butyl methacrylate	<chem>CCCCOC(=O)C(C)=C</chem>	Ester	apprentissage	324,6	2,6

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
{2-[(2-phenylpropan-2-yl)peroxy]propan-2-yl}benzene	<chem>CC(C)(OOC(C)(C)c1ccccc1)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	393,2	7,0
1,2,3-trimethylbenzene	<chem>Cc1ccc(C)c1C</chem>	Aromatique	apprentissage	319,0	1,9
butyl methanoate	<chem>CCCCOC=O</chem>	Ester	apprentissage	293,3	2,3
2,3,4,5-tetramethylhexane	<chem>CC(C)C(C)C(C)C(C)C</chem>	Paraffine	apprentissage	305,3	1,3
4-methyloctane	<chem>CCCC(C)CCC</chem>	Paraffine	apprentissage	297,1	2,1
[(benzyloxy)methyl]benzene	<chem>C(OCc1ccccc1)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	414,2	6,1
5-ethylidinenorbornene	<chem>CC=C1CC2CC1C=C2</chem>	Cyclique	apprentissage	304,2	7,0
methyl (9Z)-octadec-9-enoate	<chem>CCCCCCCC=CCCCCCCC(=O)OC</chem>	Ester	apprentissage	390,9	4,9
2-ethyl-1-butanol	<chem>CCC(CC)CO</chem>	Alcool	apprentissage	322,5	7,5
1-heptanol	<chem>CCCCCCC</chem>	Alcool	apprentissage	340,6	5,4
decanal	<chem>CCCCCCCCC=O</chem>	Aldehyde	apprentissage	357,3	0,8
ethyl 3-methylbutanoate	<chem>CCOC(=O)CC(C)C</chem>	Ester	apprentissage	310,9	12,7
2,3,3-trimethyl-1-butene	<chem>CC(=C)C(C)(C)C</chem>	Alcène	apprentissage	254,5	1,6
alpha-alpha-dimethyl benzene methanol	<chem>CC(C)(O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	359,2	1,0
triethylene glycol monomethyl ether	<chem>COCCOCCOCCO</chem>	Ether	apprentissage	398,5	7,0
3,3-dimethylhexane	<chem>CCCC(C)(C)CC</chem>	Paraffine	apprentissage	274,3	2,3
prop-1-ene	<chem>CC=C</chem>	Alcène	apprentissage	194,1	29,1
cyclohexyl methanecarboxylate	<chem>CC(=O)OC1CCCCC1</chem>	Cyclique (Oxygéné)	apprentissage	338,0	7,9
ethyl heptadecanoate	<chem>CCCCCCCCCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	425,5	2,5
3-ethylhex-1-ene	<chem>CCCC(CC)C=C</chem>	Alcène	apprentissage	279,2	5,1
1,4-hexadiene	<chem>CC=CCC=C</chem>	Alcène	apprentissage	252,3	4,1
diethyl phthalate	<chem>CCOC(=O)c1ccccc1C(=O)OCC</chem>	Aromatique (Oxygéné)	apprentissage	397,1	7,0
ethyl buta-2,3-dienoate	<chem>CCOC(=O)C=C=C</chem>	Ester	apprentissage	318,3	1,7
propanol	<chem>CCCO</chem>	Alcool	apprentissage	288,6	9,4
2-(butan-2-yloxy)butane	<chem>CCC(C)OC(C)CC</chem>	Ether	apprentissage	286,2	1,8
1-(hexyloxy)hexane	<chem>CCCCCCOCCCCC</chem>	Ether	apprentissage	357,1	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
4-methylpent-1-ene	<chem>CC(C)CC=C</chem>	Alcène	apprentissage	242,2	0,2
1-(2-butoxy-isopropoxy)propan-2-ol	<chem>CCCCOCC(C)OCC(C)O</chem>	Ether	apprentissage	376,4	2,9
2,4-dimethylheptane	<chem>CCCC(C)CC(C)C</chem>	Paraffine	apprentissage	293,2	5,2
heptanal	<chem>CCCCCCC=O</chem>	Aldehyde	apprentissage	313,2	5,0
1,3-cyclohexadiene	<chem>C1CC=CC=C1</chem>	Cyclique	apprentissage	260,5	3,5
isoprene	<chem>CC(=C)C=C</chem>	Alcène	apprentissage	225,1	5,8
7-methyloct-1-ene	<chem>CC(C)CCCCC=C</chem>	Alcène	apprentissage	293,3	2,3
2-ethylhexane-1,3-diol	<chem>CCCC(O)C(CC)CO</chem>	Alcool	apprentissage	404,0	5,0
3-ethyl-2,3-dimethylpentane	<chem>CCC(C)(CC)C(C)C</chem>	Paraffine	apprentissage	290,6	2,6
2,3-dimethylpentane	<chem>CCC(C)C(C)C</chem>	Paraffine	apprentissage	261,3	3,3
2-{2-[2-(2-ethylhexanoyloxy)ethoxy]ethoxy}ethyl 2-ethylhexanoate	<chem>CCCC(CC)C(=O)OCCOCCOCCOC(=O)C(C)CCCC</chem>	Ester	apprentissage	470,8	1,3
hexane	<chem>CCCCCC</chem>	Paraffine	apprentissage	250,8	0,7
nonan-5-one	<chem>CCCCC(=O)CCCC</chem>	Cétone	apprentissage	332,2	1,0
2,5,8,11-tetraoxadodecane	<chem>COCCOCCOCCOC</chem>	Ether	apprentissage	380,0	3,2
2-ethoxyethyl acetate	<chem>CCOCCOC(C)=O</chem>	Ester	apprentissage	321,0	7,0
cis-2-heptene	<chem>CCCC=CC</chem>	Alcène	apprentissage	264,9	0,1
1-[2-(2-butoxyethoxy)ethoxy]butane	<chem>CCCCOCCOCCOCCCC</chem>	Ether	apprentissage	384,2	7,0
cycloheptene	<chem>C1CCC=CCC1</chem>	Cyclique	apprentissage	274,1	7,0
2,2,3,3-tetramethylbutane	<chem>CC(C)(C)C(C)(C)C</chem>	Paraffine	apprentissage	267,7	10,3
ethyl (9Z)-octadec-9-enoate	<chem>CCCCCCCCC=CCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	409,8	44,8
2-methyl-1-pentanol	<chem>CCCC(C)CO</chem>	Alcool	apprentissage	321,9	2,9
methyl hexanoate	<chem>CCCCCC(=O)OC</chem>	Ester	apprentissage	314,2	1,8
2-methylbut-2-ene	<chem>CC=C(C)C</chem>	Alcène	apprentissage	235,1	7,0
phenol	<chem>Oc1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	346,2	6,8
1-heptyne	<chem>CCCCCC#C</chem>	Alcyne	apprentissage	278,0	7,0
methyl hexadecanoate	<chem>CCCCCCCCCCCCCCCC(=O)OC</chem>	Ester	apprentissage	397,9	27,1

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
3-methylbutyl propanoate	<chem>CCC(=O)OCCC(C)C</chem>	Ester	apprentissage	324,4	3,4
2-hydroxypropyl methacrylate	<chem>CC(O)COC(=O)C(C)=C</chem>	Ester	apprentissage	365,6	3,5
2-oxacyclobutanone	<chem>O=C1CCO1</chem>	Cyclique (Oxygéné)	apprentissage	295,6	47,6
1-nonanol	<chem>CCCCCCCCO</chem>	Alcool	apprentissage	364,4	11,4
propane-1,1-diol	<chem>CCC(O)O</chem>	Alcool	apprentissage	365,0	7,0
glycerol	<chem>OCC(O)CO</chem>	Alcool	apprentissage	439,8	6,6
diethyl succinate	<chem>CCOC(=O)CCC(=O)OCC</chem>	Ester	apprentissage	370,0	7,0
ethyl 2-hydroxypropanoate	<chem>CCOC(=O)C(C)O</chem>	Ester	apprentissage	326,1	7,0
1,4-dimethylbenzene	<chem>Cc1ccc(C)cc1</chem>	Aromatique	apprentissage	296,7	1,5
2-(2-(2-hydroxypropoxy)propoxy)-1-propyl alcohol	<chem>CC(O)COCC(C)OCC(C)O</chem>	Ether	apprentissage	406,2	7,0
2,3,4-trimethylpentane	<chem>CC(C)C(C)C(C)C</chem>	Paraffine	apprentissage	275,2	2,2
3-methylheptane	<chem>CCCCC(C)CC</chem>	Paraffine	apprentissage	280,9	1,9
octan-2-one	<chem>CCCCCCC(C)=O</chem>	Cétone	apprentissage	324,3	0,1
2,2-diethylpropane-1,3-diol	<chem>CCC(CC)(CO)CO</chem>	Alcool	apprentissage	398,9	24,9
ethylene glycol diacetate	<chem>CC(=O)OCCOC(C)=O</chem>	Ester	apprentissage	356,8	4,3
methoxymethane	<chem>COC</chem>	Ether	apprentissage	220,6	11,4
1-acetyloxyprop-2-enyl acetate	<chem>CC(=O)OC(OC(C)=O)C=C</chem>	Ester	apprentissage	349,3	6,1
neopentane	<chem>CC(C)(C)C</chem>	Paraffine	apprentissage	219,9	11,9
2-methylhex-1-ene	<chem>CCCCC(C)=C</chem>	Alcène	apprentissage	260,2	7,0
ethylcyclohexane	<chem>CCC1CCCCC1</chem>	Cyclique	apprentissage	290,1	5,0
2-methylpropanoic acid	<chem>CC(C)C(O)=O</chem>	Acide carboxylique	apprentissage	330,4	1,2
trans-decahydronaphthalene	<chem>C1CCC2CCCCC2C1</chem>	Cyclique	apprentissage	323,5	1,6
ethyl benzoate	<chem>CCOC(=O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	353,7	3,5
vinylcyclohexene	<chem>C=CC1CCC=CC1</chem>	Cyclique	apprentissage	292,4	3,4
2-propyn-1-ol	<chem>OCC#C</chem>	Alcool	apprentissage	302,3	7,0
butylcyclopentane	<chem>CCCC1CCCC1</chem>	Cyclique	apprentissage	302,3	5,7
2,2,3,4-tetramethylpentane	<chem>CC(C)C(C)C(C)(C)C</chem>	Paraffine	apprentissage	286,1	2,1



Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2,4,5-trimethylheptane	<chem>CCC(C)C(C)CC(C)C</chem>	Paraffine	apprentissage	307,1	3,1
neopentyl ethanoate	<chem>CC(=O)OCC(C)(C)C</chem>	Ester	apprentissage	309,0	7,0
2-butoxyethylacetate	<chem>CCCCOCCOC(C)=O</chem>	Ester	apprentissage	345,7	1,6
1,3-propylene glycol	<chem>OCCCO</chem>	Alcool	apprentissage	382,1	30,0
2-norbornene	<chem>C1CC2CC1C=C2</chem>	Cyclique	apprentissage	267,6	9,4
3-ethyl-4-methylhexane	<chem>CCC(C)C(CC)CC</chem>	Paraffine	apprentissage	295,9	7,9
methyl(13Z)docos-13-enoate	<chem>CCCCCCCCC=CCCCCCCCCCCCC(=O)OC</chem>	Ester	apprentissage	393,2	87,2
2-methyloxirane	<chem>CC1CO1</chem>	Cyclique (Oxygéné)	apprentissage	230,3	5,7
ethyl heptanoate	<chem>CCCCCCC(=O)OCC</chem>	Ester	apprentissage	339,8	7,2
3-benzenepropanol	<chem>OCCc1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	378,3	5,2
2,5-dimethylhexane	<chem>CC(C)CCC(C)C</chem>	Paraffine	apprentissage	276,0	5,0
1-decanol	<chem>CCCCCCCCCO</chem>	Alcool	apprentissage	375,5	2,5
propyl ethanoate	<chem>CCCOC(C)=O</chem>	Ester	apprentissage	288,1	0,1
dodecyl propanoate	<chem>CCCCCCCCCCCCOC(=O)CC</chem>	Ester	apprentissage	402,6	18,4
2-(4-methylcyclohex-3-enyl)propan-2-ol	<chem>CC1=CCC(CC1)C(C)(C)O</chem>	Cyclique (Oxygéné)	apprentissage	360,2	7,0
2-methylpropyl methanoate	<chem>CC(C)COC=O</chem>	Ester	apprentissage	294,9	7,1
3,4-dimethylheptane	<chem>CCCC(C)C(C)CC</chem>	Paraffine	apprentissage	295,0	7,0
anthracene	<chem>c1ccc2cc3ccccc3cc2c1</chem>	Aromatique	apprentissage	426,4	32,4
p-diethylbenzene	<chem>CCc1ccc(CC)cc1</chem>	Aromatique	apprentissage	326,3	2,9
m-ethyltoluene	<chem>CCc1cccc(C)c1</chem>	Aromatique	apprentissage	311,6	0,4
salicylaldehyde	<chem>Oc1ccccc1C=O</chem>	Aromatique (Oxygéné)	apprentissage	356,1	7,0
1,2,4-triethylbenzene	<chem>CCc1ccc(CC)c(CC)c1</chem>	Aromatique	apprentissage	352,6	0,5
nonan-3-one	<chem>CCCCCCC(=O)CC</chem>	Cétone	apprentissage	332,2	7,0
4,6,6-trimethylbicyclo[3.1.1]hept-3-ene	<chem>CC1=CCC2CC1C2(C)C</chem>	Cyclique	apprentissage	310,1	7,0
3-methylbutyl 3-methylbutanoate	<chem>CC(C)CCOC(=O)CC(C)C</chem>	Ester	apprentissage	346,9	1,9
1-methylcyclohexanol	<chem>CC1(O)CCCCC1</chem>	Cyclique (Oxygéné)	apprentissage	331,2	7,0
1,2,3,4-tetrahydronaphthalene	<chem>C1CCc2ccccc2C1</chem>	Aromatique	apprentissage	333,1	10,9

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
ethyl hexadecanoate	<chem>CCCCCCCCCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	420,5	34,5
2,3-dimethyl-1,3-butadiene	<chem>CC(=C)C(C)=C</chem>	Alcène	apprentissage	244,2	7,0
furan	<chem>c1ccoc1</chem>	Aromatique (Oxygéné)	apprentissage	244,0	7,0
methyl (2E)-oct-2-enoate	<chem>CCCCC=CC(=O)OC</chem>	Ester	apprentissage	349,4	6,6
cis-1,2-dimethylcyclohexane	<chem>CC1CCCCC1C</chem>	Cyclique	apprentissage	286,9	2,9
dodecyl butanoate	<chem>CCCCCCCCCCCCOC(=O)CCC</chem>	Ester	apprentissage	410,1	22,9
cyclohex-3-ene-1-carbaldehyde	<chem>O=CC1CCC=CC1</chem>	Cyclique (Oxygéné)	apprentissage	315,2	4,9
1-hexanol	<chem>CCCCCCO</chem>	Alcool	apprentissage	328,0	7,0
butan-2-one	<chem>CCC(C)=O</chem>	Cétone	apprentissage	272,8	5,8
methyl methacrylate	<chem>COC(=O)C(C)=C</chem>	Ester	apprentissage	289,6	5,5
dioctyladipate	<chem>CCCCCCCCOC(=O)CCCCC(=O)OCCCCCCCC</chem>	Ester	apprentissage	474,2	26,0
butyl butanoate	<chem>CCCCOC(=O)CCC</chem>	Ester	apprentissage	327,6	1,1
1-methyl-4-(prop-1-en-2-yl)cyclohex-1-ene	<chem>CC(=C)C1CCC(C)=CC1</chem>	Cyclique	apprentissage	312,9	5,3
1-phenyl-1-propanol	<chem>CCC(O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	370,1	7,0
2-ethyl hexanol	<chem>CCCC(CC)CO</chem>	Alcool	apprentissage	347,2	1,2
phthalic acid, diisobutyl ester	<chem>CC(C)COC(=O)c1ccccc1C(=O)OCC(C)C</chem>	Aromatique (Oxygéné)	apprentissage	438,5	4,4
pentaerythritol	<chem>OCC(CO)(CO)CO</chem>	Alcool	apprentissage	526,2	7,0
1,2,3,4-tetramethylbenzene	<chem>Cc1ccc(C)c(C)c1C</chem>	Aromatique	apprentissage	335,3	5,9
cyclohex-3-enylmethan-1-ol	<chem>OCC1CCC=CC1</chem>	Cyclique (Oxygéné)	apprentissage	344,7	1,5
2,4,6-trimethyl-1,3,5-trioxane	<chem>CC1OC(C)OC(C)O1</chem>	Cyclique (Oxygéné)	apprentissage	302,0	7,0
2,6-xylénol	<chem>Cc1cccc(C)c1O</chem>	Aromatique (Oxygéné)	apprentissage	353,1	7,0
2-methylpropyl propanoate	<chem>CCC(=O)OCC(C)C</chem>	Ester	apprentissage	311,6	8,4
ethyl methacrylate	<chem>CCOC(=O)C(C)=C</chem>	Ester	apprentissage	296,9	3,8
2,5-dimethyl-2,4-hexadiene	<chem>CC(C)=CC=C(C)C</chem>	Alcène	apprentissage	285,0	12,2
methyl (9Z,12Z)-octadec-9,12-dienoate	<chem>CCCCC=CCC=CCCCCCCC(=O)OC</chem>	Ester	apprentissage	382,0	4,1
2-ethylhexanal	<chem>CCCC(CC)C=O</chem>	Aldehyde	apprentissage	322,2	5,0
3-methylnonane	<chem>CCCCCC(C)CC</chem>	Paraffine	apprentissage	312,8	1,8

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
3-ethyl-2,3,4-trimethylpentane	<chem>CCC(C)(C(C)C)C(C)C</chem>	Paraffine	apprentissage	304,4	0,4
3-ethyl-5-methylheptane	<chem>CCC(C)CC(CC)CC</chem>	Paraffine	apprentissage	310,6	6,6
2,3,5-trimethylheptane	<chem>CCC(C)CC(C)C(C)C</chem>	Paraffine	apprentissage	307,1	3,1
2-methyl-1-benzofuran	<chem>Cc1cc2ccccc2o1</chem>	Aromatique (Oxygéné)	apprentissage	347,1	7,0
1,3,5-triethylbenzene	<chem>CCc1cc(CC)cc(CC)c1</chem>	Aromatique	apprentissage	352,4	1,7
5-methylheptan-3-ol	<chem>CCC(C)CC(O)CC</chem>	Alcool	apprentissage	341,1	14,1
2-(tert-butylperoxy)-2-methylpropane	<chem>CC(C)(C)OOC(C)(C)C</chem>	Peroxyde	apprentissage	298,0	7,0
3,3,4,4-tetramethylhexane	<chem>CCC(C)(C)C(C)(C)CC</chem>	Paraffine	apprentissage	301,7	2,3
2,2,3,3-tetramethylpentane	<chem>CCC(C)(C)C(C)(C)C</chem>	Paraffine	apprentissage	285,1	3,9
terpinolene	<chem>CC(C)=C1CCC(C)=CC1</chem>	Cyclique	apprentissage	323,9	12,9
1-butene	<chem>CCC=C</chem>	Alcène	apprentissage	211,6	18,6
4-methylnonane	<chem>CCCCC(C)CCC</chem>	Paraffine	apprentissage	312,8	1,8
methyl 2-hydroxypropanoate	<chem>COC(=O)C(C)O</chem>	Ester	apprentissage	321,6	0,6
cyclononane	<chem>C1CCCCCCC1</chem>	Cyclique	apprentissage	308,2	7,8
hexanal	<chem>CCCCCC=O</chem>	Aldehyde	apprentissage	297,6	2,4
diisodecyl phthalate	<chem>CC(C)CCCCCOC(=O)c1cccc1C(=O)OCC CCCCC(C)C</chem>	Aromatique (Oxygéné)	apprentissage	498,4	7,0
2-heptanol	<chem>CCCCCC(C)O</chem>	Alcool	apprentissage	333,5	1,3
3-methylhex-1-ene	<chem>CCCC(C)C=C</chem>	Alcène	apprentissage	261,3	5,8
ethyl undec-10-enoate	<chem>CCOC(=O)CCCCCCCCC=C</chem>	Ester	apprentissage	381,3	4,7
2-ethoxyethanol	<chem>CCOCCO</chem>	Ether	apprentissage	319,7	3,7
3-methylbutan-2-one	<chem>CC(C)C(C)=O</chem>	Cétone	apprentissage	281,1	1,9
cis-4-octene	<chem>CCCC=CCCC</chem>	Alcène	apprentissage	281,1	8,9
1-undecene	<chem>CCCCCCCCC=C</chem>	Alcène	apprentissage	328,7	1,7
methyl (9Z)-hexadec-9-enoate	<chem>CCCCC=C(C)CCCCC(=O)OC</chem>	Ester	apprentissage	386,3	0,1
2,2-dimethylpropyl propanoate	<chem>CCC(=O)OCC(C)(C)C</chem>	Ester	apprentissage	320,9	13,1
3,5,5-trimethylcyclohex-2-en-1-one	<chem>CC1=CC(=O)CC(C)(C)C1</chem>	Cyclique (Oxygéné)	apprentissage	350,2	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
3-methylbut-3-en-2-one	<chem>CC(=C)C(C)=O</chem>	Cétone	apprentissage	287,2	7,0
cis-2-hexene	<chem>CCCC=CC</chem>	Alcène	apprentissage	248,7	1,7
1-heptene	<chem>CCCCCC=C</chem>	Alcène	apprentissage	263,7	1,5
1,4-cyclohexadiene	<chem>C1C=CCC=C1</chem>	Cyclique	apprentissage	266,2	4,1
heptane	<chem>CCCCCCC</chem>	Paraffine	apprentissage	267,7	1,3
methyl but-3-enoate	<chem>COC(=O)CC=C</chem>	Ester	apprentissage	289,8	3,2
ethyl undecanoate	<chem>CCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	383,6	2,4
pentadecane	<chem>CCCCCCCCCCCCC</chem>	Paraffine	apprentissage	383,8	4,2
2,3,3,4-tetramethylpentane	<chem>CC(C)C(C)(C)C(C)C</chem>	Paraffine	apprentissage	287,7	16,3
ethyl butanoate	<chem>CCCC(=O)OCC</chem>	Ester	apprentissage	300,6	3,4
methyl (3E)-pent-3-enoate	<chem>COC(=O)CC=CC</chem>	Ester	apprentissage	302,8	5,8
1-trans-3,5-trimethylcyclohexane	<chem>CC1CC(C)CC(C)C1</chem>	Cyclique	apprentissage	298,4	6,3
3-pentanol	<chem>CCC(O)CC</chem>	Alcool	apprentissage	303,9	9,1
ethyl 4-oxopentanoate	<chem>CCOC(=O)CCC(C)=O</chem>	Ester	apprentissage	366,3	3,1
tert-butyl 2-methylacrylate	<chem>CC(=C)C(=O)OC(C)(C)C</chem>	Ester	apprentissage	302,2	2,0
3,3-dimethylbutan-1-ol	<chem>CC(C)(C)CCO</chem>	Alcool	apprentissage	315,5	5,5
1,2,3,4,5-pentamethyl benzene	<chem>Cc1cc(C)c(C)c(C)c1C</chem>	Aromatique	apprentissage	354,6	9,5
4-methylhex-1-ene	<chem>CCC(C)CC=C</chem>	Alcène	apprentissage	260,6	2,6
(2E)-4-methylpent-2-ene	<chem>CC=CC(C)C</chem>	Alcène	apprentissage	245,3	6,3
l-menthol	<chem>CC(C)C1CCC(C)CC1O</chem>	Cyclique (Oxygéné)	apprentissage	363,6	2,6
p-cresol	<chem>Cc1ccc(O)cc1</chem>	Aromatique (Oxygéné)	apprentissage	361,0	7,0
cis-2-pentene	<chem>CCC=CC</chem>	Alcène	apprentissage	232,4	6,4
5-vinylbicyclo(2.2.1)hept-2-ene	<chem>C=CC1CC2CC1C=C2</chem>	Cyclique	apprentissage	301,2	1,1
8-methyl-1-nonanol	<chem>CC(C)CCCCCCO</chem>	Alcool	apprentissage	371,5	6,1
1,5-pentanediol	<chem>OCCCCO</chem>	Alcool	apprentissage	400,1	2,0
3-methylpentan-2-one	<chem>CCC(C)C(C)=O</chem>	Cétone	apprentissage	294,7	9,5
cycloheptane	<chem>C1CCCCC1</chem>	Cyclique	apprentissage	276,8	9,7

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2-hydroxyethyl acrylate	<chem>OCCOC(=O)C=C</chem>	Ester	apprentissage	364,2	7,0
2,3,4-trimethylhexane	<chem>CCC(C)C(C)C(C)C</chem>	Paraffine	apprentissage	292,2	4,2
p-terphenyl	<chem>c1ccc(cc1)-c1ccc(cc1)-c1ccccc1</chem>	Aromatique	apprentissage	473,0	7,0
hexylbenzene	<chem>CCCCCCc1ccccc1</chem>	Aromatique	apprentissage	349,7	3,5
o-ethyltoluene	<chem>CCc1ccccc1C</chem>	Aromatique	apprentissage	312,2	0,0
phenanthrene	<chem>c1ccc2c(c1)ccc1ccccc21</chem>	Aromatique	apprentissage	426,6	17,4
mesitylene	<chem>Cc1cc(C)cc(C)c1</chem>	Aromatique	apprentissage	314,0	3,5
1-octadecene	<chem>CCCCCCCCCCCCCCCC=C</chem>	Alcène	apprentissage	413,1	8,1
ethyl eicosanoate	<chem>CCCCCCCCCCCCCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	437,0	7,0
2,2,3-trimethylbutane	<chem>CC(C)C(C)(C)C</chem>	Paraffine	apprentissage	254,6	5,6
nonanal	<chem>CCCCCCCCC=O</chem>	Aldehyde	apprentissage	343,1	7,0
methoxydihydropyran	<chem>COC1CCC=CO1</chem>	Cyclique (Oxygéné)	apprentissage	294,7	5,6
n-nonylbenzene	<chem>CCCCCCCCCc1ccccc1</chem>	Aromatique	apprentissage	383,7	11,7
2,3,5-trimethylhexane	<chem>CC(C)CC(C)C(C)C</chem>	Paraffine	apprentissage	290,6	2,6
methyl tridecanoate	<chem>CCCCCCCCCCCCC(=O)OC</chem>	Ester	apprentissage	380,8	5,2
anisole	<chem>COc1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	318,2	7,0
propan-2-one	<chem>CC(C)=O</chem>	Cétone	apprentissage	261,1	6,1
1,2,4-trimethylbenzene	<chem>Cc1ccc(C)c(C)c1</chem>	Aromatique	apprentissage	314,5	4,2
2-methylnaphthalene	<chem>Cc1ccc2ccccc2c1</chem>	Aromatique	apprentissage	364,0	6,1
cis-1,2-dimethylcyclopentane	<chem>CC1CCCC1C</chem>	Cyclique	apprentissage	266,9	2,1
1,2-ethanediol	<chem>OCCO</chem>	Alcool	apprentissage	373,5	9,7
2-ethyl-1,4-dimethylbenzene	<chem>CCc1cc(C)ccc1C</chem>	Aromatique	apprentissage	327,7	1,3
dimethoxymethane	<chem>COCOC</chem>	Ether	apprentissage	262,0	7,0
methyl (11E)-octadec-11-enoate	<chem>CCCCC=CCCCCCCCC(=O)OC</chem>	Ester	apprentissage	390,9	55,9
dimethyl phthalate	<chem>COC(=O)c1ccccc1C(=O)OC</chem>	Aromatique (Oxygéné)	apprentissage	412,0	7,0
2,2,4-trimethyl-1,3-pentanediol	<chem>CC(C)C(O)C(C)(C)CO</chem>	Alcool	apprentissage	390,1	7,0
1,5,9-cyclododecatriene	<chem>C1CC=CCCC=CCCC=C1</chem>	Cyclique	apprentissage	353,2	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
tert-butylbenzene	<chem>CC(C)(C)c1ccccc1</chem>	Aromatique	apprentissage	320,9	12,3
dimethyl-1,4-cyclohexanedicarboxylate	<chem>COC(=O)C1CCC(CC1)C(=O)OC</chem>	Cyclique (Oxygéné)	apprentissage	387,5	4,3
allyl alcohol	<chem>OCC=C</chem>	Alcool	apprentissage	292,6	1,4
1,2-diphenylethane	<chem>C(Cc1ccccc1)c1ccccc1</chem>	Aromatique	apprentissage	401,5	0,5
ethoxyethane	<chem>CCOCC</chem>	Ether	apprentissage	239,2	11,0
ethyl pentanoate	<chem>CCCCC(=O)OCC</chem>	Ester	apprentissage	314,1	7,1
2-methyl-5-propan-2-ylcyclohexa-1,3-diene	<chem>CC(C)C1CC=C(C)C=C1</chem>	Cyclique	apprentissage	316,3	5,7
indan	<chem>C1Cc2ccccc2C1</chem>	Aromatique	apprentissage	318,1	3,5
2,3-dimethylbutane-2,3-diol	<chem>CC(C)(O)C(C)(C)O</chem>	Alcool	apprentissage	343,0	7,0
methoxyethene	<chem>COC=C</chem>	Ether	apprentissage	224,1	7,0
dimethyl maleate	<chem>COC(=O)C=CC(=O)OC</chem>	Ester	apprentissage	369,9	1,7
3,4-dimethylhexane	<chem>CCC(C)C(C)CC</chem>	Paraffine	apprentissage	278,9	1,9
alpha-methylstyrene	<chem>CC(=C)c1ccccc1</chem>	Aromatique	apprentissage	312,6	0,5
ethyl 2-methylpropanoate	<chem>CCOC(=O)C(C)C</chem>	Ester	apprentissage	294,0	7,0
2-(2-pentoxyethoxy)ethanol	<chem>CCCCOCCOCCO</chem>	Ether	apprentissage	386,6	3,6
1,1-dimethylcyclohexane	<chem>CC1(C)CCCCC1</chem>	Cyclique	apprentissage	282,2	2,2
2-ethylphenol	<chem>CCc1ccccc1O</chem>	Aromatique (Oxygéné)	apprentissage	362,0	10,9
(2E)-but-2-enal	<chem>CC=CC=O</chem>	Aldehyde	apprentissage	274,2	7,0
4-methylpent-3-en-2-one	<chem>CC(C)=CC(C)=O</chem>	Cétone	apprentissage	308,0	7,0
nonylphenol	<chem>CCCCCCCCCc1ccccc1O</chem>	Aromatique (Oxygéné)	apprentissage	418,0	7,0
2,2,4-trimethylhexane	<chem>CCC(C)CC(C)(C)C</chem>	Paraffine	apprentissage	286,6	1,4
alpha-methylbenzyl alcohol	<chem>CC(O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	362,4	4,2
methyl (3E)-hex-3-enoate	<chem>CCC=CCC(=O)OC</chem>	Ester	apprentissage	313,7	5,3
3-ethyl-2,2-dimethylhexane	<chem>CCCC(CC)C(C)(C)C</chem>	Paraffine	apprentissage	304,6	6,4
propylbenzene	<chem>CCCc1ccccc1</chem>	Aromatique	apprentissage	309,5	6,3
1-ethyl-2,4-dimethylbenzene	<chem>CCc1ccc(C)cc1C</chem>	Aromatique	apprentissage	327,7	2,3
4-ethyloctane	<chem>CCCCC(CC)CCC</chem>	Paraffine	apprentissage	313,6	0,4

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2,6-dimethyloctane	<chem>CCC(C)CCCC(C)C</chem>	Paraffine	apprentissage	309,0	2,0
ethyl (2E)-oct-4-enoate	<chem>CCCC=CCCC(=O)OCC</chem>	Ester	apprentissage	345,0	7,0
isopropylcyclopentane	<chem>CC(C)C1CCCC1</chem>	Cyclique	apprentissage	283,1	3,9
3-methyl-1,2-butadiene	<chem>CC(C)=C=C</chem>	Alcène	apprentissage	268,1	7,0
gamma-valerolactone	<chem>CC1CCC(=O)O1</chem>	Cyclique (Oxygéné)	apprentissage	314,6	39,4
2-oxopropanoic acid	<chem>CC(=O)C(O)=O</chem>	Acide carboxylique	apprentissage	362,1	7,0
dipropylene glycol propyl ether	<chem>CCCOCC(C)OCC(C)O</chem>	Ether	apprentissage	365,9	5,1
dibutyl sebacate	<chem>CCCCOC(=O)CCCCCCCC(=O)OCCCC</chem>	Ester	apprentissage	458,1	7,0
dihexyl phthalate	<chem>CCCCCOC(=O)c1cccc1C(=O)OCCCCC</chem>	Aromatique (Oxygéné)	apprentissage	466,2	0,1
2,5,8,11,14-pentaoxapentadecane	<chem>COCCOCCOCCOCCOC</chem>	Ether	apprentissage	421,1	7,0
methyl hex-5-enoate	<chem>COC(=O)CCCC=C</chem>	Ester	apprentissage	314,6	4,4
1-methyl-3-propylbenzene	<chem>CCc1cccc(C)c1</chem>	Aromatique	apprentissage	324,4	0,6
5-methylnonane	<chem>CCCC(C)CCCC</chem>	Paraffine	apprentissage	312,8	0,8
octane	<chem>CCCCCCCC</chem>	Paraffine	apprentissage	284,2	1,8
methyl (6E,9E,12E)-octadeca-6,9,12-trienoate	<chem>CCCCC=CCC=CCC=CCCCC(=O)OC</chem>	Ester	apprentissage	379,0	44,0
2,3,4-trimethylpent-2-ene	<chem>CC(C)C(C)=C(C)C</chem>	Alcène	apprentissage	282,0	7,0
4-hydroxyacetophenone	<chem>CC(=O)c1ccc(O)cc1</chem>	Aromatique (Oxygéné)	apprentissage	396,8	50,2
sec-butyl acetate	<chem>CCC(C)OC(C)=O</chem>	Ester	apprentissage	295,4	40,2
hexan-3-one	<chem>CCCC(=O)CC</chem>	Cétone	apprentissage	296,5	9,4
didecylphthalate	<chem>CCCCCCCCCOC(=O)c1cccc1C(=O)OCCCCCCCCC</chem>	Aromatique (Oxygéné)	apprentissage	498,0	7,0
m-diisopropylbenzene	<chem>CC(C)c1cccc(c1)C(C)C</chem>	Aromatique	apprentissage	351,2	2,2
benzyl benzoate	<chem>O=C(OCc1cccc1)c1cccc1</chem>	Aromatique (Oxygéné)	apprentissage	418,0	0,1
(ethenyloxy)ethene	<chem>C=COC=C</chem>	Ether	apprentissage	233,1	7,0
decylbenzene	<chem>CCCCCCCCCc1cccc1</chem>	Aromatique	apprentissage	393,3	13,3
1,2,3,5-tetramethylbenzene	<chem>Cc1cc(C)c(C)c(C)c1</chem>	Aromatique	apprentissage	335,0	1,5
glycerol triethanoate	<chem>CC(=O)OCC(COC(C)=O)OC(C)=O</chem>	Ester	apprentissage	404,0	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
furfural	<chem>O=Cc1ccco1</chem>	Aromatique (Oxygéné)	apprentissage	326,2	7,0
methyl methanoate	<chem>COC=O</chem>	Ester	apprentissage	254,6	0,6
2-methylpropyl butanoate	<chem>CCCC(=O)OCC(C)C</chem>	Ester	apprentissage	325,2	4,8
2,2,4,4-tetramethylpentane	<chem>CC(C)(C)CC(C)(C)C</chem>	Paraffine	apprentissage	281,0	5,0
2-ethyloxirane	<chem>CCC1CO1</chem>	Cyclique (Oxygéné)	apprentissage	244,2	7,0
2-methyl-2-butanol	<chem>CCC(C)(C)O</chem>	Alcool	apprentissage	293,5	0,5
isobutylbenzene	<chem>CC(C)Cc1ccccc1</chem>	Aromatique	apprentissage	320,8	4,4
3-ethyl-3-methylpentane	<chem>CCC(C)(CC)CC</chem>	Paraffine	apprentissage	276,2	0,2
1,2-bis(2-ethylhexyl) benzene-1,2-dicarboxylate	<chem>CCCC(CC)COC(=O)c1ccccc1C(=O)OCC(C)C</chem>	Aromatique (Oxygéné)	apprentissage	488,2	0,8
vinyl acetate	<chem>CC(=O)OC=C</chem>	Ester	apprentissage	272,0	7,0
o-cymene	<chem>CC(C)c1ccccc1C</chem>	Aromatique	apprentissage	324,9	6,1
1-propyl acrylate	<chem>CCCOC(=O)C=C</chem>	Ester	apprentissage	303,1	7,9
2,3-dimethyloctane	<chem>CCCCC(C)C(C)C</chem>	Paraffine	apprentissage	309,8	0,8
pentanal	<chem>CCCC=O</chem>	Aldehyde	apprentissage	281,8	3,4
2-methoxyethanol	<chem>COCCO</chem>	Ether	apprentissage	314,5	2,5
2,3,3-trimethylhexane	<chem>CCCC(C)(C)C(C)C</chem>	Paraffine	apprentissage	289,0	1,0
2-methylpent-2-ene	<chem>CCC=C(C)C</chem>	Alcène	apprentissage	249,7	0,3
non-1-yne	<chem>CCCCCCCC#C</chem>	Alcyne	apprentissage	303,5	2,6
2-ethylhexanoic acid	<chem>CCCC(CC)C(O)=O</chem>	Acide carboxylique	apprentissage	380,2	7,0
2-methylbut-1-ene	<chem>CCC(C)=C</chem>	Alcène	apprentissage	228,4	2,4
cyclooct-1,3,5,7-tetraene	<chem>C1=CC=CC=C1</chem>	Cyclique	apprentissage	300,9	5,9
methylcyclopentane	<chem>CC1CCCC1</chem>	Cyclique	apprentissage	252,6	6,6
1-ethoxybutane	<chem>CCCCOCC</chem>	Ether	apprentissage	269,0	8,2
prop-2-enal	<chem>C=CC=O</chem>	Aldehyde	apprentissage	254,1	7,0
prop-2-enylcyclopentane	<chem>C=CCC1CCCC1</chem>	Cyclique	apprentissage	282,7	2,3
2,2,5-trimethylhexane	<chem>CC(C)CCC(C)(C)C</chem>	Paraffine	apprentissage	285,8	0,2



Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
1,2-dimethylpropyl ethanoate	<chem>CC(C)C(C)OC(C)=O</chem>	Ester	apprentissage	307,9	8,1
furfuryl alcohol	<chem>OCC1CCCO1</chem>	Aromatique (Oxygéné)	apprentissage	337,9	0,2
oxolane	<chem>C1CCOC1</chem>	Cyclique (Oxygéné)	apprentissage	266,6	7,6
m-diethylbenzene	<chem>CCc1cccc(CC)c1</chem>	Aromatique	apprentissage	326,3	2,7
1-pentene	<chem>CCCC=C</chem>	Alcène	apprentissage	229,0	7,0
ethylbenzene	<chem>CCc1ccccc1</chem>	Aromatique	apprentissage	295,1	7,0
1,3-dimethylbenzene	<chem>Cc1cccc(C)c1</chem>	Aromatique	apprentissage	296,7	1,5
dec-1-yne	<chem>CCCCCCCC#C</chem>	Alcyne	apprentissage	316,2	7,0
pentylcyclohexane	<chem>CCCCC1CCCCC1</chem>	Cyclique	apprentissage	334,4	4,6
sec-butyl alcohol	<chem>CCC(C)O</chem>	Alcool	apprentissage	291,7	4,5
n-pentylbenzene	<chem>CCCCC1CCCCC1</chem>	Aromatique	apprentissage	336,8	1,3
tripropylene glycol monomethyl ether	<chem>COC(C)COC(C)COC(C)CO</chem>	Ether	apprentissage	391,0	7,0
cis-1-propenylbenzene	<chem>CC=Cc1ccccc1</chem>	Aromatique	apprentissage	317,3	6,3
nonan-2-one	<chem>CCCCCCCC(C)=O</chem>	Cétone	apprentissage	335,9	1,3
ethyl tetradecanoate	<chem>CCCCCCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	408,1	22,1
2-(propan-2-yloxy)propane	<chem>CC(C)OC(C)C</chem>	Ether	apprentissage	252,0	7,0
2,4-dimethylhexane	<chem>CCC(C)CC(C)C</chem>	Paraffine	apprentissage	277,0	6,2
methyl (4Z,7Z,10Z,13Z,16Z,19Z)-docosa-4,7,10,13,16,19-hexaenoate	<chem>COC(=O)CCCC=CCC=CCC=CCC=CCC=CCC=CC</chem>	Ester	apprentissage	373,0	7,0
propyl methanoate	<chem>CCCOC=O</chem>	Ester	apprentissage	277,9	7,9
2-methylprop-2-enoic acid	<chem>CC(=C)C(O)=O</chem>	Acide carboxylique	apprentissage	347,0	7,0
octanal	<chem>CCCCCCCC=O</chem>	Aldehyde	apprentissage	328,4	4,2
2,2,3,4-tetramethylhexane	<chem>CCC(C)C(C)C(C)C</chem>	Paraffine	apprentissage	302,4	1,6
1,2-propylene glycol	<chem>CC(O)CO</chem>	Alcool	apprentissage	365,0	7,0
2,2,5,5-tetramethylhexane	<chem>CC(C)(C)CCC(C)(C)C</chem>	Paraffine	apprentissage	296,8	7,8
2-propoxyethanol	<chem>CCCOCCO</chem>	Ether	apprentissage	329,1	7,0
undecyl methanoate	<chem>CCCCCCCCCCCOC=O</chem>	Ester	apprentissage	392,9	7,9

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2-(2-butoxyethoxy)ethan-1-ol	CCCCOCCOCCO	Ether	apprentissage	377,6	26,6
2,6,10,15,19,23-hexamethyltetracosane	CC(C)CCCC(C)CCCC(C)CCCC(C)CCCC(C)C	Paraffine	apprentissage	483,0	7,0
butyl acrylate	CCCCOC(=O)C=C	Ester	apprentissage	318,7	6,7
1-phenyl-2-propanol	CC(O)Cc1ccccc1	Aromatique (Oxygéné)	apprentissage	365,1	7,0
2-methylhexane	CCCCC(C)C	Paraffine	apprentissage	263,3	8,3
heptan-2-one	CCCCCC(C)=O	Cétone	apprentissage	312,1	0,1
bicyclohexyl	C1CCC(CC1)C1CCCCC1	Cyclique	apprentissage	348,9	1,7
2,2-dimethylheptane	CCCCC(C)(C)C	Paraffine	apprentissage	288,9	8,1
1-methylcyclohex-1-ene	CC1=CCCCC1	Cyclique	apprentissage	274,1	2,1
1,2-dimethoxyethane	COCCOC	Ether	apprentissage	276,7	5,2
octyl butanoate	CCCCCCCCOC(=O)CCC	Ester	apprentissage	375,0	3,0
1,2,4,5-tetramethylbenzene	Cc1cc(C)c(C)cc1C	Aromatique	apprentissage	335,0	7,0
2,4-pentadediol	CC(O)CC(C)O	Alcool	apprentissage	371,9	2,3
ethyl decanoate	CCCCCCCCC(=O)OCC	Ester	apprentissage	373,8	1,2
m-cresol	Cc1ccc(O)c1	Aromatique (Oxygéné)	apprentissage	361,0	7,0
butyl ethanoate	CCCCOC(C)=O	Ester	apprentissage	302,0	7,0
3-ethylphenol	CCc1ccc(O)c1	Aromatique (Oxygéné)	apprentissage	369,8	2,6
2-cyclohexyl-2-methylpropane	CC(C)(C)C1CCCCC1	Cyclique	apprentissage	310,6	4,4
cyclohexyl carboxylate	O=COC1CCCCC1	Cyclique (Oxygéné)	apprentissage	331,1	7,0
acetaldehyde	CC=O	Aldehyde	apprentissage	238,2	3,2
hepta-1,6-diene	C=CCCC=C	Alcène	apprentissage	268,3	5,3
1-octylbenzene	CCCCCCCCc1ccccc1	Aromatique	apprentissage	373,2	7,0
1-tert-butoxy-propan-2-ol	CC(O)COC(C)(C)C	Ether	apprentissage	315,7	1,4
2-ethyl-1-pentene	CCCC(=C)CC	Alcène	apprentissage	261,5	1,5
diallyl maleate	C=CCOC(=O)C=CC(=O)OCC=C	Ester	apprentissage	386,0	7,0
1-propoxy-2-propanol	CCCOCC(C)O	Ether	apprentissage	328,1	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
butylbenzene	<chem>CCCCc1ccccc1</chem>	Aromatique	apprentissage	323,4	0,2
2-phenyl-1-propanol	<chem>CC(CO)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	373,1	6,9
2,3-dimethyl-2-butene	<chem>CC(C)=C(C)C</chem>	Alcène	apprentissage	256,5	0,6
1-methyl-4-(propan-2-yl)cyclohexa-1,3-diene	<chem>CC(C)C1=CC=C(C)CC1</chem>	Cyclique	apprentissage	317,8	1,3
3-methylbut-1-ene	<chem>CC(C)C=C</chem>	Alcène	apprentissage	225,7	8,6
dimethyl propane-1,3-dioate	<chem>COC(=O)CC(=O)OC</chem>	Ester	apprentissage	356,2	7,0
2,4,4-trimethylhexane	<chem>CCC(C)(C)CC(C)C</chem>	Paraffine	apprentissage	287,4	0,6
4-ethyl-3-methylheptane	<chem>CCCC(CC)C(C)CC</chem>	Paraffine	apprentissage	311,3	2,7
hexanedioic acid	<chem>OC(=O)CCCCCC(=O)O</chem>	Acide carboxylique	apprentissage	443,0	7,0
methyl salicylate	<chem>COC(=O)c1ccccc1O</chem>	Aromatique (Oxygéné)	apprentissage	376,0	7,0
methoxyethane	<chem>CCOC</chem>	Ether	apprentissage	229,0	7,0
methyl acrylate	<chem>COC(=O)C=C</chem>	Ester	apprentissage	279,8	9,8
2,2,3-trimethylpentane	<chem>CCC(C)C(C)(C)C</chem>	Paraffine	apprentissage	272,0	2,0
2,2'-oxydiethanol	<chem>OCCOCCO</chem>	Ether	apprentissage	402,1	5,1
2-ethyl-1-butene	<chem>CCC(=C)CC</chem>	Alcène	apprentissage	245,8	1,4
nona-1,8-diene	<chem>C=CCCCCCC=C</chem>	Alcène	apprentissage	300,9	1,9
4-methylheptane	<chem>CCCC(C)CCC</chem>	Paraffine	apprentissage	280,9	1,8
3-methylpentane	<chem>CCC(C)CC</chem>	Paraffine	apprentissage	247,4	6,4
methyl 10-undecenoate	<chem>COC(=O)CCCCCCCCC=C</chem>	Ester	apprentissage	366,0	7,0
o-cresol	<chem>Cc1ccccc1O</chem>	Aromatique (Oxygéné)	apprentissage	353,3	0,7
methyl 2-methylpropanoate	<chem>COC(=O)C(C)C</chem>	Ester	apprentissage	283,1	7,0
3,3-dimethyl-1-butene	<chem>CC(C)(C)C=C</chem>	Alcène	apprentissage	238,2	7,0
2,3,4-trimethylpentan-1-ol	<chem>CC(C)C(C)C(C)CO</chem>	Alcool	apprentissage	339,8	6,8
undecyl ethanoate	<chem>CCCCCCCCCCCOC(=O)C</chem>	Ester	apprentissage	385,9	12,1
2,4-dimethyl-3-ethylpentane	<chem>CCC(C)(C)C(C)C</chem>	Paraffine	apprentissage	292,2	4,2
hexyl methanoate	<chem>CCCCCCOC=O</chem>	Ester	apprentissage	323,8	6,2
dicyclopentadiene	<chem>C1C=CC2C3CC(C=C3)C12</chem>	Cyclique	apprentissage	312,4	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
dipropylene glycol	<chem>CC(O)COCC(C)O</chem>	Ether	apprentissage	384,0	7,0
m-cymene	<chem>CC(C)c1cccc(C)c1</chem>	Aromatique	apprentissage	324,4	1,4
ethyl propanoate	<chem>CCOC(=O)CC</chem>	Ester	apprentissage	286,7	1,7
ethyl acrylate	<chem>CCOC(=O)C=C</chem>	Ester	apprentissage	289,0	6,9
1-hexene	<chem>CCCCCC=C</chem>	Alcène	apprentissage	246,4	0,6
1-ethylpropyl ethanoate	<chem>CCC(CC)OC(C)=O</chem>	Ester	apprentissage	311,0	7,0
isobutyl methacrylate	<chem>CC(C)COC(=O)C(C)=C</chem>	Ester	apprentissage	321,1	7,0
2-methyloctane	<chem>CCCCCCC(C)C</chem>	Paraffine	apprentissage	296,2	0,8
2-pentyne	<chem>CCC#CC</chem>	Alcyne	apprentissage	250,0	7,0
2,4-dimethylpentane	<chem>CC(C)CC(C)C</chem>	Paraffine	apprentissage	259,3	1,7
3-methylhexane	<chem>CCCC(C)CC</chem>	Paraffine	apprentissage	264,3	4,7
methyl butanoate	<chem>CCCC(=O)OC</chem>	Ester	apprentissage	289,6	2,6
4-methylpenta-1,3-diene	<chem>CC(C)=CC=C</chem>	Alcène	apprentissage	247,4	7,6
acetic acid	<chem>CC(O)=O</chem>	Acide carboxylique	apprentissage	315,9	0,1
2-methylnonane	<chem>CCCCCCCC(C)C</chem>	Paraffine	apprentissage	312,0	2,0
2-propanol	<chem>CC(C)O</chem>	Alcool	apprentissage	280,0	5,1
1-methylcyclopentene	<chem>CC1=CCCC1</chem>	Cyclique	apprentissage	251,0	5,0
tert-butyl butanoate	<chem>CCCC(=O)OC(C)(C)C</chem>	Ester	apprentissage	313,0	7,0
p-diisopropylbenzene	<chem>CC(C)c1ccc(cc1)C(C)C</chem>	Aromatique	apprentissage	351,2	2,2
2-methyl-2,4-pentandiol	<chem>CC(O)CC(C)(C)O</chem>	Alcool	apprentissage	365,3	0,7
(3E)-hexa-1,3-diene	<chem>CCC=CC=C</chem>	Alcène	apprentissage	248,2	6,0
1-(pentyloxy)pentane	<chem>CCCCOCCCC</chem>	Ether	apprentissage	329,4	0,8
cyclohexanone	<chem>O=C1CCCCC1</chem>	Cyclique (Oxygéné)	apprentissage	310,0	7,0
3,5-dimethylheptane	<chem>CCC(C)CC(C)CC</chem>	Paraffine	apprentissage	294,1	6,1
3-methylbutyl methanoate	<chem>CC(C)CCOC=O</chem>	Ester	apprentissage	307,0	7,0
2-octanol	<chem>CCCCCCC(C)O</chem>	Alcool	apprentissage	346,8	11,2
dihexyl adipate	<chem>CCCCCCOC(=O)CCCCC(=O)OCCCCCC</chem>	Ester	apprentissage	458,1	5,9

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2,3-dimethylhexane	<chem>CCCC(C)C(C)C</chem>	Paraffine	apprentissage	277,9	0,6
dibutyl maleate	<chem>CCCCOC(=O)C=CC(=O)OCCCC</chem>	Ester	apprentissage	420,0	7,0
1-octene	<chem>CCCCCCC=C</chem>	Alcène	apprentissage	280,7	0,3
2-methylpent-2-enal	<chem>CCC=C(C)C=O</chem>	Aldehyde	apprentissage	303,7	0,4
3,3-dimethylbutan-2-one	<chem>CC(=O)C(C)(C)C</chem>	Cétone	apprentissage	286,0	0,9
3-methylpentanol	<chem>CCC(C)CCO</chem>	Alcool	apprentissage	324,2	7,0
2,3-dimethyl hexene	<chem>CCCC(C)C(C)=C</chem>	Alcène	apprentissage	274,3	6,7
benzene	<chem>c1ccccc1</chem>	Aromatique	apprentissage	261,8	0,2
4-methylideneoxetan-2-one	<chem>C=C1CC(=O)O1</chem>	Cyclique (Oxygéné)	apprentissage	300,0	7,0
1-propoxypropane	<chem>CCCOCCC</chem>	Ether	apprentissage	268,1	26,0
1-(ethenyloxy)butane	<chem>CCCCOC=C</chem>	Ether	apprentissage	265,0	7,0
(1-methyl ethyl)cyclohexane	<chem>CC(C)C1CCCCC1</chem>	Cyclique	apprentissage	302,2	6,0
(2R,3R,4R,5S)-hexane-1,2,3,4,5,6-hexol	<chem>OCC(O)C(O)C(O)C(O)CO</chem>	Alcool	apprentissage	429,0	7,0
2-hexanol	<chem>CCCCC(C)O</chem>	Alcool	apprentissage	319,8	5,8
tert-butyl alcohol	<chem>CC(C)(C)O</chem>	Alcool	apprentissage	280,7	3,6
2-methylprop-1-ene	<chem>CC(C)=C</chem>	Alcène	apprentissage	211,6	14,6
butanoic acid	<chem>CCCC(O)=O</chem>	Acide carboxylique	apprentissage	339,3	5,7
benzoic acid	<chem>OC(=O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	401,2	7,0
1-ethylnaphthalene	<chem>CCc1cccc2ccccc12</chem>	Aromatique	apprentissage	377,2	7,0
methyl pentadecanoate	<chem>CCCCCCCCCCCCCCCC(=O)OC</chem>	Ester	apprentissage	393,0	7,0
2,2,4-trimethylpentan-1-ol	<chem>CC(C)CC(C)(C)CO</chem>	Alcool	apprentissage	333,8	3,8
1,2-diethoxyethane	<chem>CCOCCOCC</chem>	Ether	apprentissage	292,2	7,9
1,1-diphenylethane	<chem>CC(c1ccccc1)c1ccccc1</chem>	Aromatique	apprentissage	406,0	3,9
3-methylbutyl ethanoate	<chem>CC(C)CCOC(C)=O</chem>	Ester	apprentissage	312,6	14,4
dibutyl phthalate	<chem>CCCCOC(=O)c1ccccc1C(=O)OCCCC</chem>	Aromatique (Oxygéné)	apprentissage	436,6	6,5
biphenyl	<chem>c1ccc(cc1)-c1ccccc1</chem>	Aromatique	apprentissage	383,8	2,2
ethynylbenzene	<chem>C#Cc1ccccc1</chem>	Aromatique	apprentissage	301,2	2,9

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2,7-dimethyloctane	<chem>CC(C)CCCC(C)C</chem>	Paraffine	apprentissage	308,3	5,7
diethyl malonate	<chem>CCOC(=O)CC(=O)OCC</chem>	Ester	apprentissage	359,4	6,7
1,2-dimethylbenzene	<chem>Cc1ccccc1C</chem>	Aromatique	apprentissage	297,3	7,3
isobutyl acrylate	<chem>CC(C)COC(=O)C=C</chem>	Ester	apprentissage	314,2	10,0
(2E)-3-methylpent-2-ene	<chem>CCC(C)=CC</chem>	Alcène	apprentissage	250,6	4,5
3-methylbenzenemethanol	<chem>Cc1cccc(CO)c1</chem>	Aromatique (Oxygéné)	apprentissage	379,7	1,6
2-hydroxyacetophenone	<chem>CC(=O)c1ccccc1O</chem>	Aromatique (Oxygéné)	apprentissage	388,8	17,7
isopropyl acetate	<chem>CC(C)OC(C)=O</chem>	Ester	apprentissage	280,4	5,4
(4-methylcyclohexyl)-methan-1-ol	<chem>CC1CCC(CO)CC1</chem>	Cyclique (Oxygéné)	apprentissage	350,4	3,6
pentane	<chem>CCCCC</chem>	Paraffine	apprentissage	233,7	0,5
(2E)-but-2-enoic acid	<chem>CC=CC(O)=O</chem>	Acide carboxylique	apprentissage	357,2	10,8
trans-3-heptene	<chem>CCCC=CCC</chem>	Alcène	apprentissage	265,3	1,9
3-methyl-2-butanol	<chem>CC(C)C(C)O</chem>	Alcool	apprentissage	299,3	0,2
triethylene glycol ethyl ether	<chem>CCOCCOCCOCCO</chem>	Ether	apprentissage	402,6	5,6
2-pentanol	<chem>CCCC(C)O</chem>	Alcool	apprentissage	305,8	8,2
2-methylbenzenemethanol	<chem>Cc1ccccc1CO</chem>	Aromatique (Oxygéné)	apprentissage	380,0	2,9
1-nonene	<chem>CCCCCCC=C</chem>	Alcène	apprentissage	297,3	0,7
8-oxatricyclo[7.4.0.0 <sup>2,7</sup> ]trideca-1(13),2,4,6,9,11-hexaene	<chem>c1ccc2c(c1)oc1ccccc21</chem>	Aromatique (Oxygéné)	apprentissage	396,2	7,0
isophthalic acid, dimethyl ester	<chem>COC(=O)c1cccc(c1)C(=O)OC</chem>	Aromatique (Oxygéné)	apprentissage	412,1	1,0
3,3,4-trimethylhexane	<chem>CCC(C)C(C)(C)CC</chem>	Paraffine	apprentissage	289,8	1,8
2-hexyne	<chem>CCCC#CC</chem>	Alcyne	apprentissage	259,4	2,8
methyl octadecanoate	<chem>CCCCCCCCCCCCCCCC(=O)OC</chem>	Ester	apprentissage	405,8	17,8
heptan-3-one	<chem>CCCCC(=O)CC</chem>	Cétone	apprentissage	309,0	1,2
cis-1,4-dimethylcyclohexane	<chem>CC1CCC(C)CC1</chem>	Cyclique	apprentissage	286,0	7,0
2,2-dimethyl-3-methylenenorbornane	<chem>CC1(C)C2CCC(C2)C1=C</chem>	Cyclique	apprentissage	309,3	0,2
1-(2-propenoxy)-2-propanol	<chem>CC(O)COCC=C</chem>	Ether	apprentissage	331,5	4,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
1,2-butanediol	<chem>CCC(O)CO</chem>	Alcool	apprentissage	372,2	9,0
2,4,6-triméthylheptane	<chem>CC(C)CC(C)CC(C)C</chem>	Paraffine	apprentissage	305,6	1,6
1,3-diméthyladamantane	<chem>CC12CC3CC(C1)CC(C)(C3)C2</chem>	Cyclique	apprentissage	332,1	7,0
2,2-diméthylhexane	<chem>CCCCC(C)(C)C</chem>	Paraffine	apprentissage	272,5	3,5
1-butoxybutane	<chem>CCCCOCCCC</chem>	Ether	apprentissage	299,4	1,3
pentyl ethanoate	<chem>CCCCCOC(C)=O</chem>	Ester	apprentissage	315,6	19,5
2,2-diméthyl-3-éthylpentane	<chem>CCC(CC)C(C)(C)C</chem>	Paraffine	apprentissage	289,0	3,0
octa-1,7-diene	<chem>C=CCCCC=C</chem>	Alcène	apprentissage	284,8	2,8
méthyl 3-oxobutanoate	<chem>COC(=O)CC(C)=O</chem>	Ester	apprentissage	350,0	7,0
octyl ethanoate	<chem>CCCCCCCCOC(C)=O</chem>	Ester	apprentissage	353,7	5,4
cis-2-octene	<chem>CCCCC=CC</chem>	Alcène	apprentissage	280,8	13,2
dodecane	<chem>CCCCCCCCCCCC</chem>	Paraffine	apprentissage	345,4	1,6
triéthylène glycol butyl ether	<chem>CCCCOCCOCCOCCO</chem>	Ether	apprentissage	414,5	2,0
1-méthylvinyl acetate	<chem>CC(=C)OC(C)=O</chem>	Ester	apprentissage	286,1	3,1
1-heptadecene	<chem>CCCCCCCCCCCCCCC=C</chem>	Alcène	apprentissage	403,7	4,3
éthyl nonanoate	<chem>CCCCCCCCC(=O)OCC</chem>	Ester	apprentissage	363,2	3,8
diéthyl maleate	<chem>CCOC(=O)C=CC(=O)OCC</chem>	Ester	apprentissage	368,1	2,0
pyrene	<chem>c1cc2ccc3cccc4ccc(c1)c2c34</chem>	Aromatique	apprentissage	466,6	5,4
(2S)-butane-1,2,4-triol	<chem>OCCC(O)CO</chem>	Alcool	apprentissage	448,7	63,7
nonyl méthanoate	<chem>CCCCCCCCCOC=O</chem>	Ester	apprentissage	367,0	7,0
tétradécyl méthanoate	<chem>CCCCCCCCCCCCCOC=O</chem>	Ester	apprentissage	426,0	7,0
2-éthylhexyl acetate	<chem>CCCCC(CC)COC(C)=O</chem>	Ester	apprentissage	351,8	7,6
éthyl-3-éthoxypropionate	<chem>CCOCCC(=O)OCC</chem>	Ester	apprentissage	329,8	1,3
2,6-diméthyl-1,3-dioxan-4-yl acetate	<chem>CC1CC(OC(C)=O)OC(C)O1</chem>	Cyclique (Oxygéné)	apprentissage	332,1	7,0
benzyl formate	<chem>O=COCC1CCCCC1</chem>	Aromatique (Oxygéné)	apprentissage	348,0	7,0
2-(2-éthoxyéthoxy)éthanol	<chem>CCOCCOCCO</chem>	Ether	apprentissage	360,9	4,5
2,6-di-tert-butyl-p-cresol	<chem>Cc1cc(c(O)c(c1)C(C)(C)C)C(C)(C)C</chem>	Aromatique (Oxygéné)	apprentissage	385,6	4,6

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
cyclopentanone	<chem>O=C1CCCC1</chem>	Cyclique (Oxygéné)	apprentissage	293,3	5,7
(2,2-diméthylpropyl)benzene	<chem>CC(C)(C)Cc1ccccc1</chem>	Aromatique	apprentissage	330,0	7,0
1-méthoxybutane	<chem>CCCCOC</chem>	Ether	apprentissage	256,7	6,4
ethyl methanoate	<chem>CCOC=O</chem>	Ester	apprentissage	260,1	7,0
octyl methanoate	<chem>CCCCCCCCOC=O</chem>	Ester	apprentissage	353,1	0,1
1,3-diméthylcyclohex-1-ene	<chem>CC1CCCC(C)=C1</chem>	Cyclique	apprentissage	286,3	1,3
2-(2-éthoxyéthoxy)ethyl acetate	<chem>CCOCCOCCOC(C)=O</chem>	Ester	apprentissage	365,9	3,2
1-tétradécanol	<chem>CCCCCCCCCCCCCO</chem>	Alcool	apprentissage	413,1	4,9
ethylene glycol diacrylate	<chem>C=CC(=O)OCCOC(=O)C=C</chem>	Ester	apprentissage	371,9	1,3
2,4-diméthylpentan-3-one	<chem>CC(C)C(=O)C(C)C</chem>	Cétone	apprentissage	295,1	7,0
3,3-diméthyl-octane	<chem>CCCCC(C)(C)CC</chem>	Paraffine	apprentissage	306,3	7,7
4-oxopentanoic acid	<chem>CC(=O)CCC(O)=O</chem>	Acide carboxylique	apprentissage	403,0	7,0
(1-méthylheptyl)benzene	<chem>CCCCCCC(C)c1ccccc1</chem>	Aromatique	apprentissage	372,3	0,7
n-butyl benzoate	<chem>CCCCOC(=O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	376,2	3,0
2,4,4-triméthyl-2-pentène	<chem>CC(C)=CC(C)(C)C</chem>	Alcène	apprentissage	272,6	0,5
butyl propanoate	<chem>CCCCOC(=O)CC</chem>	Ester	apprentissage	314,3	8,9
propyl methacrylate	<chem>CCCOC(=O)C(C)=C</chem>	Ester	apprentissage	310,4	11,6
undécanal	<chem>CCCCCCCCCCC=O</chem>	Aldehyde	apprentissage	370,8	1,7
2-(2-propoxyéthoxy)ethanol	<chem>CCCOCCOCCO</chem>	Ether	apprentissage	368,4	3,6
phenylmethan-1-ol	<chem>OCc1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	364,2	9,8
3,3,5-triméthylheptane	<chem>CCC(C)CC(C)(C)CC</chem>	Paraffine	apprentissage	303,9	0,1
2-méthylpropane-2-peroxyde	<chem>CC(C)(C)OO</chem>	Peroxyde	apprentissage	296,4	3,4
tetraethylene glycol	<chem>OCCOCCOCCOCCO</chem>	Ether	apprentissage	459,5	9,5
3-éthylhexane	<chem>CCCC(CC)CC</chem>	Paraffine	apprentissage	281,9	3,9
prop-2-enylcyclohexane	<chem>C=CCC1CCCCC1</chem>	Cyclique	apprentissage	302,2	0,2
methyl (5Z,8Z,11Z,14Z,17Z)-eicosa-5,8,11,14,17-pentaénoate	<chem>CCC=CCC=CCC=CCC=CCC=CCCCC(=O)OC</chem>	Ester	apprentissage	377,9	0,9



Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
undecyl propanoate	<chem>CCCCCCCCCOC(=O)CC</chem>	Ester	apprentissage	394,2	15,8
methyl pent-4-enoate	<chem>COC(=O)CCC=C</chem>	Ester	apprentissage	302,4	0,4
1,3,5-trioxane	<chem>C1OCOCO1</chem>	Cyclique (Oxygéné)	apprentissage	311,2	7,0
methyl (11Z)-eicos-11-enoate	<chem>CCCCCCCC=CCCCCCCCC(=O)OC</chem>	Ester	apprentissage	393,0	7,0
2,2,4,4,6,8,8-heptamethylnonane	<chem>CC(CC(C)(C)C)CC(C)(C)CC(C)(C)C</chem>	Paraffine	apprentissage	372,3	4,1
2,3-dimethylpent-1-ene	<chem>CCC(C)C=C</chem>	Alcène	apprentissage	258,9	2,9
2,4-xylénol	<chem>Cc1ccc(O)c(C)c1</chem>	Aromatique (Oxygéné)	apprentissage	367,0	1,2
n-undecylbenzene	<chem>CCCCCCCCCc1ccccc1</chem>	Aromatique	apprentissage	402,1	14,9
diethyl carbonate	<chem>CCOC(=O)OCC</chem>	Ester	apprentissage	297,8	0,2
2,4,4-trimethylpentan-1-ol	<chem>CC(CO)CC(C)(C)C</chem>	Alcool	apprentissage	335,2	2,2
propyl benzoate	<chem>CCCOC(=O)c1ccccc1</chem>	Aromatique (Oxygéné)	apprentissage	364,2	7,0
3,3-diethylpentane	<chem>CCC(CC)(CC)CC</chem>	Paraffine	apprentissage	294,0	0,0
tetrahydrofurfuryl alcohol	<chem>OCC1CCCO1</chem>	Cyclique (Oxygéné)	apprentissage	349,1	6,0
allyl methacrylate	<chem>CC(=C)C(=O)OCC=C</chem>	Ester	apprentissage	313,1	7,0
2,3,3-trimethylpentane	<chem>CCC(C)(C)C(C)C</chem>	Paraffine	apprentissage	272,9	0,1
allyl acetate	<chem>CC(=O)OCC=C</chem>	Ester	apprentissage	290,3	5,1
1,3-benzenediol	<chem>Oc1cccc(O)c1</chem>	Aromatique (Oxygéné)	apprentissage	396,1	3,9
2-methylpropyl 2-methylpropanoate	<chem>CC(C)COC(=O)C(C)C</chem>	Ester	apprentissage	318,6	7,5
ethyl (2E,4E)-hexa-2,4-dienoate	<chem>CCOC(=O)C=CC=CC</chem>	Ester	apprentissage	335,0	7,0
propanoic acid	<chem>CCC(O)=O</chem>	Acide carboxylique	apprentissage	325,8	2,2
2,6-dimethylheptan-4-one	<chem>CC(C)CC(=O)CC(C)C</chem>	Cétone	apprentissage	327,8	5,7
ethylcyclopentane	<chem>CCC1CCCC1</chem>	Cyclique	apprentissage	270,2	1,2
acetol	<chem>CC(=O)CO</chem>	Cétone	apprentissage	322,0	7,0
guaiacol	<chem>COc1ccccc1O</chem>	Aromatique (Oxygéné)	apprentissage	362,0	7,0
benzene-1,4-diAcide carboxylique	<chem>OC(=O)c1ccc(cc1)C(O)=O</chem>	Aromatique (Oxygéné)	apprentissage	526,2	7,0
2,2-dimethylpentane	<chem>CCCC(C)(C)C</chem>	Paraffine	apprentissage	255,8	2,3
1-ethenyl-3-methyl benzene	<chem>Cc1ccc(C=C)c1</chem>	Aromatique	apprentissage	317,2	7,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
1-octanol	CCCCCCCCO	Alcool	apprentissage	352,8	1,2
2,3,3,4-tetramethylhexane	CCC(C)C(C)(C)C(C)C	Paraffine	apprentissage	303,7	0,3
dodecyl ethanoate	CCCCCCCCCCCCOC(C)=O	Ester	apprentissage	395,0	6,0
2,3-dimethylheptane	CCCC(C)C(C)C	Paraffine	apprentissage	294,1	6,1
3-methylbutyl hexanoate	CCCCCC(=O)OCCC(C)C	Ester	apprentissage	361,3	3,3
1,6-hexanediol	OCCCCCO	Alcool	apprentissage	408,5	36,5
2-methylpentane	CCCC(C)C	Paraffine	apprentissage	246,3	3,7
heptan-4-one	CCCC(=O)CCC	Cétone	apprentissage	309,0	12,2
2-(2-butoxyethoxy)ethyl acetate	CCCCOCCOCCOC(C)=O	Ester	apprentissage	386,4	5,7
1-dodecene	CCCCCCCCCCC=C	Alcène	apprentissage	343,4	2,6
ethyl (2E,4Z)-deca-2,4-dienoate	CCCCCC=CC=CC(=O)OCC	Ester	apprentissage	385,5	0,5
1-pentadecene	CCCCCCCCCCCCC=C	Alcène	apprentissage	382,3	2,7
1,1-diethylcyclohexane	CCC1(CC)CCCCC1	Cyclique	apprentissage	315,2	7,0
tridecane	CCCCCCCCCCCCC	Paraffine	apprentissage	359,0	7,0
9,10-dihydroanthracene-9,10-dione	O=C1c2cccc2C(=O)c2cccc12	Aromatique (Oxygéné)	apprentissage	458,9	0,8
1,2-dimethyl-3-ethylbenzene	CCc1cccc(C)c1C	Aromatique	apprentissage	332,5	5,5
2-hydroxybenzoic acid	OC(=O)c1ccccc1O	Aromatique (Oxygéné)	apprentissage	435,0	5,0
4-propylheptane	CCCC(CCC)CCC	Paraffine	apprentissage	313,6	0,4
3,3-dimethylpentane	CCC(C)(C)CC	Paraffine	apprentissage	257,7	3,7
3-methyloctane	CCCCC(C)CC	Paraffine	apprentissage	297,1	0,1
ethyl octadecanoate	CCCCCCCCCCCCCCCC(=O)OCC	Ester	apprentissage	429,9	11,1
2-(3-oxobutanolyoxy) ethyl-2-methyl prop-2-enoate	CC(=O)CC(=O)OCCOC(=O)C(C)=C	Ester	apprentissage	408,6	1,6
diethyl oxalate	CCOC(=O)C(=O)OCC	Ester	apprentissage	388,0	7,0
4,4-dimethylheptane	CCCC(C)(C)CCC	Paraffine	apprentissage	290,6	2,6
5-methylhexan-2-one	CC(C)CCC(C)=O	Cétone	apprentissage	309,7	0,6
p-tert-butylphenol	CC(C)(C)c1ccc(O)cc1	Aromatique (Oxygéné)	apprentissage	381,8	4,2

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
propylcyclopentane	CCCC1CCCC1	Cyclique	apprentissage	286,4	2,6
2-methylprop-2-enal	CC(=C)C=O	Aldehyde	apprentissage	268,2	7,0
2,3-butanediol	CC(O)C(C)O	Alcool	apprentissage	358,2	0,0
2,2,3-trimethylhexane	CCCC(C)C(C)(C)C	Paraffine	apprentissage	288,2	0,2
3-methylheptan-3-ol	CCCC(C)(O)CC	Alcool	apprentissage	335,3	8,3
1-methyl-4-(propan-2-yl)cyclohexa-1,4-diene	CC(C)C1=CCC(C)=CC1	Cyclique	apprentissage	320,5	3,6
{4-[(benzoyloxy)methyl]cyclohexyl}methyl benzoate	O=C(OCC1CCC(COC(=O)c2ccccc2)CC1)c1ccccc1	Aromatique (Oxygéné)	apprentissage	441,0	7,0
4-methyl-2-pentanol	CC(C)CC(C)O	Alcool	apprentissage	315,3	1,3
2-oxacycloheptanone	O=C1CCCCO1	Cyclique (Oxygéné)	apprentissage	314,4	67,8
2,4,4-trimethyl-1-pentene	CC(=C)CC(C)(C)C	Alcène	apprentissage	266,7	0,5
tetradecane	CCCCCCCCCCCC	Paraffine	apprentissage	371,8	1,3
tert-butyl formate	CC(C)(C)OC=O	Ester	apprentissage	269,6	5,4
3-methyl-3-pentanol	CCC(C)(O)CC	Alcool	apprentissage	306,7	9,5
3-methylpent-1-ene	CCC(C)C=C	Alcène	apprentissage	244,1	1,0
1-methylnaphthalene	Cc1cccc2ccccc12	Aromatique	apprentissage	364,4	9,3
nonyl ethanoate	CCCCCCCCCOC(C)=O	Ester	apprentissage	365,2	7,0
methyl (9Z,12Z,15Z)-octadec-9,12,15-trienoate	CCC=CCC=CCC=CCCCCCCC(=O)OC	Ester	apprentissage	379,0	7,0
cyclohexanol	OC1CCCCC1	Cyclique (Oxygéné)	apprentissage	327,9	13,1
undecyl butanoate	CCCCCCCCCOC(=O)CCC	Ester	test	402,6	18,4
cyclooctane	C1CCCCCCC1	Cyclique	test	292,7	10,4
ethyl 3-oxobutanoate	CCOC(=O)CC(C)=O	Ester	test	356,5	26,1
hexyl ethanoate	CCCCCCOC(C)=O	Ester	test	328,8	18,7
cis-3-methylcyclohexanol	CC1CCCC(O)C1	Cyclique (Oxygéné)	test	336,9	1,7
2,5-dihydrofuran	C1=CCOC1	Cyclique (Oxygéné)	test	259,9	2,7
1-methoxy-2-(2-methoxyethoxy)ethane	COCCOCCOC	Ether	test	330,8	5,1
1-tridecanol	CCCCCCCCCCCCO	Alcool	test	404,8	10,8

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
2,5-dimethylheptane	<chem>CCC(C)CCC(C)C</chem>	Paraffine	test	293,2	5,2
cis-2-butene-1,4-diol	<chem>OCC=CCO</chem>	Alcool	test	415,2	14,0
diphenylmethane	<chem>c1ccc(Cc2ccccc2)cc1</chem>	Aromatique	test	392,8	10,3
1,3,5-tris(methylethyl)benzene	<chem>CC(C)c1cc(C(C)C)cc(C(C)C)c1</chem>	Aromatique	test	382,7	22,9
2-hydroxyethyl methacrylate	<chem>C=C(C)C(=O)OCCO</chem>	Ester	test	368,3	1,9
pentyl propanoate	<chem>CCCCCOC(=O)CC</chem>	Ester	test	327,6	8,4
di(2-ethylhexyl)adipate	<chem>CCCCCC(CC)COC(=O)CCCCCC(=O)OCC(CC)C CCC</chem>	Ester	test	476,9	10,9
propylene glycol monomethyl ether acetate	<chem>COCC(C)OC(C)=O</chem>	Ester	test	318,0	2,0
methyl (3E)-non-3-enoate	<chem>CCCCC=CCC(=O)OC</chem>	Ester	test	343,9	17,1
2,3-dimethyl-1-propylbenzene	<chem>CCCc1cccc(C)c1C</chem>	Aromatique	test	343,6	1,4
3-ethyl-3-methylheptane	<chem>CCCC(C)(CC)CC</chem>	Paraffine	test	307,8	6,2
ethyl octanoate	<chem>CCCCCCCC(=O)OCC</chem>	Ester	test	351,8	2,2
3-methylbutyl dodecanoate	<chem>CCCCCCCCCCCC(=O)OCCC(C)C</chem>	Ester	test	415,1	29,1
1-ethenyl-2-methyl benzene	<chem>C=Cc1cccc1C</chem>	Aromatique	test	317,9	2,3
diacetone alcohol	<chem>CC(=O)CC(C)(C)O</chem>	Cétone	test	362,0	31,0
dinonyl phthalate	<chem>CCCCCCCCCOC(=O)c1cccc1C(=O)OCCCC CCCC</chem>	Aromatique (Oxygéné)	test	493,3	4,1
2-methoxy-2-methylbutane	<chem>CCC(C)(C)OC</chem>	Ether	test	252,8	9,4
decyl methanoate	<chem>CCCCCCCCCOC=O</chem>	Ester	test	380,3	7,3
2,2-dimethylbutanoic acid	<chem>CCC(C)(C)C(=O)O</chem>	Acide carboxylique	test	343,0	9,1
[4-(hydroxymethyl)phenyl]methanol	<chem>OCC1ccc(CO)cc1</chem>	Aromatique (Oxygéné)	test	462,2	1,2
vinyl 2,2-dimethylpropanoate	<chem>C=COC(=O)C(C)(C)C</chem>	Ester	test	286,9	4,8
2-methoxymethylethoxy-propan-2-ol	<chem>COC(C)COC(C)CO</chem>	Ether	test	351,9	4,8
p-cymene	<chem>Cc1ccc(C(C)C)cc1</chem>	Aromatique	test	324,4	4,4
(1-methyl-1-phenylethyl)benzene	<chem>CC(C)(c1cccc1)c1cccc1</chem>	Aromatique	test	418,3	35,2
methyl decanoate	<chem>CCCCCCCCC(=O)OC</chem>	Ester	test	356,7	10,5

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
sec-butylbenzene	<chem>CCC(C)c1ccccc1</chem>	Aromatique	test	323,2	1,8
ethyl ethanoate	<chem>CCOC(C)=O</chem>	Ester	test	274,6	5,6
cyclohexylbenzene	<chem>c1ccc(C2CCCCC2)cc1</chem>	Aromatique	test	354,2	17,8
2,2-bis(4-hydroxyphenyl)propane	<chem>CC(C)(c1ccc(O)cc1)c1ccc(O)cc1</chem>	Aromatique (Oxygéné)	test	434,5	45,7
phenylethene	<chem>C=Cc1ccccc1</chem>	Aromatique	test	301,5	3,5
ethyl dodecanoate	<chem>CCCCCCCCCCCC(=O)OCC</chem>	Ester	test	392,6	6,6
1-(octyloxy)octane	<chem>CCCCCCCCOCCCCCCCC</chem>	Ether	test	404,2	19,1
heptadecane	<chem>CCCCCCCCCCCCCCCCCC</chem>	Paraffine	test	405,4	15,7
2-methylpropyl ethanoate	<chem>CC(=O)OCC(C)C</chem>	Ester	test	299,3	8,3
2-methyl-1-propanol	<chem>CC(C)CO</chem>	Alcool	test	295,1	5,9
4-(1,1-dimethylpropyl)phenol	<chem>CCC(C)(C)c1ccc(O)cc1</chem>	Aromatique (Oxygéné)	test	390,2	6,2
2,2-dimethylpropane-1,3-diol	<chem>CC(C)(CO)CO</chem>	Alcool	test	383,8	12,7
cis-4-methylcyclohexanol	<chem>CC1CCC(O)CC1</chem>	Cyclique (Oxygéné)	test	336,9	6,3
3-methylcyclohex-1-ene	<chem>CC1C=CCCC1</chem>	Cyclique	test	272,1	2,1
2-(2-hexoxyethoxy)ethanol	<chem>CCCCCOCCOCCO</chem>	Ether	test	395,2	13,0
o-terphenyl	<chem>c1ccc(-c2ccccc2-c2ccccc2)cc1</chem>	Aromatique	test	473,1	37,1
nonadecane	<chem>CCCCCCCCCCCCCCCCCC</chem>	Paraffine	test	423,9	17,1
3,4,5-trimethylheptane	<chem>CCC(C)(C)(C)CC</chem>	Paraffine	test	308,5	4,5
methyl nonanoate	<chem>CCCCCCCC(=O)OC</chem>	Ester	test	347,1	12,9
butanal	<chem>CCCC=O</chem>	Aldehyde	test	265,9	3,7
cis-2-methylcyclohexanol	<chem>CC1CCCC1O</chem>	Cyclique (Oxygéné)	test	335,1	3,9
3-methylbutyl butanoate	<chem>CCCC(=O)OCCC(C)C</chem>	Ester	test	337,3	6,3
methyl nonadecanoate	<chem>CCCCCCCCCCCCCCCCCC(=O)OC</chem>	Ester	test	408,9	22,9
1-ethoxy-2-(2-ethoxyethoxy)ethane	<chem>CCOCCOCCOCC</chem>	Ether	test	344,6	10,6
cyclopentane	<chem>C1CCCC1</chem>	Cyclique	test	239,4	3,4
benzyl acetate	<chem>CC(=O)OCc1ccccc1</chem>	Aromatique (Oxygéné)	test	355,3	7,8
isopentane	<chem>CCC(C)C</chem>	Paraffine	test	229,2	12,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
1,3-butanediol	<chem>CC(O)CCO</chem>	Alcool	test	376,3	5,7
1-heptylbenzene	<chem>CCCCCCCc1ccccc1</chem>	Aromatique	test	361,8	6,3
1-tetradecene	<chem>C=CCCCCCCCCCCC</chem>	Alcène	test	370,2	0,2
ethyl (2E)-oct-2-enoate	<chem>CCCCC=CC(=O)OCC</chem>	Ester	test	360,7	10,3
methyl benzoate	<chem>COC(=O)c1ccccc1</chem>	Aromatique (Oxygéné)	test	350,1	5,1
3-ethylpent-2-ene	<chem>CC=C(CC)CC</chem>	Alcène	test	266,4	0,6
cyclohexene	<chem>C1=CCCCC1</chem>	Cyclique	test	258,5	15,4
1-ethoxypropan-2-ol	<chem>CCOCC(C)O</chem>	Ether	test	317,9	2,7
2-butoxyethanol	<chem>CCCCOCCO</chem>	Ether	test	340,7	7,6
cis-1,3-dimethylcyclohexane	<chem>CC1CCCC(C)C1</chem>	Cyclique	test	286,0	7,0
nonane	<chem>CCCCCCCCC</chem>	Paraffine	test	300,4	3,6
2-methyl-, 3-hydroxy-2,2,4-trimethylpentyl propanoate	<chem>CC(C)C(=O)OCC(C)(C)C(O)C(C)C</chem>	Ester	test	406,5	27,4
p-methoxyphenol	<chem>COc1ccc(O)cc1</chem>	Aromatique (Oxygéné)	test	367,1	31,0
2-butyl-2-ethylpropane-1,3-diol	<chem>CCCC(CC)(CO)CO</chem>	Alcool	test	412,5	26,5
isopropylbenzene	<chem>CC(C)c1ccccc1</chem>	Aromatique	test	308,8	4,6
sec-butyl formate	<chem>CCC(C)OC=O</chem>	Ester	test	284,4	14,6
benzaldehyde	<chem>O=Cc1ccccc1</chem>	Aromatique (Oxygéné)	test	324,2	11,0
1-pentyne	<chem>C#CCCC</chem>	Alcyne	test	252,9	13,9
2,3-dimethyl-1-butene	<chem>C=C(C)C(C)C</chem>	Alcène	test	242,3	12,8
pentyl butanoate	<chem>CCCCCOC(=O)CCC</chem>	Ester	test	340,4	4,6
cis-3-hexene	<chem>CCC=CCC</chem>	Alcène	test	249,1	4,1
dodecanal	<chem>CCCCCCCCCCCC=O</chem>	Aldehyde	test	383,6	9,5
isopentylbenzene	<chem>CC(C)CCc1ccccc1</chem>	Aromatique	test	334,3	0,7
2,3-dimethylbutane	<chem>CC(C)C(C)C</chem>	Paraffine	test	243,4	0,6
methyl eicosanoate	<chem>CCCCCCCCCCCCCCCCCCCC(=O)OC</chem>	Ester	test	411,5	25,3
2-methyloxolane	<chem>CC1CCCO1</chem>	Cyclique (Oxygéné)	test	274,1	12,0

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
3-ethyl-2-methylpentane	<chem>CCC(CC)C(C)C</chem>	Paraffine	test	278,9	2,9
2,2'-ethylenedioxydiethanol	<chem>OCCOCCOCCO</chem>	Ether	test	429,6	0,5
2-(2-methoxyethoxy)ethanol	<chem>COCCOCCO</chem>	Ether	test	357,1	0,0
p-hydroquinone	<chem>Oc1ccc(O)cc1</chem>	Aromatique (Oxygéné)	test	396,1	41,9
2,2-dimethylpropyl methanoate	<chem>CC(C)(C)COC=O</chem>	Ester	test	311,7	0,7
3,5-dimethyloctane	<chem>CCCC(C)CC(C)CC</chem>	Paraffine	test	309,8	4,2
cyclohexane-1,4-diAcide carboxylique	<chem>O=C(O)C1CCC(C(=O)O)CC1</chem>	Cyclique (Oxygéné)	test	444,3	63,9
ethyl (2E)-but-2-enoate	<chem>CC=CC(=O)OCC</chem>	Ester	test	305,8	30,8
propylcyclohexane	<chem>CCCC1CCCC1</chem>	Cyclique	test	305,5	1,5
hexanoic acid	<chem>CCCCCC(=O)O</chem>	Acide carboxylique	test	365,2	9,8
methyl heptanoate	<chem>CCCCCCC(=O)OC</chem>	Ester	test	325,7	0,7
2-methylpentan-3-one	<chem>CCC(=O)C(C)C</chem>	Cétone	test	290,2	4,1
3-methylpenta-1,3-diene	<chem>C=CC(C)=CC</chem>	Alcène	test	248,3	3,3
tert-butyl acetate	<chem>CC(=O)OC(C)(C)C</chem>	Ester	test	283,6	3,6
heptyl ethanoate	<chem>CCCCCCCOC(C)=O</chem>	Ester	test	341,6	7,4
3-ethylpentane	<chem>CCC(CC)CC</chem>	Paraffine	test	265,4	4,4
2-phenyl-2-methylbutane	<chem>CCC(C)(C)c1ccccc1</chem>	Aromatique	test	335,0	3,0
3-methylidene-6-(propan-2-yl)cyclohex-1-ene	<chem>C=C1C=CC(C(C)C)CC1</chem>	Cyclique	test	313,0	9,2
1-pentanol	<chem>CCCCCO</chem>	Alcool	test	315,1	14,9
methyl octanoate	<chem>CCCCCCCC(=O)OC</chem>	Ester	test	336,7	5,3
2-butyl-octan-1-ol	<chem>CCCCCCC(CO)CCCC</chem>	Alcool	test	390,3	8,2
nonanoic acid	<chem>CCCCCCCCC(=O)O</chem>	Acide carboxylique	test	399,8	2,3
isobutane	<chem>CC(C)C</chem>	Paraffine	test	210,9	20,7
2-ethylhexyl acrylate	<chem>C=CC(=O)OCC(CC)CCCC</chem>	Ester	test	373,8	18,6
3-hydroxybutyraldehyde	<chem>CC(O)CC=O</chem>	Aldehyde	test	325,4	13,6
1,2-benzenediol	<chem>Oc1ccccc1O</chem>	Aromatique (Oxygéné)	test	389,9	10,1
1-octyne	<chem>C#CCCCCCC</chem>	Alcyne	test	290,7	1,6

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
methanol	CO	Alcool	test	275,0	9,0
1-hexadecene	C=CCCCCCCCCCCCCCC	Alcène	test	393,5	11,7
m-terphenyl	c1ccc(-c2ccc(-c3ccccc3)c2)cc1	Aromatique	test	473,0	9,0
hexan-2-one	CCCC(C)=O	Cétone	test	299,4	3,2
p-ethylphenol	CCc1ccc(O)cc1	Aromatique (Oxygéné)	test	369,8	3,4
vinyl formate	C=COC=O	Ester	test	267,8	11,8
vinyl propionate	C=COC(=O)CC	Ester	test	280,4	6,4
propanal	CCC=O	Aldehyde	test	250,0	6,9
2-methylpropanal	CC(C)C=O	Aldehyde	test	260,2	6,1
1,4-cyclohexanedimethanol	OCC1CCC(CO)CC1	Cyclique (Oxygéné)	test	417,4	16,7
octadecane	CCCCCCCCCCCCCCCCCC	Paraffine	test	415,1	23,1
2,5-dimethyl-1,5-hexadiene	C=C(C)CCC(=C)C	Alcène	test	275,6	4,4
methyl heneicosanoate	CCCCCCCCCCCCCCCCCCCC(=O)OC	Ester	test	413,6	27,6
o-diethylbenzene	CCc1ccccc1CC	Aromatique	test	326,8	3,6
cyclopentadiene	C1=CCC=C1	Cyclique	test	230,1	42,9
4,4-dimethylpent-1-ene	C=CCC(C)(C)C	Alcène	test	252,6	8,4
1-tridecene	C=CCCCCCCCCCCC	Alcène	test	357,2	5,1
3-ethyl-1-pentene	C=CC(CC)CC	Alcène	test	262,3	6,3
dipropylene glycol monomethyl ether acetate	COCC(C)OCC(C)OC(C)=O	Ester	test	364,5	5,5
2-phenylethanol	OCCc1ccccc1	Aromatique (Oxygéné)	test	367,4	1,6
gamma-butyrolactone	O=C1CCCO1	Cyclique (Oxygéné)	test	307,5	64,0
4-methylhept-1-ene	C=CCC(C)CCC	Alcène	test	277,6	2,6
pentan-3-one	CCC(=O)CC	Cétone	test	283,7	2,3
hexane-2,5-diol	CC(O)CCC(C)O	Alcool	test	382,1	8,1
p-methylstyrene	C=Cc1ccc(C)cc1	Aromatique	test	317,2	1,8
butanol	CCCCO	Alcool	test	301,9	8,1
4-methylpentan-2-one	CC(=O)CC(C)C	Cétone	test	297,0	10,8



Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
undecane	CCCCCCCCCCC	Paraffine	test	331,1	7,1
2-methylhept-2-ene	CCCC=C(C)C	Alcène	test	279,3	1,7
toluene	Cc1ccccc1	Aromatique	test	278,9	1,8
2-ethylnaphthalene	CCc1ccc2ccccc2c1	Aromatique	test	376,8	0,3
butyl octadecanoate	CCCCCCCCCCCCCCCC(=O)OCCCC	Ester	test	440,0	6,8
decane	CCCCCCCCCC	Paraffine	test	316,0	3,0
2-methylpent-1-ene	C=C(C)CCC	Alcène	test	244,3	2,7
hexadecane	CCCCCCCCCCCCCCCC	Paraffine	test	395,0	13,1
cyclohexane	C1CCCCC1	Cyclique	test	260,5	7,4
2,2-dimethylbutane	CCC(C)(C)C	Paraffine	test	238,8	13,8
cyclodecane	C1CCCCCCCCC1	Cyclique	test	323,1	15,0
1,3-butadiene	C=CC=C	Alcène	test	212,4	15,2
oct-1-en-3-ol	C=CC(O)CCCC	Alcool	test	348,7	7,7
methylenecyclopentane	C=C1CCCC1	Cyclique	test	250,9	3,1
1,5-hexadiene	C=CCCC=C	Alcène	test	251,6	24,4
1-hexyne	C#CCCC	Alcyne	test	265,3	13,3
(2-methylbutyl)benzene	CCC(C)Cc1ccccc1	Aromatique	test	334,8	2,2
methylcyclohexane	CC1CCCCC1	Cyclique	test	273,3	6,1
ethyl (2E)-hex-3-enoate	CCC=CCC(=O)OCC	Ester	test	322,6	9,4
1,1-dimethylcyclopentane	CC1(C)CCCC1	Cyclique	test	262,1	5,1
2-hexoxyethanol	CCCCCOCCO	Ether	test	363,0	8,0
1-methyl-1,3-cyclopentadiene	CC1=CC=CC1	Cyclique	test	248,1	73,9
dodecylbenzene	CCCCCCCCCCCCCc1ccccc1	Aromatique	test	409,9	4,1
prop-2-enoic acid	C=CC(=O)O	Acide carboxylique	test	335,1	11,1
propyl propanoate	CCCOC(=O)CC	Ester	test	300,6	8,4
(9Z)-octadec-9-enoic acid	CCCCCCCC=CCCCCCCC(=O)O	Acide carboxylique	test	450,2	11,8
ethyl (2E)-dec-2-enoate	CCCCCCC=CC(=O)OCC	Ester	test	385,2	21,2

Nom	SMILES Canonique	Famille Chimique	Jeu	PE Prédit (K)	Ecart* (K)
5-methylhex-1-ene	<chem>C=CCCC(C)C</chem>	Alcène	test	259,5	3,5
pentan-2-one	<chem>CCCC(C)=O</chem>	Cétone	test	286,2	6,2
cis-2-butene	<chem>CC=CC</chem>	Alcène	test	216,2	16,2
trans-3-octene	<chem>CCC=CCCC</chem>	Alcène	test	281,1	0,9
1-phenylethan-1-one	<chem>CC(=O)c1ccccc1</chem>	Aromatique (Oxygéné)	test	347,5	7,9
2-methyl-1-butanol	<chem>CCC(C)CO</chem>	Alcool	test	308,9	7,2
phenoxybenzene	<chem>c1ccc(Oc2ccccc2)cc1</chem>	Aromatique (Oxygéné)	test	389,3	20,3
decyl ethanoate	<chem>CCCCCCCCCOC(C)=O</chem>	Ester	test	375,9	2,1
ethanol	<chem>CCO</chem>	Alcool	test	278,2	7,8
2,2-dimethyloctane	<chem>CCCCCCC(C)(C)C</chem>	Paraffine	test	304,9	0,9
ethyl hexanoate	<chem>CCCCCC(=O)OCC</chem>	Ester	test	327,2	1,2
2-methylhept-1-ene	<chem>C=C(C)CCCC</chem>	Alcène	test	275,8	7,4
1-decene	<chem>C=CCCCCCCC</chem>	Alcène	test	313,3	2,3
2,3,4,4-tetramethylhexane	<chem>CCC(C)(C)C(C)C(C)C</chem>	Paraffine	test	303,0	1,0
methyl (2E)-penta-2,4-dienoate	<chem>C=CC=CC(=O)OC</chem>	Ester	test	311,4	1,4
2,3-epoxypropanol	<chem>OCC1CO1</chem>	Cyclique (Oxygéné)	test	317,6	26,6
methyl dodecanoate	<chem>CCCCCCCCCCCC(=O)OC</chem>	Ester	test	373,6	12,4
2,2,4-trimethylpentane	<chem>CC(C)CC(C)(C)C</chem>	Paraffine	test	269,5	8,5
2,2-dimethylpropanoic acid	<chem>CC(C)(C)C(=O)O</chem>	Acide carboxylique	test	329,8	6,4

\* écart absolu entre valeur prédite et expérimentale

# Annexe B. Inverse-QSPR for *de novo* Design: A Review

Philippe Gantzer, Benoit Creton,\* Carlos Nieto-Draghi

DOI: 10.1002/minf.201900087

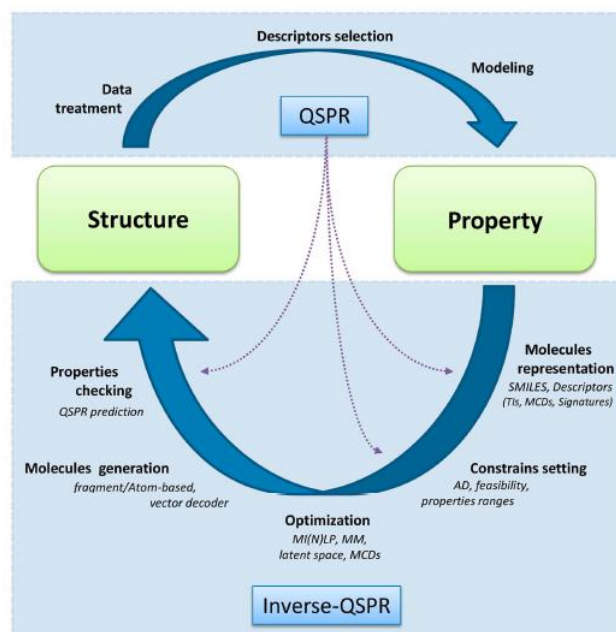
**Abstract:** The use of computer tools to solve chemistry-related problems has given rise to a large and increasing number of publications these last decades. This new field of science is now well recognized and labelled Chemoinformatics. Among all chemoinformatics techniques, the use of statistical based approaches for property predictions has been the subject of numerous research reflecting both new developments and many cases of applications. The so obtained predictive models relating a property to molecular features – descriptors – are gathered under the acronym QSPR, for Quantitative Structure Property Relationships. Apart from the obvious use of such models to predict property values for new compounds, their use to virtually synthesize new molecules – *de novo* design – is currently a high-interest subject. Inverse-QSPR (i-QSPR) methods have hence been developed to accelerate the discovery of new materials that meet a set of specifications. In the proposed manuscript, we review existing i-QSPR methodologies published in the open literature in a way to highlight developments, applications, improvements and limitations of each.

Review www.molinf.com

DOI: 10.1002/minf.201900087  Check for updates

**Inverse-QSPR for *de novo* Design: A Review**

Philippe Gantzer,<sup>[a]</sup> Benoit Creton,<sup>\*[a]</sup> and Carlos Nieto-Draghi<sup>[a]</sup>



## Annexe C. Génération de molécules par assemblages de fragments et contraintes sur la propriété

Dans cette Annexe, nous présentons les résultats obtenus pour la génération de molécules par assemblage de fragments (F), en imposant une contrainte sur la valeur de point d'éclair (PE). Deux méthodes de génération inspirées de la méthode F1b (voir partie 4.2 « Génération par assemblage de fragments »), F2a et F2b, sont d'abord présentées. Ensuite, les performances des méthodes F0, F1a, F1b, F2a et F2b sont évaluées pour la génération de molécules respectant la contrainte imposée sur la propriété.

### C.1. Méthodes F supplémentaires

Comme discuté dans le Chapitre 3 « Méthodes pour la génération moléculaire et pour leur comparaison », la génération de molécules pour une propriété donnée peut être restreinte par des descripteurs de type MCD (Monotonically Changing Descriptors),<sup>59</sup> dont font partie les fragments SMF. Les méthodes F supplémentaires décrites dans cette partie utilisent de telles contraintes.

#### C.1.1. Méthode F2a

Le Tableau S1 présente les intervalles des valeurs prises par certains fragments (ou descripteurs SMF de type 3, tels que définis dans la section 2.1.1.1.2 « Descripteurs moléculaires ») dans les trois sous-ensembles de molécules (présentés dans le Tableau 30). Certains fragments ne sont pas présents dans toutes les molécules de chaque intervalle de FP. Le fragment C-C, par exemple, n'est pas présent dans les structures du méthanol et du diméthyléther (les SMILES de ces molécules sont : CO et COC, leur valeur de PE est comprise dans l'intervalle [200 K ; 300 K]) ni dans celle du trioxyméthylène (le SMILES de cette molécule est C1OCOCO1 et sa valeur de PE est comprise dans l'intervalle [300 K ; 400 K]). Les bornes de ces 5 descripteurs varient en fonction du sous-ensemble, et donc en fonction des valeurs de la propriété des molécules. Comme la valeur de PE est liée à la taille moléculaire<sup>91,134,135</sup>, les valeurs des descripteurs ont tendance à augmenter lorsque la valeur de PE des molécules augmente.

Intervalle de PE	Intervalle de descripteurs				
	C-C	C-C-C	C=C	C-O	C-C-O
[200 K ; 300K]	[0 ; 9]	[0 ; 14]	[0 ; 4]	[0 ; 4]	[0 ; 6]
[300 K ; 400 K]	[0 ; 30]	[0 ; 24]	[0 ; 7]	[0 ; 8]	[0 ; 9]
[400 K ; 500 K]	[2 ; 29]	[0 ; 34]	[0 ; 9]	[0 ; 9]	[0 ; 10]

Tableau S1 : Intervalles des valeurs de descripteurs observées pour les molécules des différents intervalles de point d'éclair du jeu initial.

La méthode de génération F1b, qui assemble des fragments SMF de type 3, a été modifiée pour diriger la génération vers des molécules possédant une valeur de PE dans un intervalle ciblé. Lors de la génération, les probabilités de sélection des fragments sont redéfinies en fonction des bornes des descripteurs dans le sous-ensemble ciblé. Cette variante de F1b a été nommée F2a.

### C.1.2. Méthode F2b

Nous nous sommes également intéressés à pondérer le choix des fragments à assembler de manière alternative à la méthode F2a, en considérant la carte GTM construite dans la partie 2.1.2.3 « Représentation des données et modèles GTM ».

Sur la carte GTM de la Figure S1a, les molécules projetées sont colorées (vert, bleu, ou rouge) en fonction de leur appartenance à un des intervalles de PE. Globalement, les molécules possédant leur valeur de PE comprise dans le premier intervalle sont projetées au milieu de la carte et celles avec leur valeur de PE comprise dans les autres intervalles sont projetées davantage aux extrémités. Nous avons ensuite considéré le nœud d'affectation de chaque molécule comme le nœud pour lequel la molécule possède la plus haute *responsabilité* (la plus forte probabilité de projection sur le nœud).

Comme proposé par Kaneko, il est possible d'identifier sur une carte GTM les nœuds dans lesquels se trouvent des molécules, avec une forte *responsabilité*, et possédant la propriété souhaitée.<sup>143</sup> Sur la carte GTM de la Figure S1b, nous avons représenté les nœuds dans lesquels au moins une molécule est affectée, et les avons colorés (vert, bleu, ou rouge) en fonction du ou des intervalles de PE de toutes les molécules qui y sont affectées. Cette coloration par nœud

diffère de la coloration de la carte GTM. La coloration de la carte GTM considère pour chaque nœud une valeur moyenne de propriété (définie comme la somme des valeurs de propriété de toutes les molécules pondérées par leurs *responsabilités* dans le nœud), et colore chaque nœud par une couleur qui est fonction de cette propriété moyenne. Dans notre approche, chaque nœud peut posséder plusieurs couleurs en fonction du ou des intervalles de propriété des molécules affectées au nœud.

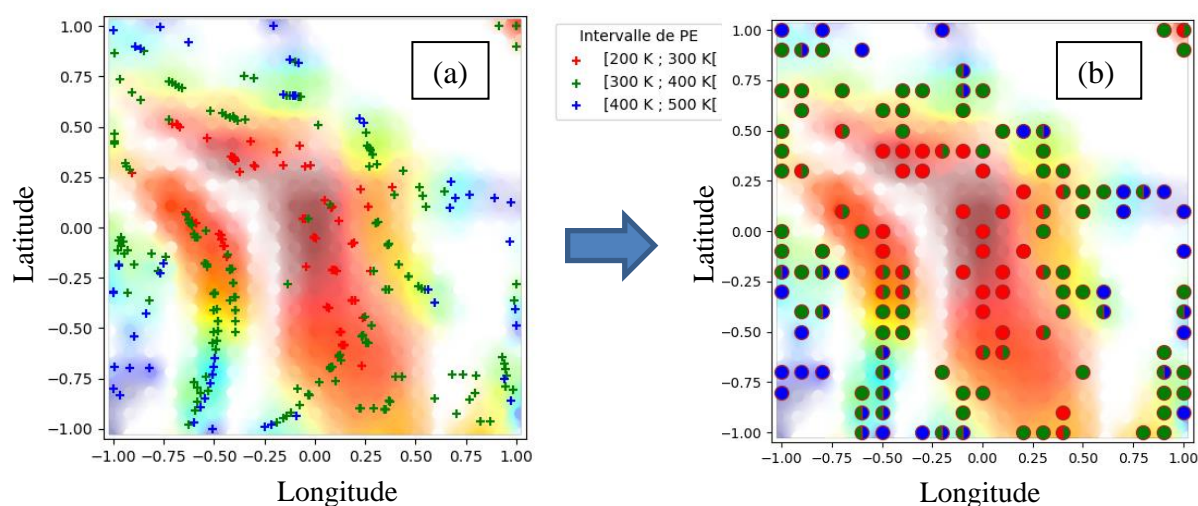


Figure S1 : Carte GTM issue de la Figure 3, avec (a) projection des molécules initiales, colorées en fonction de leur intervalle de propriété, ou avec (b) projection des nœuds, colorés en fonction du ou des intervalles de propriétés des molécules y étant affectées.

Nous avons listé pour chaque intervalle de PE les nœuds de la carte GTM dans lesquels au moins une molécule de l'intervalle est affectée. Une fois les nœuds listés, toutes les molécules de ces nœuds – possédant ou non la propriété souhaitée – sont extraites avec leurs valeurs de descripteurs. Les bornes de descripteurs sont ensuite redéfinies à partir des valeurs de descripteurs de ces molécules. Comme avec F2a, ces bornes sont ensuite utilisées pour pondérer la sélection des fragments à ajouter. Cette variante de génération a été nommée F2b. Avec F2b, l'idée est de profiter de la généralisation de l'espace chimique procurée par la carte GTM pour pondérer le choix des fragments, contrairement à F1a qui pondère plus strictement ce choix selon les seules molécules appartenant à chaque sous-ensemble de PE.

## C.2. Performances des générations de F avec contraintes sur la propriété

Dans cette section, nous analysons les performances des variations de génération par assemblage de fragments pour obtenir des molécules dans chaque intervalle de PE. Le Tableau S2 présente les caractéristiques des différentes variations présentées précédemment et considérées dans cette section.

Méthode	Fragments	Pondération du choix	
		Liaisons	Fragments
F0	Simple	X	
F1a	SMF	X	
F1b	SMF	X	X <sup>a</sup>
F2a	SMF	X	X <sup>b</sup>
F2b	SMF	X	X <sup>c</sup>

<sup>a</sup> en fonction de l'intégralité du jeu initial de données – <sup>b</sup> en fonction d'un sous-ensemble du jeu initial de données sélectionné en fonction de la propriété – <sup>c</sup> en fonction d'un sous-ensemble du jeu initial de données sélectionné par la méthode GTM

Tableau S2 : Les différentes méthodes de génération par assemblage de fragments considérées.

Nous avons généré jusqu'à 1 million de structures prédictibles avec chaque méthode et calculé l'indice de spécificité (tel que défini dans la partie 5.2.5 « Spécificité de la génération ») des molécules obtenues. Les méthodes F2a et F2b ont été utilisées de manière distincte pour générer des molécules dans chaque intervalle de PE, car la pondération du choix des fragments varie d'un intervalle à l'autre. Le Tableau S3 présente les valeurs de l'indice de spécificité pour les molécules issues des différentes méthodes, après avoir généré 1 million de molécules prédictibles (au sens défini dans la partie 4.1 « Evaluation de la qualité des générations »), en fonction de l'intervalle de PE ciblé.

Molécules générées par la méthode	Indice de spécificité pour chaque intervalle de FP		
	[200 K ; 300 K[	[300 K ; 400 K[	[400 K ; 500 K[
F0	0,004	0,63	0,36
F1a	0,003	0,67	0,32
F1b	0,003	0,55	0,45
F2a	0,003	0,50	0,49
F2b	0,003	0,52	0,50

Tableau S3 : Valeurs de l'indice de spécificité pour différents jeux de molécules (jeu initial et des molécules générées) en fonction de l'intervalle de point d'éclair ciblé.

Les différentes approches étudiées ici ne génèrent pas plus de 0,4% de molécules possédant une valeur prédite de PE dans l'intervalle [200 K ; 300 K[, malgré le fait que la base de départ soit composée de 33% de molécules avec une valeur expérimentale de PE comprise dans cet intervalle. Nous avons au début de cette partie évoqué le fait que les molécules avec une telle valeur de PE étaient généralement plus petites que celles avec une valeur de PE supérieure à 300 K. De ce fait, la combinatoire est davantage limitée que pour des molécules de plus grande taille. Le phénomène opposé se produit pour la génération de molécules possédant une valeur de PE supérieure à 300 K ; les molécules étant de plus grandes tailles et contenant davantage d'atomes, le nombre de molécules possibles pouvant être générées par combinaison de fragments est supérieur à celui des molécules qui ont une valeur de PE plus faible.

La Figure S2 présente le nombre de molécules générées par les différentes variations pour chaque intervalle de propriété, en fonction du nombre total de molécules générées. Pour générer un maximum de molécules possédant leur valeur de PE dans l'intervalle [200 K ; 300 K[, les méthodes F0 et F1a, ne pondérant pas le choix des fragments, s'avèrent les plus efficaces. Cependant, les méthodes F2a et F2b, pondérant le choix des fragments en fonction de la propriété ciblée, produisent plus rapidement des molécules possédant leur propriété dans cet intervalle : 2 500 structures satisfaisantes sont obtenues en générant un total de  $10^5$  molécules, alors que les autres méthodes nécessitent de générer près du double de molécules pour obtenir le même nombre de structures satisfaisantes. Pour générer des molécules possédant leur valeur



de PE dans l'intervalle [300 K ; 400 K], les variations ne pondérant pas le choix des fragments s'avèrent également les plus efficaces. Pour générer des molécules possédant une valeur de PE comprise dans le dernier intervalle, les variations F2a et F2b, pondérant le choix des fragments en fonction de la propriété, sont les plus efficaces.

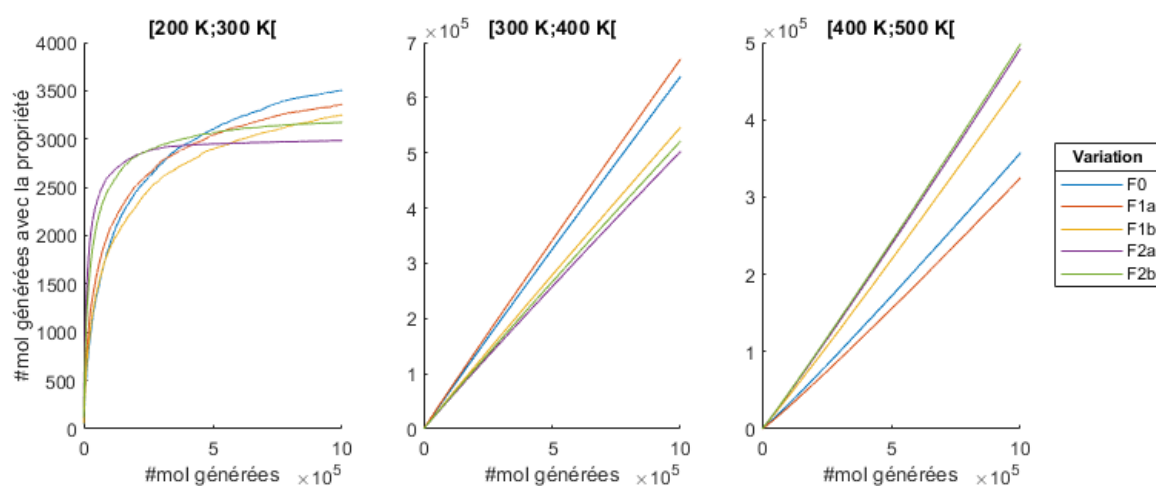


Figure S2 : Nombre de molécules générées par les différentes variations de la méthode d'assemblage de fragments, possédant leur valeur de point d'éclair dans l'intervalle ciblé, en fonction du nombre total de molécules générées (#mol générées).

## Annexe D. **Choix de la taille des cubes discrétisant $\mathbb{C}$**

Nous détaillons dans cette Annexe l'approche employée pour fixer la taille des cubes discrétisant l'espace  $\mathbb{C}$ .

Le nombre de cubes dans  $\mathbb{C}$ ,  $N_{cubes}$ , est corrélé à la taille d'un cube,  $T_{cube}$ . En considérant l'espace  $\mathbb{C}$  comme un cube de taille  $T_{\mathbb{C}}$ ,  $N_{cubes}$  est approximé par la relation (23).

$$N_{cubes} \approx (T_{\mathbb{C}}/T_{cube})^3 \quad (23)$$

Le choix de la valeur pour  $T_{cubes}$  est un compromis entre : (i) la capacité à pouvoir identifier de manière précise les zones  $\mathbb{C}$  (les cubes) dans lesquelles des molécules peuvent être générées, et (ii) la capacité à pouvoir comparer l'occupation relative de l'espace occupé de  $\mathbb{C}$  par les différents jeux de molécules. La capacité à pouvoir identifier de manière précise les zones  $\mathbb{C}$  augmente avec  $N_{cubes}$ , car la résolution de  $\mathbb{C}$  augmente. Au contraire, la capacité à pouvoir comparer l'occupation relative de l'espace occupé de  $\mathbb{C}$  par les différents jeux de molécules diminue avec  $N_{cubes}$ , car des molécules projetées proches l'une de l'autre ont une plus grande probabilité d'être localisées dans des cubes distincts. Dans le cas extrême où  $T_{\mathbb{C}} = T_{cubes}$ , le seul cube de  $\mathbb{C}$  est occupé par toutes les molécules et les valeurs des indices  $I_1$  à  $I_3$  sont toujours égales à leur valeur maximale et aucune comparaison n'est réalisable. Au contraire, si la résolution de  $\mathbb{C}$  est trop élevée, chaque molécule générée, quelle que soit la méthode employée, est projetée dans un cube différent. Dans ce cas, les valeurs des indices sont basses et stables avec le nombre de molécules et aucune comparaison n'est réalisable.

Pour choisir la valeur de  $T_{cubes}$ , dans la suite de cette Annexe, nous avons considéré la méthode G1b, avec laquelle nous avons généré  $5.10^4$  (50k) molécules. G1b a été utilisée, car c'est la seule méthode implémentée durant le travail de thèse qui permet de générer des molécules comportant des cycles. Nous avons considéré  $5.10^4$  molécules, car nous n'avons pas besoin de concaténer plusieurs générations ensemble pour les obtenir. Les générations ont été répétées dix fois, de manière identique, pour obtenir les valeurs moyennes.

### **D.1. Couverture du sous-espace de $\mathbb{C}$ regroupant les molécules initiales**

Dans cette section, nous avons uniquement considéré H, le sous-espace de  $\mathbb{C}$  regroupant les molécules initiales (représenté par la forme convexe violette sur la Figure 11) pour identifier une valeur optimale  $T_{cubes}$ . En effet, suite à la vérification du respect de l'AD des molécules

généérées, c'est la zone principale de projection de ces molécules ; et par conséquent la zone qui doit être discrétisée de manière optimale.

Nous avons testé des valeurs de  $T_{cubes}$  comprises entre 0,5 et 10 pour discrétiser H.  $T_{cubes}$  n'a pas d'unité, car  $\mathbb{C}$  est construit par ACP à partir de descripteurs SMF (qui dénombrent des fragments et n'ont pas d'unité). Des valeurs  $T_{cubes}$  inférieures à 0,5 n'ont pas pu être étudiées, car les ressources en mémoire vive nécessaires pour manipuler de tels ensembles de cubes (dont le nombre est supérieur à  $5 \cdot 10^5$ ) étaient trop importantes. La Figure S3 présente le taux moyen de cubes occupés dans H en fonction du nombre de molécules générées et de  $T_{cubes}$  ; et le Tableau S4 donne les valeurs des taux moyens à la fin des générations. Avec une valeur  $T_{cubes}$  égale ou supérieure à 3, quasiment tous les cubes de H sont occupés après avoir généré  $10^4$  molécules, ce qui démontre une résolution trop faible. Avec une valeur  $T_{cubes}$  égale à 2, l'évolution du taux de cubes occupés en fonction du nombre de molécules générées est plus lente que le cas précédent, mais la majorité des cubes sont également occupés à la fin des générations. Une valeur  $T_{cubes}$  comprise entre 1 et 2 permet une évolution encore plus lente de ce taux, à la fin des générations environ 83% des cubes de H sont occupés. Une valeur  $T_{cubes}$  égale à 0,5 ne permet que d'occuper 46% des cubes, la résolution semble alors trop élevée. Nous avons alors fixé la valeur  $T_{cubes}$  égale à 1.

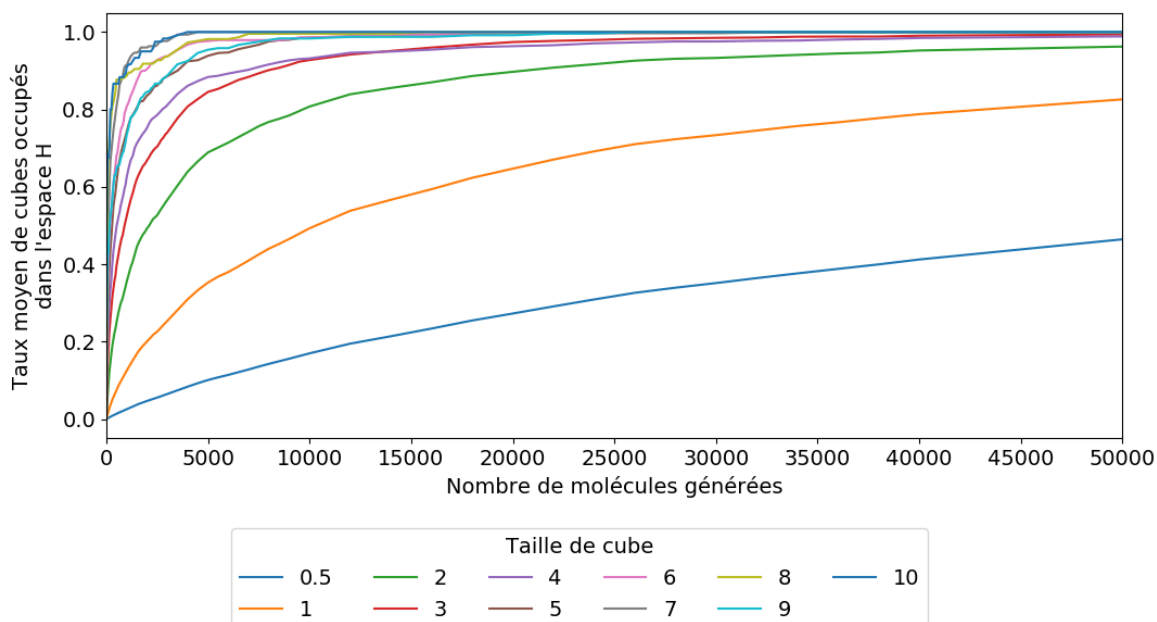


Figure S3 : Taux moyen de cubes occupés dans l'espace H des molécules initiales en fonction du nombre de molécules générées et de la taille des cubes, moyenne sur les dix générations avec la méthode G1b.

$T_{cubes}$	Taux moyen de cubes occupés
0,5	0,464 ± 0,004
1	0,826 ± 0,005
2	0,962 ± 0,003
3	0,994 ± 0,005
4	0,989 ± 0,007
5	1,000 ± 0,000
6	1,000 ± 0,000
7	1,000 ± 0,000
8	1,000 ± 0,000
9	1,000 ± 0,000
10	1,000 ± 0,000

Tableau S4 : Taux moyen de cubes occupés dans l'espace H en fonction de la taille des cubes  $T_{cubes}$ , sur dix générations de  $5.10^4$  molécules avec la méthode G1b.

## D.2. Variation des valeurs d'indices pour une même méthode de génération

Dans cette section, tout comme dans le Chapitre 5, l'espace  $\mathbb{C}$  complet a été considéré (et non plus uniquement le sous-espace H, qui a uniquement été utilisé pour fixer  $T_{cubes}$ ). Nous évaluons ci-après la stabilité des indices *collectifs*  $I_1$  à  $I_3$  sur dix générations identiques utilisant la méthode G1b. Suivant nos précédentes observations, la valeur de  $T_{cubes}$  a été fixée à 1. Les dix générations ont été comparées à l'aide des indices *collectifs*  $I_1$  à  $I_3$  et les évolutions des valeurs d'indices avec le nombre de molécules générées sont présentées sur la Figure S4. Le Tableau S5 présente les valeurs moyennes des indices à 50k molécules générées. Dans un premier temps, chaque génération produit des molécules projetées dans des zones différentes de  $\mathbb{C}$  (les valeurs des indices  $I_1$  et  $I_{2a}$  à  $I_{2d}$  sont faibles). Nous attribuons ce caractère au fait

que la diversité d'opérations réalisables pour produire de nouvelles molécules est élevée. Ensuite, les différentes générations commencent à produire des molécules de manière similaire, avec une distribution similaire des molécules dans les cubes (les valeurs des indices  $I_1$  à  $I_{2d}$  augmentent avec le nombre de molécules générées). Les valeurs des indices entre les différentes générations sont similaires dès  $10^2$  à  $10^3$  molécules, ce qui démontre que la méthode G1b produit des molécules projetées dans  $\mathbb{C}$  de manière similaire sur plusieurs générations. L'information donnée par nos indices est reproductible (écarts moyens inférieurs à 1% des valeurs moyennes observées).

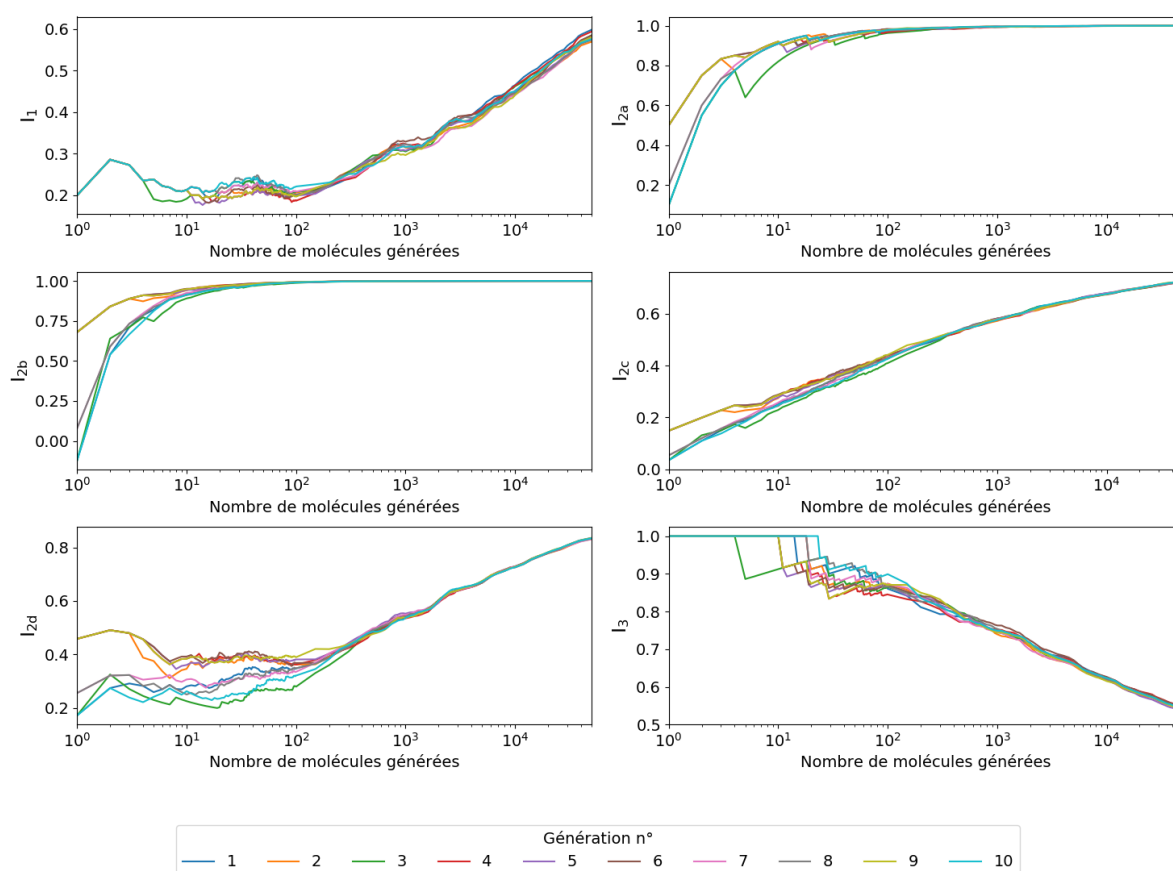


Figure S4 : Évolutions des indices  $I_1$  à  $I_3$  avec le nombre de molécules générées par dix générations utilisant la méthode G1b.

<b>Indice</b>	<b>Valeur moyenne</b>
$I_1$	$0,591 \pm 0,007$
$I_{2a}$	$0,999 \pm 0,001$
$I_{2b}$	$0,999 \pm 0,001$
$I_{2c}$	$0,724 \pm 0,001$
$I_{2d}$	$0,838 \pm 0,001$
$I_3$	$0,542 \pm 0,002$

Tableau S5 : Valeurs moyennes des indices et écarts moyens, à  $5.10^4$  molécules générées, sur dix générations avec la méthode G1b.

## Chapitre 7. Bibliographie

---

- (1) Loi "Grenelle II" portant engagement national pour l'environnement : L. n° 2010-788, 12 jui. 2010.
- (2) IFP Energies Nouvelles. *Domaines d'activités*. Consulté le 23/02/2021 à l'adresse <https://www.ifpenouvelles.fr/ifpen/domaines-dactivites>.
- (3) Brown, F. K. Chemoinformatics: What is it and How does it Impact Drug Discovery. In ; Annual Reports in Medicinal Chemistry; Elsevier, 1998; pp 375–384. DOI: 10.1016/S0065-7743(08)61100-8.
- (4) Hann, M.; Green, R. Chemoinformatics — a new name for an old problem? *Current Opinion in Chemical Biology* **1999**, *3*, 379–383. DOI: 10.1016/S1367-5931(99)80057-X.
- (5) Participants du Workshop "Chemoinformatics in Europe: Research and Teaching". *The Obernai Declaration*. Consulté le 21/07/2021 à l'adresse <http://infochim.u-strasbg.fr/chemoinformatics/Obernai%20Declaration.pdf>.
- (6) Apodaca, R. L. *Sixty-Four Free Chemistry Databases*. Consulté le 26/04/2021 à l'adresse <https://depth-first.com/articles/2011/10/12/sixty-four-free-chemistry-databases/>.
- (7) Auteur(s) de Wikipédia. *List of chemical databases*. Consulté le 21/07/2021 à l'adresse [https://en.wikipedia.org/wiki/List\\_of\\_chemical\\_databases](https://en.wikipedia.org/wiki/List_of_chemical_databases).
- (8) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic acids research* **2016**, *44*, D1202-13. DOI: 10.1093/nar/gkv951.
- (9) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC : A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **2012**, *52*, 1757–1768. DOI: 10.1021/ci3001277.
- (10) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL : A Large-scale Bioactivity Database for Drug Discovery. *Nucleic acids research* **2012**, *40*, D1100-7. DOI: 10.1093/nar/gkr777.

- (11) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13*, 540–554. DOI: 10.1002/cmdc.201700561.
- (12) Chemical Abstracts Service. *SciFinder*. Consulté le 21/07/2021 à l'adresse <https://scifinder.cas.org/>.
- (13) Gabrielson, S. W. *SciFinder*. *Journal of the Medical Library Association* **2018**, *106*. DOI: 10.5195/jmla.2018.515.
- (14) Westmoreland, P.; Kollman, P.; Chaka, A.; Cummings, P.; Morokuma, K.; Neurock, M.; Stechel, E.; Vashishta, P. *Applications of Molecular and Materials Modeling*. Consulté le 07/02/2021 à l'adresse <https://apps.dtic.mil/sti/pdfs/ADA467500.pdf>.
- (15) *EC Regulation no. 1907/2006 concerning the registration, evaluation, authorization and restriction of chemicals (REACH)*, 2006.
- (16) Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; Hemptinne, J.-C. de; Ungerer, P.; Rousseau, B.; Adamo, C. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chemical reviews* **2015**, *115*, 13093–13164. DOI: 10.1021/acs.chemrev.5b00215.
- (17) *Globally harmonized system of classification and labelling of chemicals (GHS)*, Eighth revised edition; United Nations, New York, 2019.
- (18) Quintero, F. A.; Patel, S. J.; Muñoz, F.; Sam Mannan, M. Review of Existing QSAR/QSPR Models Developed for Properties Used in Hazardous Chemicals Classification System. *Industrial & Engineering Chemistry Research* **2012**, *51*, 16101–16115. DOI: 10.1021/ie301079r.
- (19) Lunghini, F.; Marcou, G.; Gantzer, P.; Azam, P.; Horvath, D.; van Miert, E.; Varnek, A. Modelling of Ready Biodegradability Based on Combined Public and Industrial Data Sources. *SAR and QSAR in environmental research* **2019**, *31*, 171–186. DOI: 10.1080/1062936X.2019.1697360.
- (20) Lunghini, F.; Marcou, G.; Azam, P.; Patoux, R.; Enrici, M. H.; Bonachera, F.; Horvath, D.; Varnek, A. QSPR models for bioconcentration factor (BCF): are they able to predict data of industrial interest? *SAR and QSAR in environmental research* **2019**, *30*, 507–524. DOI: 10.1080/1062936X.2019.1626278.



- (21) Neves, B. J.; Braga, R. C.; Melo-Filho, C. C.; Moreira-Filho, J. T.; Muratov, E. N.; Andrade, C. H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Frontiers in pharmacology* **2018**, *9*, 1275. DOI: 10.3389/fphar.2018.01275.
- (22) Cheng, F.; Li, W.; Liu, G.; Tang, Y. In silico ADMET prediction: recent advances, current challenges and future trends. *Current topics in medicinal chemistry* **2013**, *13*, 1273–1289. DOI: 10.2174/15680266113139990033.
- (23) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *Journal of cheminformatics* **2009**, *1*, 8. DOI: 10.1186/1758-2946-1-8.
- (24) Creton, B. Chemoinformatics at IFP Energies Nouvelles : Applications in the Fields of Energy, Transport, and Environment. *Molecular informatics* **2017**, *36*, 1700028. DOI: 10.1002/minf.201700028.
- (25) Donaldson, E. C.; Chilingarian, G. V.; Yen, T. F. *Enhanced Oil Recovery, II : Processes and Operations*; Developments in Petroleum Science, v. 17B; Elsevier Science, Amsterdam, 2014.
- (26) Bourgeois, M.; Darche, G. *A l'avant-garde de la simulation des procédés de récupération améliorée*. Consulté le 16/01/2020 à l'adresse <https://www.ep.total.com/fr/lavant-garde-de-la-simulation-des-procedes-de-recuperation-amelioree>.
- (27) Muller, C.; Maldonado, A. G.; Varnek, A.; Creton, B. Prediction of Optimal Salinities for Surfactant Formulations Using a Quantitative Structure–Property Relationships Approach. *Energy & Fuels* **2015**, *29*, 4281–4288. DOI: 10.1021/acs.energyfuels.5b00825.
- (28) Phan, A.; Doonan, C. J.; Uribe-Romo, F. J.; Knobler, C. B.; O'Keeffe, M.; Yaghi, O. M. Synthesis, structure, and carbon dioxide capture properties of zeolitic imidazolate frameworks. *Accounts of chemical research* **2010**, *43*, 58–67. DOI: 10.1021/ar900116g.
- (29) Amrouche, H.; Creton, B.; Siperstein, F.; Nieto-Draghi, C. Prediction of thermodynamic properties of adsorbed gases in zeolitic imidazolate frameworks. *RSC Advances* **2012**, *2*, 6028. DOI: 10.1039/c2ra00025c.
- (30) Galvelis, R.; Slater, B.; Chaudret, R.; Creton, B.; Nieto-Draghi, C.; Mellot-Draznieks, C. Impact of functionalized linkers on the energy landscape of ZIFs. *CrystEngComm* **2013**, *15*, 9603. DOI: 10.1039/C3CE41103F.

- (31) Saldana, D. A. Méthodes d'apprentissage automatique pour l'aide à la formulation : Carburants Alternatifs pour l'Aéronautique. Thèse de doctorat, Paris, 2013. Consulté le 01/09/2019 à l'adresse <http://www.theses.fr/2013PA066346>.
- (32) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design : A Molecular Modeling Perspective. *Medicinal Research Reviews* **1996**, *16*, 3–50. DOI: 10.1002/(SICI)1098-1128(199601)16:1<3:AID-MED1>3.0.CO;2-6.
- (33) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 675–679. DOI: 10.1007/s10822-013-9672-4.
- (34) Gantzer, P.; Creton, B.; Nieto-Draghi, C. Inverse-QSPR for De Novo Design : A Review. *Molecular informatics* **2020**, *39*, 1900087. DOI: 10.1002/minf.201900087.
- (35) Guyton-Morveau, L. B. Mémoire sur les dénominations chimiques, la nécessité d'en perfectionner le système et les règles pour y parvenir. *Observations sur la Physique* **1782**, *19*, 370–382.
- (36) Baudrimont, A.; Trébuchet, A. *Du sucre et de sa fabrication: suivi d'un précis de la législation qui régit cette industrie*; J.B. Baillière, Paris, 1841.
- (37) Guyton-Morveau, L. B.; Lavoisier, A. L. de; Berthollet, C. L.; Cuchet, G. J. *Méthode de nomenclature chimique, proposée par MM. de Morveau, Lavoisier, Bertholet, & de Fourcroy*; Cuchet, Paris, 1787.
- (38) Doré, M. *Leçons de chimie élémentaire appliquées aux arts industriels et faites aux Ouvriers du XIIIe arrondissement : A l'usage des élèves de rhétorique scientifique, des aspirants aux grades des facultés et aux écoles du gouvernement*; Carilian-Goeury et Dalmont, Paris, 1857.
- (39) Kersaint, G. Aperçu sur les nomenclatures en chimie. *Revue d'histoire de la pharmacie* **1968**, *56*, 203–206. DOI: 10.3406/pharm.1968.7790.
- (40) Favre, H. A.; Powell, W. H. *Nomenclature of organic chemistry : IUPAC recommendations and preferred names 2013*; Royal Society of Chemistry, Cambridge, 2014.
- (41) Capitolis, J.; Delacroix, S.; Frogneux, X.; Medina, E.; Rey, N.; Tinat, L.; Carencó, S. Précis de nomenclature en chimie inorganique. *L'actualité chimique* **2019**, *437*, 12–18.

- (42) Agence européenne des produits chimiques. *Ethanol - Brief ECHA profile*. Consulté le 03/05/2021 à l'adresse <https://echa.europa.eu/brief-profile/-/briefprofile/100.000.526>.
- (43) American Chemical Society. *CAS REGISTRY - The gold standard for chemical substance information*. Consulté le 17/01/2020 à l'adresse <https://www.cas.org/support/documentation/chemical-substances>.
- (44) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order - An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *Journal of Chemical Information and Modeling* **2015**, *55*, 2111–2120. DOI: 10.1021/acs.jcim.5b00543.
- (45) Landrum, G. *RDKit: Open-Source Cheminformatics*. Consulté le 26/04/2021 à l'adresse <http://www.rdkit.org/>.
- (46) Dalke, A.; Hert, J.; Kramer, C. mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. *Journal of Chemical Information and Modeling* **2018**, *58*, 902–910. DOI: 10.1021/acs.jcim.8b00173.
- (47) O'Boyle, N. M. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics* **2012**, *4*, 22. DOI: 10.1186/1758-2946-4-22.
- (48) Lo, Y.-C.; Rensi, S. E.; Tornø, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug discovery today* **2018**, *23*, 1538–1546. DOI: 10.1016/j.drudis.2018.05.010.
- (49) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Molecular informatics* **2010**, *29*, 855–868. DOI: 10.1002/minf.201000099.
- (50) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Current Computer Aided-Drug Design* **2008**, *4*, 191–198. DOI: 10.2174/157340908785747465.
- (51) Daylight. *SMARTS - A Language for Describing Molecular Patterns*. Consulté le 01/09/2019 à l'adresse <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
- (52) Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. The distance matrix in chemistry. *Journal of Mathematical Chemistry* **1992**, *11*, 223–258. DOI: 10.1007/BF01164206.

- (53) Randić, M.; Zupan, J. On Interpretation of Well-Known Topological Indices. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 550–560. DOI: 10.1021/ci000095o.
- (54) Roy, K. *Advances in QSAR modeling : Applications in pharmaceutical, chemical, food, agricultural and environmental sciences*; Challenges and advances in computational chemistry and physics, volume 24; Springer, Cham, Switzerland, 2017.
- (55) Devillers, J.; Balaban, A. T. *Topological indices and related descriptors in QSAR and QSPR*; Gordon and Breach, Amsterdam, 1999.
- (56) Kier, L. B. A Shape Index from Molecular Graphs. *Quantitative Structure-Activity Relationships* **1985**, *4*, 109–116. DOI: 10.1002/qsar.19850040303.
- (57) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bulletin of the Chemical Society of Japan* **1971**, *44*, 2332–2339. DOI: 10.1246/bcsj.44.2332.
- (58) Randic, M. Characterization of molecular branching. *Journal of the American Chemical Society* **1975**, *97*, 6609–6615. DOI: 10.1021/ja00856a001.
- (59) Miyao, T.; Kaneko, H.; Funatsu, K. Ring system-based chemical graph generation for de novo molecular design. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 425–446. DOI: 10.1007/s10822-016-9916-1.
- (60) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A practical guide to support vector classification*. Consulté le 01/07/2021 à l'adresse <https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- (61) Cornuéjols, A.; Miclet, L. *Apprentissage artificiel : Concepts et algorithmes*, 2e éd.; Algorithmes; Eyrolles, Paris, 2015.
- (62) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297. DOI: 10.1023/A:1022627411411.
- (63) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1996; pp 155–161.
- (64) Awad, M.; Khanna, R. Support Vector Regression. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress, 2015; pp 67–80. DOI: 10.1007/978-1-4302-5990-9\_4.

- (65) Gaspar, H. A.; Baskin, I. I.; Varnek, A. Visualization of a Multidimensional Descriptor Space. In *Frontiers in Molecular Design and Chemical Information Science - Herman Skolnik Award Symposium 2015: Jürgen Bajorath*; ACS Symposium Series, Vol. 1222; American Chemical Society, 2016; pp 243–267. DOI: 10.1021/bk-2016-1222.ch012.
- (66) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, 2, 37–52. DOI: 10.1016/0169-7439(87)80084-9.
- (67) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of Chemical Information and Computer Sciences* **2001**, 41, 233–245. DOI: 10.1021/ci0001482.
- (68) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM) : Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Molecular informatics* **2012**, 31, 301–312. DOI: 10.1002/minf.201100163.
- (69) Gramatica, P.; Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. *Journal of Chemical Information and Modeling* **2016**, 56, 1127–1131. DOI: 10.1021/acs.jcim.6b00088.
- (70) Roy, K.; Kar, S.; Das, R. N. *A Primer on QSAR/QSPR Modeling : Fundamental Concepts*; EBL-Schweitzer; Springer International Publishing, Cham, Suisse, 2015.
- (71) Afendras, G.; Markatou, M. Optimality of training/test size and resampling effectiveness in cross-validation. *Journal of Statistical Planning and Inference* **2019**, 199, 286–301. DOI: 10.1016/j.jspi.2018.07.005.
- (72) Marcot, B. G.; Hanea, A. M. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics* **2020**. DOI: 10.1007/s00180-020-00999-9.
- (73) Marwala, T.; Leke, C. A. *Handbook of machine learning : Optimization and Decision Making*, Volume 2; World Scientific, Singapore, 2020.
- (74) Sra, S.; Nowozin, S.; Wright, S. J. *Optimization for machine learning*; Neural information processing series; MIT Press, Cambridge, Mass., 2012.
- (75) Langouët, H. Constraints Derivative-Free Optimization : Two Industrial Applications in Reservoir Engineering and in Engine Calibration. Thèse de doctorat, Université Nice Sophia

Antipolis, 2011. Consulté le 26/04/2021 à l'adresse <https://tel.archives-ouvertes.fr/tel-00671987>.

(76) Sinoquet, D.; Langouët, H.; Da Veiga, S. A Derivative Free Optimization Method for Reservoir Characterization Inverse Problem. In *72nd EAGE Conference and Exhibition incorporating SPE EUROPEC 2010*; European Association of Geoscientists & Engineers, 2010. DOI: 10.3997/2214-4609.201401002.

(77) Siarry, P. *Métaheuristiques*, Nouvelle édition; Algorithmes; Eyrolles, Paris, 2014.

(78) Zitzler, E.; Deb, K.; Thiele, L. Comparison of multiobjective evolutionary algorithms : Empirical results. *Evolutionary computation* **2000**, 8, 173–195. DOI: 10.1162/106365600568202.

(79) Holland, J. H., Ed. *Adaptation in natural and artificial systems : An introductory analysis with applications to biology, control, and artificial intelligence*; Complex adaptive systems; The MIT Press, 1992.

(80) Roy, K.; Kar, S.; Ambure, P. On a Simple Approach for Determining Applicability Domain of QSAR Models. *Chemometrics and Intelligent Laboratory Systems* **2015**, 145, 22–29. DOI: 10.1016/j.chemolab.2015.04.013.

(81) Ross, S. M. *Introduction to probability models*, 10th edition; Elsevier; Academic Press, Amsterdam, 2010.

(82) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: a Unified Approach to the Applicability Domain Problem of QSAR Models. *Journal of Chemical Information and Modeling* **2009**, 49, 1762–1776. DOI: 10.1021/ci9000579.

(83) Chang, C.-C.; Lin, C.-J. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* **2011**, 2, 1–27.

(84) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, 12, 2825–2830.

(85) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, 9, 90–95. DOI: 10.1109/MCSE.2007.55.

(86) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, VanderPlas, Jake; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272. DOI: 10.1038/s41592-019-0686-2.

(87) Gaspar, H. A. Cartographie de l'espace chimique. Thèse de doctorat, Université de Strasbourg, 2015. Consulté à l'adresse <https://tel.archives-ouvertes.fr/tel-01292573>.

(88) Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of libsvm Models. *Challenges* **2014**, *5*, 450–472. DOI: 10.3390/challe5020450.

(89) Phoon, L. Y.; Mustaffa, A. A.; Hashim, H.; Mat, R. A Review of Flash Point Prediction Models for Flammable Liquid Mixtures. *Industrial & Engineering Chemistry Research* **2014**, *53*, 12553–12565. DOI: 10.1021/ie501233g.

(90) *Regulation (EC) N° 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) N° 1907/2006*, L.353, 2008.

(91) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Pidol, L.; Jeuland, N.; Creton, B. Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. *Energy & Fuels* **2011**, *25*, 3900–3908. DOI: 10.1021/ef200795j.

(92) Rowley, R. L.; Wilding, W. V.; Oscarson, J. L.; Yang, Y.; Zundel, N. A.; Daubert, T. E.; Danner, R. P. DIPPR Data Compilation of Pure Compound Properties. *Design Institute for Physical Properties* **2003**.

(93) Wray, H. A. Statement of Harry A. Wray, Consultant Flammability of Liquids. In *Authorizations and Other Amendments to the Consumer Product Safety Act: Hearings Before the Subcommittee on Consumer Protection and Finance of the Committee on Interstate and Foreign Commerce, House of Representatives, Ninety-fifth Congress, Second Session*; U.S. Government Printing Office, February 24 and 28, 1978; pp 171–175.

- (94) Gantzer, P.; Creton, B.; Nieto-Draghi, C. Comparisons of Molecular Structure Generation Methods Based on Fragment Assemblies and Genetic Graphs. *Journal of Chemical Information and Modeling*, 2021.
- (95) Gani, R.; Brignole, E. A. Molecular Design of Solvents for Liquid Extraction Based on UNIFAC. *Fluid Phase Equilibria* **1983**, *13*, 331–340. DOI: 10.1016/0378-3812(83)80104-6.
- (96) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE Journal* **1975**, *21*, 1086–1099. DOI: 10.1002/aic.690210607.
- (97) Brignole, E. A.; Bottini, S. B.; Gani, R. A Strategy for the Design and Selection of Solvents for Separation Processes. *Fluid Phase Equilibria* **1986**, *29*, 125–132. DOI: 10.1016/0378-3812(86)85016-6.
- (98) Pretel, E. J.; López, P. A.; Bottini, S. B.; Brignole, E. A. Computer-Aided Molecular Design of Solvents for Separation Processes. *AIChE Journal* **1994**, *40*, 1349–1360. DOI: 10.1002/aic.690400808.
- (99) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A Method for Automatic Generation of Novel Chemical Structures and its Potential Applications to Drug Discovery. *Journal of Chemical Information and Modeling* **1991**, *31*, 527–530. DOI: 10.1021/ci00004a016.
- (100) Peironcely, J. E.; Rojas-Chertó, M.; Fichera, D.; Reijmers, T.; Coulier, L.; Faulon, J.-L.; Hankemeier, T. OMG: Open Molecule Generator. *Journal of cheminformatics* **2012**, *4*, 21. DOI: 10.1186/1758-2946-4-21.
- (101) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society* **2009**, *131*, 8732–8733. DOI: 10.1021/ja902302h.
- (102) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875. DOI: 10.1021/ci300415d.
- (103) Clark, D. E.; Firth, M. A.; Murray, C. W. MOLMAKER : De Novo Generation of 3D Databases for Use in Drug Design. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 137–145. DOI: 10.1021/ci9502055.



- (104) Makino, S.; Ewing, T. J.; Kuntz, I. D. DREAM++: Flexible docking program for virtual combinatorial libraries. *Journal of Computer-Aided Molecular Design* **1999**, *13*, 513–532. DOI: 10.1023/A:1008066310669.
- (105) ChemAxon. *Reaction Based Enumeration*. Consulté le 18/09/2019 à l'adresse <https://docs.chemaxon.com/display/docs/Reaction+Based+Enumeration>.
- (106) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm To Suggest Combinatorial Libraries. *Journal of Chemical Information and Modeling* **1995**, *35*, 310–320. DOI: 10.1021/ci00024a021.
- (107) Pilia, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *Journal of Chemical Information and Modeling* **2019**, *59*, 5013–5025. DOI: 10.1021/acs.jcim.9b00807.
- (108) Torkamanian-Afshar, M.; Nematzadeh, S.; Tabar zad, M.; Najafi, A.; Lanjanian, H.; Masoudi-Nejad, A. In silico design of novel aptamers utilizing a hybrid method of machine learning and genetic algorithm. *Molecular diversity* **2021**, *25*, 1395–1407. DOI: 10.1007/s11030-021-10192-9.
- (109) Khazaal, A. S.; Sprinborg, M.; Fan, C.; Huwig, K. Application of an inverse-design method for designing new branched thiophene oligomers for bulk-heterojunction solar cells. *Computational Condensed Matter* **2020**, *25*, e00503. DOI: 10.1016/j.cocom.2020.e00503.
- (110) Nachbar, R. Molecular evolution: a hierarchical representation for chemical topology and its automated manipulation. In *Proc. of the Third Annual Genetic Programming Conference*, 1998.
- (111) Globus, A.; Lawton, J.; Wipke, T. Automatic Molecular Design Using Evolutionary Techniques. *Nanotechnology* **1999**, *10*, 290. DOI: 10.1088/0957-4484/10/3/312.
- (112) Chu, Y.; He, X. MoleGear : A Java-Based Platform for Evolutionary De Novo Molecular Design. *Molecules* **2019**, *24*, 1444. DOI: 10.3390/molecules24071444.
- (113) Lameijer, E.-W.; Bäck, T.; Kok, J. N.; Ijzerman, A. P. Evolutionary Algorithms in Drug Design. *Natural Computing* **2005**, *4*, 177–243. DOI: 10.1007/s11047-004-5237-8.
- (114) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse problem in QSAR/QSPR studies for the case of topological indexes characterizing

- molecular shape (Kier indices). *Journal of Chemical Information and Modeling* **1993**, *33*, 630–634. DOI: 10.1021/ci00014a017.
- (115) Skvortsova, M. I.; Baskin, I. I.; Palyulin, V. A.; Slovokhotova, O. L.; Zefirov, N. S. Structural design inverse problems for topological indices in QSAR/QSPR studies. *AIP Conference Proceedings* **1995**, *330*, 486–499. DOI: 10.1063/1.47751.
- (116) Baskin, I.; Gordeeva, E. V.; Devdaria, R.; Zefirov, N.; Palyulin, V.; Stankevich, M. Solving the Inverse Problem of Structure-Property Relations for the Case of Topological Indexes. *Doklady Chemistry* **1989**, *307*, 217–220.
- (117) Funatsu, K.; Miyao, T.; Arakawa, M. Systematic Generation of Chemical Structures for Rational Drug Design Based on QSAR Models. *Current Computer Aided-Drug Design* **2011**, *7*, 1–9. DOI: 10.2174/157340911793743556.
- (118) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug discovery today* **2018**, *23*, 1241–1250. DOI: 10.1016/j.drudis.2018.01.039.
- (119) Bjerrum, E. J.; Threlfall, R. *Molecular Generation with Recurrent Neural Networks (RNNs)*. Consulté le 26/04/2021 à l'adresse <http://arxiv.org/pdf/1705.04612v2>.
- (120) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
- (121) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS central science* **2018**, *4*, 268–276. DOI: 10.1021/acscentsci.7b00572.
- (122) Sattarov, B.; Baskin, I. I.; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *Journal of Chemical Information and Modeling* **2019**, *59*, 1182–1196. DOI: 10.1021/acs.jcim.8b00751.
- (123) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. *Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models*. Consulté le 26/04/2021 à l'adresse <https://arxiv.org/abs/1705.10843v3>.

- (124) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS central science* **2018**, *4*, 120–131. DOI: 10.1021/acscentsci.7b00512.
- (125) Langevin, M.; Minoux, H.; Levesque, M.; Bianciotto, M. Scaffold-Constrained Molecular Generation. *Journal of Chemical Information and Modeling* **2020**, *60*, 5637–5646. DOI: 10.1021/acs.jcim.0c01015.
- (126) Bung, N.; Krishnan, S. R.; Bulusu, G.; Roy, A. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future medicinal chemistry* **2021**. DOI: 10.4155/fmc-2020-0262.
- (127) Santana, M. V. S.; Silva-Jr, F. P. De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. *BMC chemistry* **2021**, *15*, 8. DOI: 10.1186/s13065-021-00737-2.
- (128) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Accounts of chemical research* **2021**, *54*, 263–270. DOI: 10.1021/acs.accounts.0c00699.
- (129) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *Journal of cheminformatics* **2019**, *11*, 71. DOI: 10.1186/s13321-019-0393-0.
- (130) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. *Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models*. Consulté le 26/04/2021 à l'adresse <http://arxiv.org/pdf/1811.12823v5>.
- (131) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling* **2019**, *59*, 1096–1108. DOI: 10.1021/acs.jcim.8b00839.
- (132) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79–86. DOI: 10.1214/aoms/1177729694.
- (133) Bjerrum, E. J. *SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules*. Consulté le 26/04/2021 à l'adresse <http://arxiv.org/pdf/1703.07076v2>.

- (134) Vidal, M.; Rogers, W. J.; Holste, J. C.; Mannan, M. S. A Review of Estimation Methods for Flash Points and Flammability Limits. *Process Safety Progress* **2004**, *23*, 47–55. DOI: 10.1002/prs.10004.
- (135) Levy, J. M. *Hazmat Chemistry Study Guide (Second Edition)*; Firebelle Productions, Campbell, CA, 2005.
- (136) Bhattacharyya, A. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* **1943**, *35*, 99–109.
- (137) Shemyakin, A. Hellinger Distance and Non-informative Priors. *Bayesian Analysis* **2014**, *9*, 923–938. DOI: 10.1214/14-BA881.
- (138) Inoue, T.; Tanaka, K.; Kotera, M.; Funatsu, K. Improvement of the Structure Generator DA ECS with Respect to Structural Diversity. *Molecular informatics* **2021**, *40*, 2000225. DOI: 10.1002/minf.202000225.
- (139) Steinbeck, C. SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1500–1507. DOI: 10.1021/ci000407n.
- (140) Gaspar, H. A.; Breen, G. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC bioinformatics* **2019**, *20*, 116. DOI: 10.1186/s12859-019-2680-1.
- (141) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 1087–1108. DOI: 10.1007/s10822-015-9882-z.
- (142) Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *Journal of Chemical Information and Modeling* **2019**, *59*, 564–572. DOI: 10.1021/acs.jcim.8b00650.
- (143) Kaneko, H. Data Visualization, Regression, Applicability Domains and Inverse Analysis Based on Generative Topographic Mapping. *Molecular informatics* **2019**, *38*, 1800088. DOI: 10.1002/minf.201800088.

## Liste des figures

Figure 1 : Schéma représentatif des travaux présentés dans ce manuscrit de thèse.....	14
Figure 2 : Fonctionnement d'un modèle de classification par SVM. Les vecteurs de chaque classe sont représentés sous forme de "x" pour la première classe et de "o" pour la seconde.	23
Figure 3 : Fonctionnement d'un modèle de régression par SVR. ....	23
Figure 4 : Modèle de validation croisée proposé pour l'optimisation des paramètres QSPR..	26
Figure 5 : Méthodologie employée pour la création et l'optimisation des SVR.....	29
Figure 6 : Méthodologie employée pour la création et l'optimisation des ACP.....	30
Figure 7 : Méthodologie employée pour la création et l'optimisation des cartes GTM. ....	30
Figure 8 : Distributions des valeurs de point d'éclair dans le jeu de donnée considéré pour le travail de thèse, pour la totalité du jeu, le jeu d'entraînement, et le jeu de test. ....	32
Figure 9 : Distribution des molécules du jeu de données (point d'éclair) en fonction de leur famille chimique. Adapté avec permission, issu de Gantzer <i>et coll.</i> <sup>94</sup> Copyright 2021 American Chemical Society.....	32
Figure 10 : Comparaison entre les valeurs de point d'éclair, expérimentales et prédites par le modèle obtenu avec les descripteurs SMF t312u3, pour les molécules des ensembles d'entraînement et de test. ....	35
Figure 11 : Représentation de l'espace $\mathbb{C}$ (en vert), issu de la PCA du jeu de données concernant le point d'éclair. Les molécules initiales y sont projetées en gris et sont englobées par une enveloppe convexe représentée en violet. Réimprimé avec permission, issu de Gantzer <i>et coll.</i> <sup>94</sup> Copyright 2021 American Chemical Society.....	37
Figure 12 : Carte GTM construite à partir des descripteurs t212u2 pour la base de point d'éclair. ....	39
Figure 13 : Représentation du graphe moléculaire, notation SMILES et encodage sous forme de vecteurs « one-hot » de la molécule de 4-tert-Butylcatechol. Adapté avec permission de Gantzer <i>et coll.</i> <sup>34</sup> , Copyright Wiley 2019.....	47
Figure 14 : Représentation d'un VAE et de son fonctionnement pour la génération moléculaire. Les molécules initiales, sous forme de SMILES, sont présentées avec une couleur fonction de leur propriété (rouge : indésirable, vert : désirable). Les chaînes SMILES sont ensuite encodées	

dans un espace latent. Une région de l'espace latent est sélectionnée (ellipse verte), et les vecteurs de cette région sont décodés en SMILES. Certains SMILES décodés sont écartés, soit car ils ne correspondent pas à des molécules valides, soit car les molécules résultantes ne possèdent pas la propriété désirée. Adapté avec permission de Gantzer et coll. <sup>34</sup> , Copyright Wiley 2019.....	48
Figure 15 : Opérateurs de modifications implémentés pour application sur les graphes moléculaires. Un exemple est donné pour chaque opérateur, et les modifications effectuées sont mises en valeur par des patchs rouges.....	66
Figure 16 : Schéma de la méthode de génération par modifications successives de structures regroupant tous les opérateurs.....	77
Figure 17 : Schéma de la méthode de génération par modifications successives de structures, avec contraintes sur la propriété.....	78
Figure 18 : Nombre de molécules valides, nouvelles et uniques générées après 2 000 tentatives pour chacun des 10 vecteurs latents, ainsi qu'en moyenne, en fonction de la valeur de $\sigma_G$ . ...	81
Figure 19 : Distribution des valeurs d'Accessibilité Synthétique (AS), de poids moléculaire, et de point d'éclair prédit dans les jeux de molécules initiales (du modèle QSPR) ainsi que générées par F1b et G1a, exprimé en pourcentage de molécules de chaque jeu.....	88
Figure 20 : Projection des 50k premières molécules générées dans l'espace $\mathbb{C}$ , pour la méthode G1a (à gauche) et F1b (à droite).....	89
Figure 21 : Représentation de la densité spatiale des 10 000 premières molécules générées par la méthode G1a (à gauche) et F1b (à droite) dans l'espace $\mathbb{C}$ , à l'aide de cubes de taille unitaire. ....	90
Figure 22 : Moyennes des ratios des molécules générées par G1a et par G1b dans chaque ensemble de cubes le long de la composante principale (PC) 3 (diagramme a), PC 2 (b) et PC 1 (c). ....	90
Figure 23 : Évolutions des indices (a) $I1$ , (b) $I2a$ , (c) $I2b$ , (d) $I2c$ , (e) $I2d$ et (f) $I3$ avec le nombre de molécules générées par les méthodes F0, F1a, F1b, G1a et G1b. Adapté avec permission, issu de Gantzer <i>et coll.</i> <sup>94</sup> Copyright 2021 American Chemical Society.....	96
Figure 24 : Distribution des valeurs de point d'éclair dans l'ensemble d'apprentissage du modèle QSPR.....	100

Figure 25 : Variation des indices $I1$ à $I4$ en fonction du nombre de molécules générées dans l'intervalle [200 K ; 300 K]. Adapté avec permission, issu de Gantzer <i>et coll.</i> <sup>94</sup> Copyright 2021 American Chemical Society.....	104
Figure 26 : Variation des indices $I1$ à $I4$ en fonction du nombre de molécules générées dans l'intervalle [300 K ; 400 K]. Adapté avec permission, issu de Gantzer <i>et coll.</i> <sup>94</sup> Copyright 2021 American Chemical Society.....	105
Figure 27 : Variation des indices $I1$ à $I4$ en fonction du nombre de molécules générées dans l'intervalle [400 K ; 500 K]. Adapté avec permission, issu de Gantzer <i>et coll.</i> <sup>94</sup> Copyright 2021 American Chemical Society.....	107

## Liste des tableaux

Tableau 1 : Classification des descripteurs moléculaires en fonction du niveau de description considéré.....	19
Tableau 2 : Les différents types de descripteurs SMF considérés, avec description et exemple. ....	20
Tableau 3 : Performances des modèles (RMSE) pour prédire l'indice de cétane.....	33
Tableau 4 : Performances des modèles (RMSE) pour prédire le point d'éclair.....	34
Tableau 5 : Valeurs des hyperparamètres optimisés du modèle utilisant les descripteurs SMF t3l2u4.....	35
Tableau 6 : Performances du modèle QSPR utilisé pour la suite de la thèse et du modèle de Saldana <i>et coll.</i> , sur plusieurs jeux de molécules. ....	36
Tableau 7 : Valeurs limites des espaces initial et étendu sur chaque axe, dans l'espace formé par les trois premières PC de la PCA du jeu de données concernant le point d'éclair.....	38
Tableau 8 : Valeurs optimisées des paramètres de la carte GTM utilisant les descripteurs t3l2u2. # signifie « nombre ».....	38
Tableau 9 : Performances du modèle GTM pour la prédiction du point d'éclair. ....	40
Tableau 10 : Caractéristiques des différentes plateformes d'analyses comparatives existantes. ....	51
Tableau 11 : Les différentes catégories de molécules générées et leurs conditions. ....	55
Tableau 12 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par assemblage libre de fragments simples. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	59
Tableau 13 : Distribution du type de liaisons dans les molécules du jeu initial. ....	60
Tableau 14 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par assemblage de fragments simples avec contraintes sur le choix des liaisons entre fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen.....	60



Tableau 15 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par assemblage libre de fragments SMF. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	61
Tableau 16 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par assemblage de fragments SMF avec contraintes sur le choix des liaisons entre fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	62
Tableau 17 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par assemblage de fragments SMF avec contraintes sur le choix des fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	63
Tableau 18 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par assemblage de fragments SMF avec contraintes sur le choix des fragments et sur des liaisons entre fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	64
Tableau 19 : Les différentes variations de génération par assemblage de fragments conservées. ....	65
Tableau 20 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par la méthode de base d'ajout et suppression de fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	68
Tableau 21 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par la première variation d'ajout et suppression de fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	69
Tableau 22 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par la deuxième variation d'ajout et suppression de fragments. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	70
Tableau 23 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par la méthode de base de croisement. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	71
Tableau 24 : Valeurs moyennes des indices <i>individuels</i> pour dix générations par la deuxième variation de croisement. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	72

Tableau 25 : Valeurs moyennes (V) des indices <i>individuels</i> pour dix générations par mutation d'atomes et de liaisons dans la méthode de base et sa variation. Le nombre moyen de molécules (#) associées à chaque indice est également indiqué avec son écart moyen. ....	74
Tableau 26 : Valeurs moyennes (V) des indices <i>individuels</i> pour dix générations par cyclisations et ouvertures de cycle. La méthode de base et sa variation sont étudiées. Le nombre moyen de molécules (#) associées à chaque indice est également indiqué avec son écart moyen. ....	75
Tableau 27 : Valeurs moyennes des indices <i>individuels</i> pour dix générations de quatre itérations par la méthode VAE. Le nombre moyen de molécules associées à chaque indice est également indiqué avec son écart moyen. ....	82
Tableau 28 : Valeurs des indices <i>I1</i> à <i>I3</i> à 5M molécules générées par les méthodes F0, F1a, F1b, G1a et G1b. ....	97
Tableau 29 : Valeurs de huit descripteurs (parmi les descripteurs ISIDA encodant les fragments de 2 à 4 atomes reliés par des liaisons simples) des molécules d'hexane, de cyclohexane et de méthylcyclopentane. ....	99
Tableau 30 : Les différents intervalles de point d'éclair utilisés comme cibles de génération. ....	101
Tableau 31 : Spécificité moyenne et écart moyen des méthodes employées pour générer des molécules dans les différents intervalles de point d'éclair, à 50k molécules générées, sur dix générations. ....	102
Tableau 32 : Valeurs des indices <i>I1</i> à <i>I4</i> pour $3,53 \cdot 10^3$ molécules générées possédant leur point d'éclair dans l'intervalle [200 K ; 300 K[. ....	104
Tableau 33 : Valeurs des indices <i>I1</i> à <i>I4</i> pour $1,48 \cdot 10^6$ molécules générées possédant leur point d'éclair dans l'intervalle [300 K ; 400 K[, en fonction de la méthode de génération. ....	106
Tableau 34 : Valeurs des indices <i>I1</i> à <i>I4</i> pour $2,75 \cdot 10^6$ molécules générées possédant leur point d'éclair dans l'intervalle [400 K ; 500 K[, en fonction de la méthode de génération. ....	107

## Résumé :

Le recours aux QSPR (de l'anglais « Quantitative Structure-Property Relationship ») est devenu fréquent pour prédire rapidement certaines propriétés. Parallèlement, la génération virtuelle de molécules, dont font partie les inversions des modèles QSPR (i-QSPR), permet de concevoir des molécules pour des applications ciblées. Durant cette thèse, nous nous sommes d'abord intéressés à la mise en place et à l'amélioration de telles méthodes de génération. À la suite d'une étude de la littérature, trois méthodes, basées sur des assemblages de fragments (F), des modifications successives de graphes génétiques (G) et un autoencodeur variationnel (E), ont été sélectionnées pour enrichir une base de données constituée de valeurs de point d'éclair (PE) d'hydrocarbures et de composés oxygénés (785 molécules). Jusqu'à 5 millions de structures ont été générées avec chaque méthode. Nous avons également développé une approche de comparaison des méthodes de génération. Elle se base sur l'analyse de la répartition des molécules générées par chaque méthode dans un espace chimique simplifié à 3 dimensions. Les méthodes F et G ont été comparées ainsi, et nous avons mis en évidence les meilleures performances de G pour compléter de manière diverse et représentative notre base de données. Pour la génération de structures répondant à une contrainte sur le PE, nous avons mis en place une sélection du nombre de molécules à modifier par la méthode G. La restriction du nombre de molécules à modifier par G s'avère efficace pour générer des molécules avec une valeur de PE supérieure à 400 K ; tandis qu'elle dégrade les performances de génération de molécules avec une valeur de PE inférieure à 400 K.

Mots clés : QSPR, i-QSPR, génération moléculaire, comparaison de générations, de novo design, criblage virtuel, graphe génétique, assemblage de fragments, autoencodeur variationnel.

## [Development and comparison of inverse-QSPR approaches]

### Abstract :

The use of Quantitative Structure-Property Relationships (QSPR) has become frequent to quickly predict molecules' properties. Also, new chemical structures can be designed nowadays by virtual generation. The inversion of QSPR models (i-QSPR) is one of the existing virtual generation methods, which allows to obtain molecules with desired properties. During this PhD work, we first focussed on implementing and improving such methods. Following a study of the literature, three methods, based on fragments assemblies (F), successive modifications of genetic graphs (G) and a variational autoencoder (E), were selected to supplement our database. Our database gathers 785 hydrocarbons and oxygenated molecules, with their associated flash point (FP) value. Up to 5 million structures were generated by each method. Then, a method for comparing generation methods was proposed. Our approach analyses the distribution of molecules generated by each algorithm, in a small-dimensional space. The F and G methods were compared with this new tool, and we highlighted the best performances of G to complete our FP database in a diverse and representative way. For the generation of structures responding to a constraint on the FP value, we set up a selection of the number of molecules to be modified by G. The restriction of the number of molecules to be modified by G proves to be effective for the generation of molecules with a FP value greater than 400 K; while it degrades the performance of generation of molecules with a FP value lower than 400 K.

Keywords : QSPR, i-QSPR, molecular generation, generations comparison, de novo design, virtual screening, genetic graph, fragment assemblies, variational autoencoder