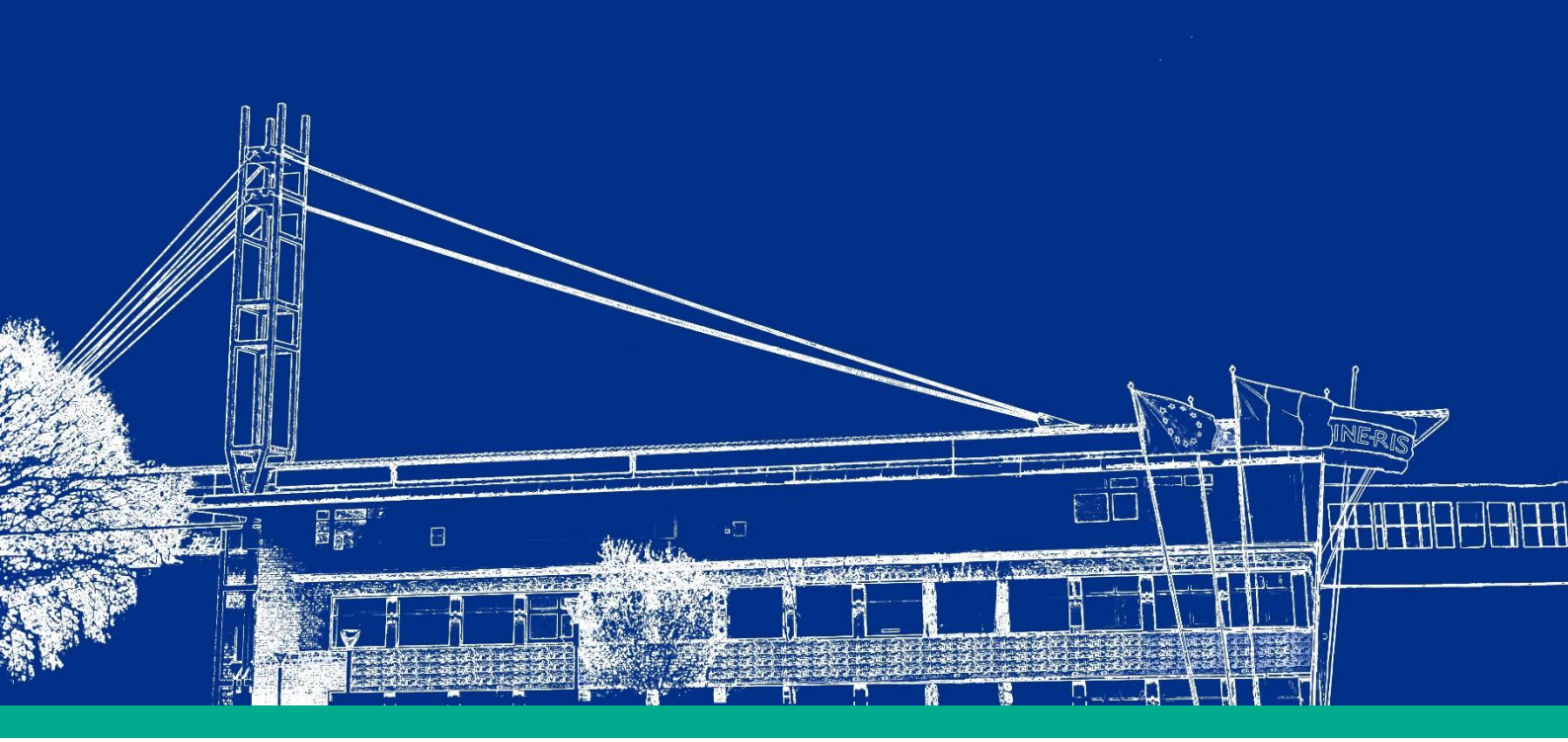

(ID Modèle = 454913)

Ineris - 206796 - 2754484 - v1.0

13/01/2023

Etat des lieux du potentiel des méthodes QSPR pour la prédiction des dangers physiques des mélanges

INERIS

maîtriser le risque
pour un développement durable

PRÉAMBULE

Le présent document a été réalisé au titre de la mission d'appui aux pouvoirs publics confiée à l'Ineris, en vertu des dispositions de l'article R131-36 du Code de l'environnement.

La responsabilité de l'Ineris ne peut pas être engagée, directement ou indirectement, du fait d'inexactitudes, d'omissions ou d'erreurs ou tous faits équivalents relatifs aux informations utilisées.

L'exactitude de ce document doit être appréciée en fonction des connaissances disponibles et objectives et, le cas échéant, de la réglementation en vigueur à la date d'établissement du document. Par conséquent, l'Ineris ne peut pas être tenu responsable en raison de l'évolution de ces éléments postérieurement à cette date. La mission ne comporte aucune obligation pour l'Ineris d'actualiser ce document après cette date.

Au vu de ses missions qui lui incombent, l'Ineris, n'est pas décideur. Les avis, recommandations, préconisations ou équivalent qui seraient proposés par l'Ineris dans le cadre des missions qui lui sont confiées, ont uniquement pour objectif de conseiller le décideur dans sa prise de décision. Par conséquent, la responsabilité de l'Ineris ne peut pas se substituer à celle du décideur qui est donc notamment seul responsable des interprétations qu'il pourrait réaliser sur la base de ce document. Tout destinataire du document utilisera les résultats qui y sont inclus intégralement ou sinon de manière objective. L'utilisation du document sous forme d'extraits ou de notes de synthèse s'effectuera également sous la seule et entière responsabilité de ce destinataire. Il en est de même pour toute autre modification qui y serait apportée. L'Ineris dégage également toute responsabilité pour chaque utilisation du document en dehors de l'objet de la mission.

Nom de la Direction en charge du rapport : DIRECTION INCENDIE DISPERSION EXPLOSION

Rédaction : FAYET Guillaume - ROTUREAU PATRICIA

Vérification : DELBAERE THIERRY; EVANNO SEBASTIEN

Approbation : Document approuvé le 13/01/2023 par PIQUETTE BERNARD

Liste des personnes ayant participé à l'étude : FAYET Guillaume, ROTUREAU Patricia, PRADAUD Isabelle

Table des matières

Résumé

Les dangers physiques de mélanges de substances chimiques, associés aux risques d'incendie ou d'explosion par exemple, sont en général caractérisés à l'aide d'outils expérimentaux. Ces essais peuvent être couteux, complexes, longs à réaliser voire dangereux pour l'opérateur.

Aussi, depuis plusieurs années et notamment avec la mise en application du règlement REACH, des méthodes prédictives de type QSAR (pour relations quantitatives structure-activité) ou QSPR (pour relations quantitatives structure-propriété) sont encouragées et utilisées comme alternatives rapides et économiques aux essais pour déterminer les dangers (éco)toxicologiques mais également physiques de substances chimiques. Or, cette approche ainsi que ses principes de développement et de validation ont été prévus pour le développement de modèles prédictifs des propriétés de produits purs. La prédiction des propriétés de mélanges (dont le développement est récent et en augmentation) présente donc de nombreux défis scientifiques et méthodologiques, additionnels à ceux de la prédiction des propriétés de produits purs.

Le présent rapport propose un état de l'art des modèles QSPR existants pour prédire les dangers physiques des mélanges. Une analyse détaillée des modèles recensés dans la littérature scientifique est proposée en focalisant sur les différents éléments clés du développement des modèles (données expérimentales disponibles et champs d'application en termes de propriétés et de familles de substances chimiques concernées, descripteurs utilisés, méthodes de développement et de validation). Sur la base de cette analyse, ce rapport dresse un bilan des potentialités et limitations des modèles actuels ainsi que des axes de progrès et perspectives vers un déploiement accru de ces nouvelles approches méthodologiques complémentaires à la caractérisation expérimentale par exemple dans la recherche de substances plus sûres (safety-by-design).

Pour citer ce document, utilisez le lien ci-après :

Institut national de l'environnement industriel et des risques, Verneuil-en-Halatte : Ineris - 206796 - v1.0 13/01/2023.

Mots-clés :

Méthodes prédictives, QSPR, dangers physiques, mélanges.

1 Introduction

La caractérisation des dangers physiques de mélanges de substances chimiques, pour appréhender le risque d'incendie ou d'explosion par exemple, est en général réalisée à l'aide d'outils expérimentaux. Ces essais de caractérisation des produits purs ou des mélanges peuvent être couteux, complexes, longs à réaliser voir dangereux pour l'opérateur. Des méthodes computationnelles sont ainsi nécessaires pour les compléter et fournir un nombre important de données pour des applications de screening dans l'industrie par exemple. Il est en effet inconcevable de tester expérimentalement un grand nombre de substances, d'autant plus lorsque qu'il s'agit de mélanges (avec une grande variation des proportions des constituants des mélanges). Aussi, depuis plusieurs années et notamment avec la mise en application du règlement REACH, des méthodes prédictives de type QSAR (pour relations quantitatives structure-activité) ou QSPR (pour relations quantitatives structure-propriété) sont encouragées et utilisées comme alternatives rapides et économiques aux essais pour déterminer les dangers (éco)toxicologiques de substances chimiques mais également physiques tels que ceux d'explosibilité et d'inflammabilité.

Ces méthodes reposent sur un principe de similitude selon lequel des molécules de structures similaires présentent des propriétés similaires. Elles consistent à mettre en place une équation mathématique entre la propriété à prédire et une série de descripteurs caractérisant la structure moléculaire des composés ciblés, comme illustré en Figure 1.

.

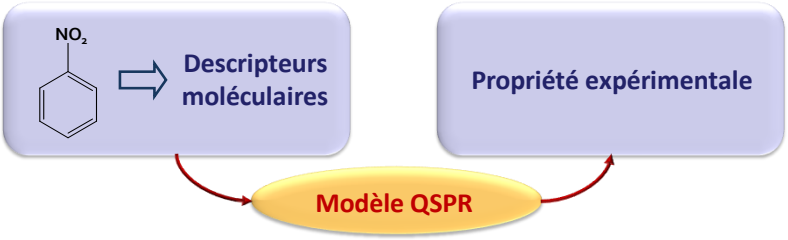

Figure 1 – Principe d'une méthode QSPR

Le développement de tels modèles est résumé brièvement ci-dessous :

Les descripteurs moléculaires peuvent être de différentes natures et plus ou moins complexes. Certains descripteurs peuvent être calculés à partir d'une simple structure semi-développée (descripteurs constitutionnels, groupes de contribution ou indices topologiques), d'autres nécessitent au préalable un calcul de structure 3D par chimie quantique (descripteurs géométriques ou quantiques). Certains logiciels permettent le calcul automatique de plusieurs centaines de descripteurs.

Différents outils de traitement de données permettent de développer de tels modèles (régressions multilinéaires, réseaux de neurones, analyses par composantes principales, arbres de décision,…) à partir d'un jeu de molécules, appelé jeu d'entraînement.

Ces outils sont alors utilisés pour mettre en place un modèle sur la base d'un jeu de données d'entraînement, puis validés selon différentes méthodes statistiques pour évaluer sa qualité d'ajustement, sa robustesse et son pouvoir prédictif sur un jeu de validation externe (de données).

Des recommandations, reposant sur cinq principes, ont été proposées par l'OCDE dès 2004 pour la validation des modèles QSPR de manière à favoriser leur utilisation dans un cadre réglementaire [1]. Des informations complémentaires sur la validation des modèles et des prédictions sont disponibles dans la note de synthèse sur les démarches de diffusion de modèles QSPR développés par l'Ineris rédigée en 2020 dans le cadre du programme d'appui [2].

L'approche QSPR ainsi que ses principes de développement et de validation ont été prévus pour le développement de modèles prédictifs des propriétés de produits purs. Il n'est donc pas étonnant que les modèles dédiés aux dangers physiques restent encore majoritairement limités aux produits purs et en particulier à la prédiction de certaines propriétés physico-chimiques dangereuses telles que le point d'éclair et le point d'ébullition dans le cas de l'inflammabilité des liquides ou des limites d'inflammabilité pour les gaz et vapeurs inflammables, des propriétés pour lesquelles les données sont le plus facilement et largement accessibles dans la littérature ou dans des bases de données [3, 4].

La prédiction des propriétés de mélanges (dont le développement est récent et en augmentation) présente en effet de nombreux défis scientifiques et méthodologiques, additionnels à ceux de la prédiction des propriétés de produits purs. En effet, il est nécessaire de tenir compte des spécificités des mélanges. Muratov [5], dès 2012, détaillait les méthodes existantes et en développement pour réaliser des prédictions QSAR/QSPR, en mettant en avant l'utilisation de descripteurs de mélanges, de type non-additifs, notamment pour prendre en compte des effets d'interaction entre constituants. Des conseils étaient également donnés concernant les méthodes d'examen et d'utilisation des bases de données mais également pour la validation appropriée de ces modèles. Dix ans plus tard, des recensements des modèles QSAR existants, capables de prédire des propriétés toxicologiques de mélanges ont été réalisés par l'Ineris [6] et par Belfield et al. [7], en examinant de manière détaillée respectivement 54 et 40 articles scientifiques pertinents (principalement la toxicité aiguë orientée davantage vers les effets sur l'environnement que sur la santé humaine pour des pesticides, produits pharmaceutiques, industriels, polluants prioritaires). A nouveau, les limites des modèles existants sont exposées et il est recommandé en conclusion de mieux prendre en compte les potentiels effets d'interaction entre constituants dans la définition des descripteurs de mélanges et de considérer des scénarios d'exposition plus réalistes (mélanges environnementaux plus complexes que le mélange théorique de différentes substances chimiques) pour améliorer la pertinence et la qualité des prédictions.

En considérant que cet état des lieux n'existe pas dans la littérature scientifique récente pour les dangers physiques d'une part, que l'Ineris a développé de premiers modèles ces dernières années pour la prédiction des points d'éclairs de mélanges organiques [8, 9] d'autre part, et qu'une grande partie des substances d'usage industriels sont des mélanges, nous proposons, dans le cadre d'un programme d'appui à l'Administration relatif à la maîtrise des risques associés aux substances, produits et procédés, un état de l'art des modèles QSPR existants pour prédire les dangers physiques des mélanges.

Dans la première partie de ce rapport, le contexte relatif aux notions de mélanges, aux spécificités associées aux dangers physiques de mélanges et aux lois de mélanges existant pour le calcul des dangers physiques des mélanges à partir de ceux de leurs constituants est synthétisé. Ensuite, le recensement des modèles QSPR existant dans la littérature scientifique est présenté et une analyse détaillée de ce recensement est réalisé en focalisant sur les différentes étapes du développement des modèles (que ce soit au niveau de la collecte de données expérimentales disponibles en termes de propriétés et de familles de substances chimiques concernées, des descripteurs utilisés, mais également des méthodologies de développement et de validation des modèles utilisées) en essayant de mettre en exergue les différences potentielles ou les similitudes avec la disponibilité des modèles existants pour la prédiction des propriétés des produits purs. Enfin, des conclusions et des perspectives sont dressées à la lueur de cet état des lieux.

2 Mélanges

2.1 Généralités

L'évaluation des risques et la caractérisation des dangers sont souvent réalisées pour des substances pures alors que l'utilisation de ces substances seules est rare dans la réalité. En effet, les êtres humains et l'environnement sont confrontés à des mélanges de substances/produits chimiques, en constante évolution. Il s'agit de produits industriels manufacturés contenant différents constituants mélangés intentionnellement (par ex, des pesticides, cosmétiques, carburants, additifs) ou accidentellement. A titre d'exemple, le secteur des carburants présente des enjeux industriels importants avec la recherche continue d'amélioration des performances (propriétés de combustion) en mélangeant des carburants dérivés de la biomasse renouvelable avec des carburants pétroliers. L'ajout d'alcools, d'éthers, de composés oxygénés est essentiel pour améliorer les propriétés (d'allumage par ex) des carburants.

L'application du règlement CLP au cas des mélanges est entrée en vigueur au 1er juin 2015. Les règlements REACH et CLP donnent pour la définition de « mélange » suivante : un mélange ou une solution constitué de deux substances ou plus.

Il s'agit dans le cadre de ce rapport de mélanges chimiques de substances pures, à savoir que ces dernières ont été mises ensemble, mélangées ou mises en contact. Bien que les mélanges soient dans l'industrie constitués d'un grand nombre de composants, les mélanges traités dans ce rapport sont souvent des mélanges binaires (contenant deux constituants) ou limités à quelques constituants. Le cas des impuretés n'est pas pris en compte dans les mélanges considérés. De même, il s'agit ici de caractériser les propriétés des mélanges de substances et non la réactivité potentielle observée lors de la mise en contact de substances (incompatibilités chimiques).

2.2 Dangers physiques des mélanges

Les dangers physiques considérés dans ce rapport sont ceux correspondants aux 16 classes de dangers physiques du règlement CLP (par exemple, explosibilité, inflammabilité). Sont associées à ces dangers physiques, un grand nombre de propriétés physico-chimiques dangereuses nécessaires à la détermination du classement des substances et des mélanges selon le règlement. Il peut s'agir :

- de la chaleur et de la température de décomposition et de la sensibilité à différents stimuli pour l'explosibilité de substances,
- du point d'éclair pour l'inflammabilité de liquides,
- des limites d'inflammabilité (LIE/LSE) pour l'inflammabilité des gaz,
- de la chaleur de combustion des aérosols.

Dans le cas d'un mélange, au-delà du danger intrinsèque de chaque constituant, d'autres paramètres sont à considérer dans l'évaluation du danger du mélange comme :

- la concentration de chaque constituant dans le mélange,
- la miscibilité dans les mélanges liquides, l'hétérogénéité des mélanges solides,
- les interactions entre constituants dans le mélange, qui peuvent engendrer des effets synergiques ou antagonistes et donc augmenter ou diminuer la valeur de la propriété.

Il est donc important de caractériser explicitement les dangers des mélanges. En effet, certains mélanges peuvent même s'avérer plus dangereux que les produits purs. Cette problématique est illustrée ci-dessous pour le cas du point d'éclair dont la valeur (associée à celle de la température d'ébullition) permet selon le règlement CLP de classer un liquide dans la classe des liquides inflammables (avec 3 catégories différentes selon le niveau de danger). Plusieurs profils de point d'éclair sont ici présentés et illustrent l'intérêt de disposer de profils de cette propriété en fonction des concentrations des produits du mélange.

Dans le premier cas (dont le profil est représenté sur la Figure 2), il s'agit de la dilution d'un liquide inflammable (à savoir l'éthanol) par de l'eau. On note, en particulier, sur ce profil qu'il y a un changement de classe du liquide inflammable et que la variation peut être importante sur un domaine de concentration étroit. En effet, la solution aqueuse d'éthanol diluée passe de la classe 3 des combustibles (le point d'éclair FP est supérieur à 60°C, selon le règlement CLP) aux produits inflammables (23°C<FP<60°C) jusqu'à une fraction molaire en éthanol de 0,2 environ. Pour les fractions molaires en éthanol supérieures, la solution d'éthanol est dans la classe 1 des liquides extrêmement inflammables. Ce profil indique également qu'il faut un taux de dilution relativement important pour que le liquide devienne non inflammable.

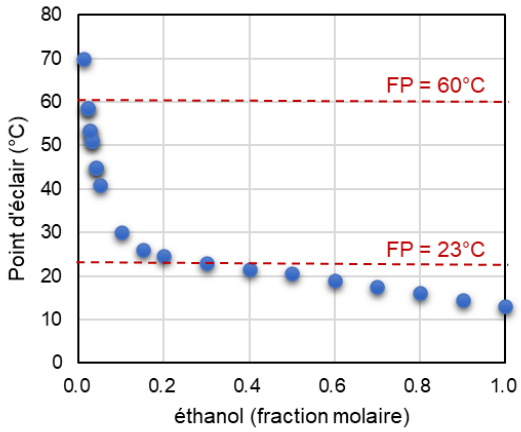

Figure 2 – Profil de point d'éclair du mélange éthanol/eau [10]

Dans le cas de mélanges entre deux produits inflammables, une approche couramment employée est de considérer le point d'éclair du produit le plus inflammable dans une approche supposée sécuritaire. Or, plusieurs types de profils sont rencontrés selon l'affinité entre les produits et leur miscibilité et cette hypothèse n'est pas toujours vérifiée (cf. exemples ci-dessous).

Parmi ces mélanges, on rencontre tout d'abord des profils idéaux (pour lesquels les constituants ne présentent pas d'interaction spécifique significative) qui présentent une évolution quasi-linéaire du point d'éclair vis-à-vis de la concentration comme présenté en Figure 3 pour le mélange octane/heptane.

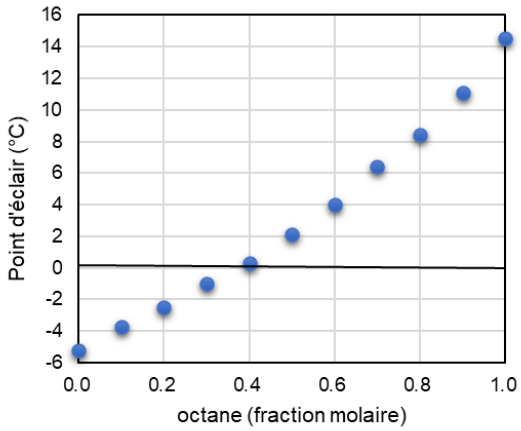

Figure 3 – Point d'éclair de mélanges octane/heptane [11]

Mais les différents composants du mélange peuvent également présenter des affinités particulières entre eux et engendrer alors des profils non idéaux. Le mélange p-picoline/phénol est, par exemple, un mélange présentant une importante déviation négative à l'idéalité amenant à un profil présentant même un point d'éclair maximal vers 0,3 en fraction molaire de p-picoline, comme montré en Figure 4. Dans de tels cas, considérer le point d'éclair du produit le plus inflammable reste une approche sécuritaire.

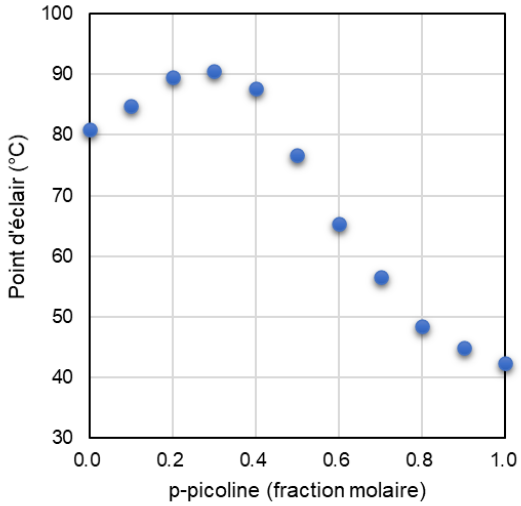

Figure 4 – Point d'éclair de mélanges p-picoline/phénol [12]

Par contre, comme dans le dernier cas (Figure 5), le mélange de deux produits purs inflammables peut s'avérer encore plus inflammable que les deux constituants du mélange pris à l'état pur. Il s'agit de mélanges présentant une déviation positive à l'idéalité pour lesquels un point d'éclair minimal inférieur à celui des produits purs peut être rencontré. Dans un tel cas, prendre en compte le point d'éclair le plus bas des deux produits mis en jeu ne représente pas une approche sécuritaire, bien au contraire.

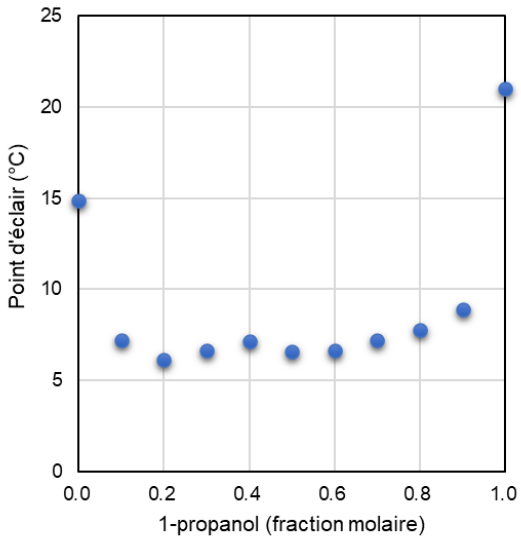

Figure 5 – Profil de point d'éclair du mélange 1-propanol/octane [13]

Les exemples, ci-dessus, montrent la diversité des situations rencontrées pour le point d'éclair. Il est donc nécessaire de caractériser explicitement le point d'éclair des mélanges pour chaque concentration et même parfois d'obtenir un profil de point d'éclair détaillé, par exemple dans le cadre de l'évaluation d'un procédé mettant en œuvre un tel mélange inflammable. Etant donné le nombre de mesures requises, le recours à des méthodes de calcul de type lois de mélanges voire QSPR (ne nécessitant même pas la connaissance des propriétés des produits purs) est indispensable pour compléter la caractérisation expérimentale. La détermination du point d'éclair de mélanges par prédiction, permettra d'obtenir au moins une estimation lorsque cette valeur n'est pas disponible et constitue un enjeu important de sécurité industrielle.

Elle sera également utile pour vérifier si une modification de composition peut augmenter par exemple l'inflammabilité d'un mélange, ce qui peut être le cas lors de l'évaporation d'un constituant au cours d'un procédé ou lors de stockages de longues durées. Ces méthodes présenteront enfin un intérêt tout particulier dans l'optimisation de campagnes expérimentales pour la définition de formulations dans une approche de type safer-by-design.

2.3 Lois de mélanges et méthodes d'estimation

Pour certains dangers physiques, des lois de mélanges existent pour estimer la propriété du mélange en fonction de la composition et de la connaissance de la valeur de cette propriété pour ses différents constituants.

Par exemple, la loi de Le Chatelier permet d'estimer la LIE_{mix} d'un mélange de produits combustibles à l'état gazeux à partir des LIE_i de ses n constituants :

$$LIE_{mix} = \frac{1}{\sum_{i=1}^{n} \frac{v_i}{LIE_i}} \quad (1)$$

où v_i est la fraction volumique des gaz purs dans l'air.

Elle s'applique avec une bonne approximation à de nombreux mélanges comme des mélanges d'hydrocarbures classiques avec néanmoins des erreurs plus importantes pour d'autres (acétone ou disulfure de carbone par exemple).

Différentes lois de mélanges ont également été proposées pour l'estimation des points d'éclairs de mélanges. La méthode de Gmehling et Rasmussen [14] est notamment citée dans les recommandations de l'ONU pour le Transport de Marchandises Dangereuses ou le règlement européen CLP. Elle repose sur la recherche de la température T pour laquelle la pression de vapeur du liquide P^{sat} égale la pression de vapeur à la limite inférieure d'inflammabilité. Cette approche a été largement adaptée au cours du temps. Le modèle en découlant le plus utilisé est la méthode de Liaw [15, 16], résumée dans l'Eq. 2.

$$\sum \frac{x_i \gamma_i P_i^{sat}(T)}{P_{i,FP}^{sat}(T)} = 1 \quad (2)$$

où x_i et γ_i sont respectivement la fraction molaire et le coefficient d'activité du composé i dans le mélange.

Ces lois de mélanges permettent parfaitement de prendre en compte les différents profils et comportements[1] mentionnés en 2.2 et représentent donc un outil puissant pour déterminer si un mélange de deux constituants peut s'avérer plus dangereux que ces derniers ou pour l'aide à la formulation de mélanges plus sûrs.

Une telle application a par exemple été réalisée pour la recherche du niveau de dilution à appliquer à une solution aqueuse à 40 % de diméthylamine (DMA) (au point d'éclair particulièrement bas de -18,5°C) pour la rendre moins inflammable [17]. Comme montré en Figure 6, la mise en place d'une loi de mélange ajustée sur des points expérimentaux fiables a permis de clarifier le profil d'évolution du point d'éclair en fonction de la concentration en DMA.

L'industriel à l'origine de cette étude a alors pu sélectionner un niveau de dilution réduisant les risques associés à ces stockages de DMA. Cette étude de cas démontre la possibilité d'aider à la sélection de formulations plus sûres pour des applications industrielles. On notera, de plus, que ces outils étaient particulièrement intéressants dans cette étude pour laquelle les mesures étaient complexes, notamment pour la solution à 40 % en DMA pour laquelle la maîtrise de l'évaporation du DMA est difficile lors de la préparation de l'essai.

[1] A ce jour, la méthode n'est pas validée pour les mélanges contenant par exemple des composants halogénés, sulfureux et/ou phosphoriques, ainsi que des acrylates réactifs.

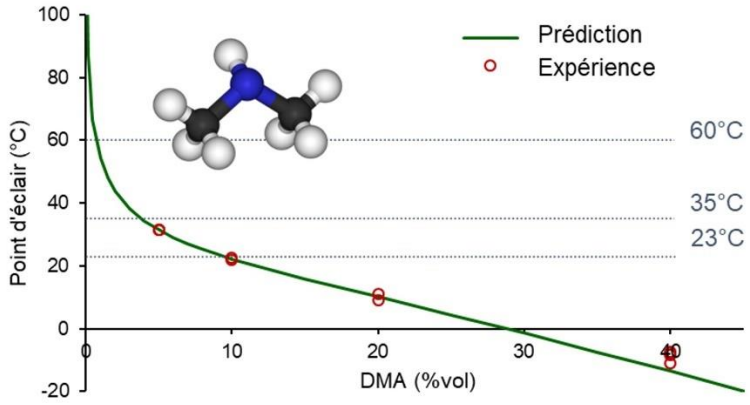

Figure 6 – Points d'éclair de solutions aqueuses de diméthylamine mesurés et prédits (selon une loi de méthode ajustée) [17]

Ces méthodes nécessitent la connaissance des propriétés des produits purs. En l'absence de cette valeur pour chaque constituant, il est possible d'utiliser des modèles QSPR existants pour la prédiction des propriétés de produits purs, comme cela a été démontré pour la prédiction du point d'éclair de mélanges binaires [18].

3 Recensement et analyse des modèles existants

3.1 Identification des modèles dans la littérature

L'état de l'art des modèles QSPR existants pour les dangers physiques des mélanges a été réalisé par l'analyse des articles disponibles dans la littérature scientifique accessible en juin 2022 via la plateforme *Web of Science*, sans limite de date de publication. Le *Web of Science* est une plateforme qui permet d'interroger plusieurs bases de données bibliographiques, dont le *Web of Science Core Collection* (qui a été en particulier exploré ici) et qui répertorie des articles de la littérature académique internationale publiés dans environ 20 000 revues à comité de lecture parmi les plus influentes (dont 12 000 avec un facteur d'impact), ainsi que dans des livres et actes de congrès.

Une requête avec des mots clés trop simples et génériques fait ressortir de très nombreuses publications ne concernant pas des modèles prédictifs de type QSPR ou des dangers physiques. Par exemple, une recherche avec les mots-clés *« Prediction »* et *« Mixture »* fait ressortir plus de 39000 références dont aucune n'est pertinente parmi les 50 premières (classées par ordre de pertinence par la plateforme).

Aussi, deux requêtes affinées ont été combinées afin d'identifier autant que possible toutes les publications relatives à des modèles QSPR dédiés à des dangers physiques tout en éliminant les articles non pertinents. La première[2] recherche toutes les références pour lesquelles les titres, résumés ou mots-clés contiennent à la fois le mot *« mixture »*, un terme relatif aux méthodes prédictives (*« QSPR »* ou *« structure-property »*) et un terme relatif aux propriétés visées (*« flammability »* ou *« explosibility »* ou *« flash point »* ou *« self-ignition »* ou *« auto-ignition »* ou *« combustion »* ou *« explosiv* »* ou *« detonation »* ou *« deflagration »* ou *« decomposition heat »* ou *« heat of decomposition »* ou *« decomposition temperature »* ou *« impact sensitivity »* ou *« friction sensitivity »* ou *« combustion heat »* ou *« heat of combustion »*). La seconde[3] est focalisée seulement sur le titre des références en ajoutant *« predict* »* parmi les termes relatifs aux méthodes prédictives.

La combinaison de ces deux requêtes fait ressortir 139 références parmi lesquelles on retrouve encore des publications ne concernant pas des modèles QSPR ou non dédiées aux propriétés ciblées (comme l'indice d'octane de carburants ou le coefficient de diffusion dans des électrolytes). Des articles ne faisant que mentionner la question de la prédiction des propriétés de mélanges sans proposer de modèles prédictifs ou ne faisant que citer des modèles existants sont également éliminés à cette étape.

Finalement, 23 publications[4] correspondant au développement de modèles QSPR pour des dangers physiques ont été retenues. Elles sont listées en Annexe 1. Ces articles traduisent un champ de recherche récent et actif puisqu'ils ont tous été publiés entre 2013 et 2022.

---

[2] TS=(mixture AND (QSPR OR "structure-property") AND (flammability OR explosibility OR "flash point" OR "self-ignition" OR "auto-ignition" OR combustion OR explosiv* OR detonation OR deflagration OR "decomposition heat" OR "heat of decomposition" OR "decomposition temperature" OR "impact sensitivity" OR "friction sensitivity" OR "combustion heat" OR "heat of combustion"))

[3] TI=(mixture AND (QSPR OR "structure-property" OR predict*) AND (flammability OR explosibility OR "flash point" OR "self-ignition" OR "auto-ignition" OR combustion OR explosiv* OR detonation OR deflagration OR "decomposition heat" OR "heat of decomposition" OR "decomposition temperature" OR "impact sensitivity" OR "friction sensitivity" OR "combustion heat" OR "heat of combustion"))

[4] Dans certaines publications, plusieurs modèles sont proposés pour une même propriété. Dans de tels cas, tous les modèles proposés ne sont pas considérés indépendamment mais un seul modèle est considéré dans les analyses statistiques proposées dans la suite de ce rapport, que les auteurs aient mentionné un modèle à privilégier ou non.

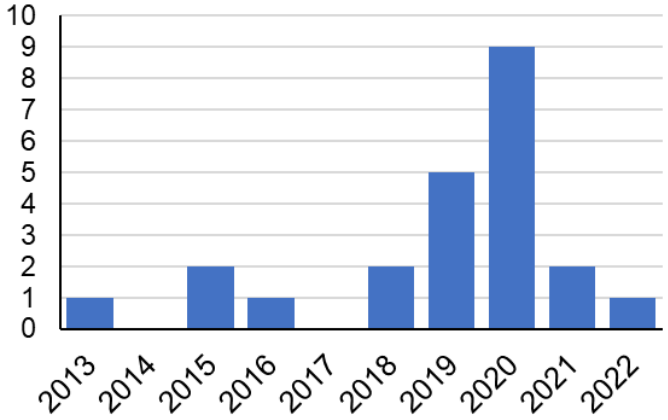

Figure 7 - Nombre de publications recensées

3.2 Propriétés et familles chimiques visées

Pour les mélanges, encore plus que pour les produits purs, la disponibilité de données expérimentales de qualité et en nombre suffisant est la limite principale au développement de modèles QSPR. Il n'est donc pas étonnant que les modèles recensés se concentrent sur certaines propriétés et sur les mêmes familles de substances.

En effet, les données expérimentales de référence utilisées pour le développement et la validation des modèles sont d'une grande importance puisque leur fiabilité influencera les performances du modèle par effet de propagation de leur incertitude. De plus, les données inclues dans le jeu d'entraînement définissent le domaine d'applicabilité du modèle. Ainsi, pour l'obtention de modèles QSPR performants, il est nécessaire de disposer d'une base de données expérimentale la plus fiable possible en nombre suffisamment important, d'une part, pour présenter une variabilité suffisante de la propriété et des structures et, d'autre part, pour permettre un partage en deux jeux de données, l'un d'entraînement et l'autre de validation.

Dans toutes les publications recensées, les données utilisées pour le développement et la validation des modèles sont issues de la compilation des résultats expérimentaux collectés dans la littérature scientifique. En effet, les quelques bases de données contenant des informations sur les dangers physiques de substances[5] concernent (principalement) des produits purs.

Ainsi, tous les modèles recensés, concernent l'inflammabilité de liquides organiques, à l'exception du modèle de He et al. (2021) [19] qui s'intéresse à la température de décomposition de liquides ioniques[6]. Les propriétés inflammables visées sont les données les plus couramment utilisées pour évaluer l'inflammabilité des liquides et des gaz, à savoir le point d'éclair (PE), les limites inférieure et supérieure d'explosivité (LIE/LSE) et la température d'auto-inflammation (TAI).

[5] On citera par exemple les bases eChemPortal de l'OCDE (https://www.echemportal.org/), la base, CarAtex de l'INRS (https://www.inrs.fr/publications/bdd/caratex.html) ou la base allemande ChemSafe (https://www.chemsafe.ptb.de/).

[6] Il est à noter que la propriété prédite par ce modèle n'est pas la température de début de décomposition (utilisée par exemple comme critère de pré-sélection pour le classement des peroxydes organiques, des matières auto-réactives ou des matières explosibles) mais la température à 5% de décomposition ($T_{d,5\%}$).

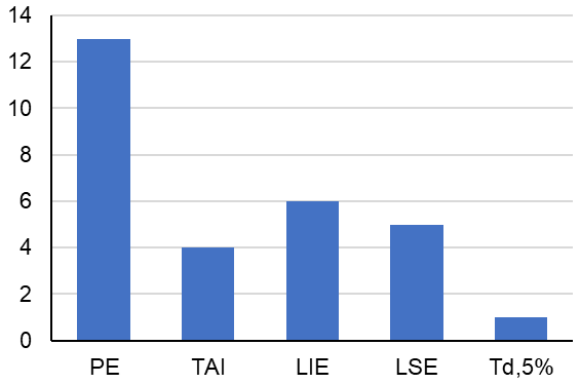

Figure 8 – Propriétés étudiées dans les publications recensées

Les mélanges liquides inflammables visés sont des mélanges organiques relativement simples, binaires et/ou ternaires. Parmi les modèles développés pour le point d'éclair de mélanges binaires, deux études ont testé leur potentiel sur des mélanges ternaires (Fayet (2019) [8] et Torabian (2019) [20]) et ont montré des performances similaires à celles observées pour les mélanges binaires.

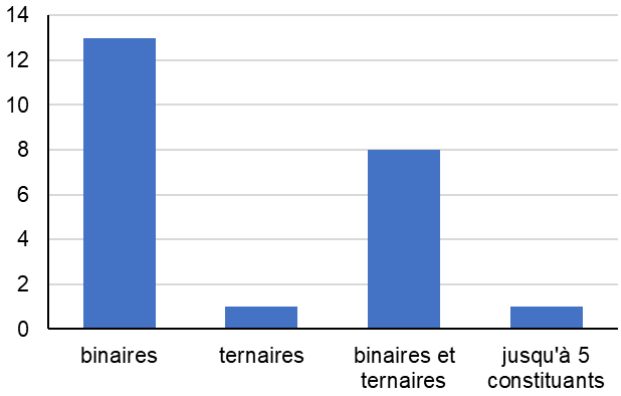

Figure 9 – Complexité des mélanges visés dans les publications recensées

D'une manière générale, les bases de données utilisées restent relativement modestes en termes de diversité des mélanges. Elles sont en général constituées autour de quelques composés organiques classiques utilisés par exemple dans la composition des carburants pour lesquels une évaluation des dangers d'inflammabilité est nécessaire dans un contexte industriel, d'où l'existence de base de données expérimentales.

Par exemple, la plus grande base de données rencontrée est celle utilisée par Jiao et al. (2020) [21] pour le développement de son modèle pour le point d'éclair. Elle contient 1 458 données mais pour des mélanges binaires et ternaires constitués à partir de (seulement) 47 produits purs différents avec une inhomogénéité de représentation de ces derniers dans les mélanges avec 3 composés très majoritaires, le méthanol, l'octane et l'éthanol, que l'on retrouve dans respectivement 36 %, 27 % et 26 % des données (comme montré en Figure 10).

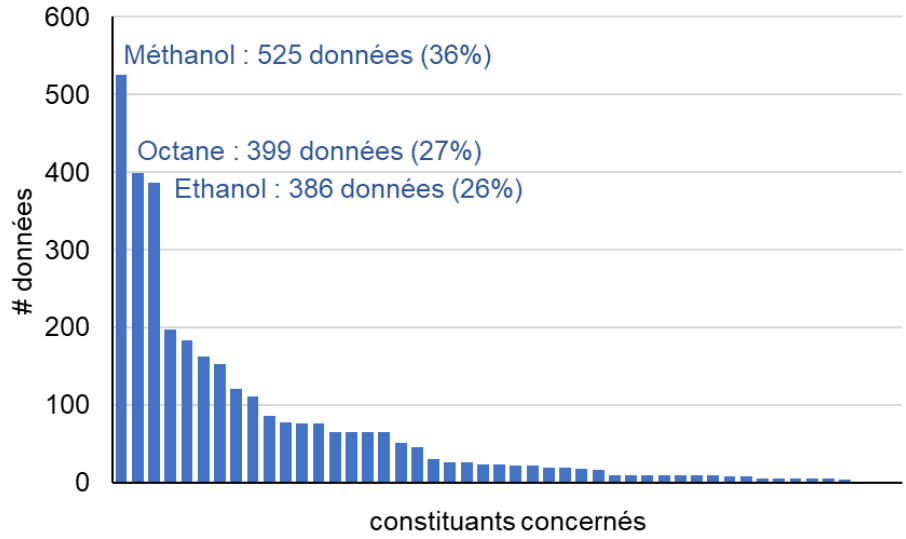

Figure 10 – Principaux constituants représentés dans les données de point d'éclair utilisées par Jiao et al. (2020) [21]

En particulier, si le nombre de données sur les mélanges ternaires peut paraître parfois important, elles concernent en général peu de mélanges différents. Par exemple, le modèle de Toropova (2020) [22] spécifiquement développé pour des mélanges ternaires est basé sur 808 données de points d'éclairs mais elles ne concernent que 8 mélanges différents.

On notera le cas particulier du modèle récemment développé par Aljaman et al. (2022) [23] sur une base de données comptant à la fois des points d'éclairs de produits purs (474 données) et des données concernant des mélanges (279 binaires, 26 ternaires, 6 quaternaires et 8 quinaires). Cette approche pourrait être un moyen de contourner le problème de manque de diversité chimique des constituants, même si l'absence de validation externe ne permet pas d'évaluer sa pertinence et sa capacité à prendre en compte les effets de concentration et d'interaction entre produits pour ceux qui sont uniquement représentés en tant que produits purs.

⇨ Les modèles recensés concernent peu de dangers physiques différents (inflammabilité principalement) et reposent sur des données parfois nombreuses mais en général peu variées en termes de diversité chimique, puisque souvent limitées ou fortement concentrées autour de mélanges formés à partir d'un nombre réduit de produits purs différents.

3.3 Descripteurs moléculaires et de mélange

Dans un modèle QSAR/QSPR classique, la structure moléculaire est représentée par une série de descripteurs moléculaires. De nombreux descripteurs ont été proposés et sont disponibles avec différents niveaux de complexités. Par exemple, on distingue des descripteurs en fonction de la complexité de la structure moléculaire considérée : les descripteurs 1D ou constitutionnels reposent sur la connaissance de la seule composition de la molécule (ex : nombres d'atomes ou de groupes spécifiques présents), les descripteurs 2D considèrent la structure topologique c'est-à-dire la manière dont les atomes sont connectés entre eux (distinguant ainsi des isomères par exemple), enfin les descripteurs 3D prennent en compte l'arrangement spatial en 3 dimensions de la molécule (descripteurs géométriques) voire même des informations issues de calculs de chimie quantique sur les propriétés électroniques et la réactivité des molécules (descripteurs quantiques). Ces descripteurs peuvent enfin considérer les molécules dans leur entièreté (descripteurs intégraux) ou seulement sur certaines parties de ces dernières (descripteurs de fragments).

Dans le cas d'un modèle de mélange, au-delà de la structure chimique de chaque constituant, d'autres paramètres peuvent influencer la propriété (comme montré au paragraphe 2.2) et en particulier :

- la concentration de chaque constituant dans le mélange ;
- les interactions entre constituants dans le mélange, qui peuvent engendrer des effets synergiques ou antagonistes.

Comme montré en Annexe 1, les modèles recensés dans cet état de l'art utilisent majoritairement des descripteurs de mélange D_m calculés en combinant des descripteurs moléculaires *d* pour chaque constituant via une formule de mélange en fonction de la concentration de chacun d'eux. Le plus couramment, il s'agit de considérer une additivité des effets via une simple pondération linéaire vis-à-vis de la fraction molaire *x* (ou plus rarement la fraction volumique), selon l'équation suivante :

$$D_m = \sum x_i d_i \quad (3)$$

D'autres formules de mélanges ont également été testées pour prendre en compte la non-linéarité des effets de concentration rencontrés dans les mélanges. Certaines ont pour but de prendre directement en compte une dépendance non linéaire de la propriété par rapport à la fraction molaire alors que d'autres traduisent une déviation par rapport au comportement linéaire vis-à-vis de la concentration. 12 formules ont par exemple été testées pour la recherche de modèles prédictifs du point d'éclair de mélanges binaires (en Table 1) [9].

Formules basées directement sur la fraction molaire de chaque constituant	$D_m = x_1 d_1 + x_2 d_2$
	$D_m = \lvert x_1 d_1 - x_2 d_2 \rvert$
	$D_m = x_1^2 d_1 + x_2^2 d_2$
	$D_m = \sqrt{x_1} d_1 + \sqrt{x_2} d_2$
	$D_m = (x_1 d_1 + x_2 d_2)^2$
	$D_m = \sqrt{(x_1 d_1)^2 + (x_2 d_2)^2}$
Formules basées sur la différence de fraction molaires	$D_m = (1 - \Delta x)\Delta d$
	$D_m = (1 - \Delta x^2)\Delta d$
	$D_m = (1 - \Delta x)^2 \Delta d$
Autres formules	$D_m = (d_1 + d_2)/2$
	$D_m = (d_1 - d_2)^2$
	$D_m = \lvert d_1 - d_2 \rvert$

Table 1 - Différentes formules de descripteurs de mélanges testées pour le développement d'un modèle QSPR pour la prédiction du point d'éclair de mélanges binaires [9]

Une partie d'entre elles ont finalement été combinées dans le modèle développé par l'Ineris en 2019 pour le point d'éclair *FP* de mélanges liquides organiques binaires [8].

$$FP\ (°C) = 20{,}3 + 28{,}6 \sum x_i T_i^E + 24{,}6 \sum \sqrt{x_i}\ {}_1\chi_i + 59{,}6 \sum \sqrt{x_i} V_{YXZ,i} - 315{,}0 \sum \sqrt{x_i} Q_{H,max,i} - 107{,}6 \sum \sqrt{x_i} V_{H,min,i} + 2{,}0 \left(\sum \sqrt{x_i}\mu\right)^2 \quad (4)$$

Dans ces travaux, des descripteurs moléculaires intégraux variés (constitutionnels, topologiques, géométriques et quantiques) ont été employés. Ces descripteurs classiques caractérisent les molécules isolées, même si certains peuvent traduire leur potentiel à créer des interactions intermoléculaires. Par exemple, dans ce modèle, $Q_{H,max}$ et $V_{H,min}$ désignent respectivement la charge maximale et la valence minimale pour un atome d'hydrogène et traduisent le potentiel des produits mis en jeu dans le mélange à former des liaisons hydrogène avec d'autres constituants.

Cette approche a également été appliquée sur des contributions de groupes ou des descripteurs de fragments. Par exemple, dans leur approche par contributions de groupe pour la prédiction de la TAI de mélanges binaires, Ye et al. [24] ont introduit, une formulation prenant en compte le nombre d'occurrence du groupe dans chaque molécule et leurs fractions volumiques respectives.

$$TAI = a + bX + cX^2 + dX^3 + eX^4 \quad (5)$$

$$X = v_1 \sum_i n_{1i} f_i + v_2 \sum_j n_{2j} f_j \quad (6)$$

où n_1 et n_2 désignent les nombres d'occurrence des groupes dans chacune des deux substances du mélange binaire, f est la contribution associée et v_i est la fraction volumique de chaque constituant.

De manière similaire, Aljaman et al. [23] ont considéré la concentration massique associée à différents groupements présents dans la molécule (CH_3, CH_2, $CH{=}CH_2$, CHO, COO, etc.) pour la prédiction du point d'éclair de carburants oxygénés.

La plupart des modèles recensés basés sur des fragments utilisent l'approche SiRMS[7]. Cette dernière propose la décomposition des molécules en fragments de 2 à 6 atomes nommés simplex. Une originalité de l'application de cette approche au cas des mélanges (binaires) est l'introduction de simplex dit non-liés dans lesquels certains atomes constituant le simplex ne sont pas liés entre eux et appartiennent donc aux deux molécules du mélange binaire. Ils caractérisent donc des interactions intermoléculaires spécifiques dans le mélange. Deux formules de mélanges différentes sont ensuite appliquées : l'Eq. 7 pour les simplex liés d_i (intramoléculaires) et l'Eq. 8 pour les simplex non liés d_{1+2} (intermoléculaires).

$$D_m = x_1 d_1 + x_1 d_1 \quad (7)$$

$$D_m = 2\, x_1 d_{1+2} \quad (8)$$

Cette approche a été utilisée par Shen et al. (2019) [25] pour la température d'auto-inflammation puis par Yao et al. (2020) [26] et Cao et al. (2020) [27] pour le point d'éclair, à chaque fois pour des mélanges de liquides organiques binaires.

De leur côté, He et al. [19] ont utilisé des descripteurs ISIDA pour prédire la température de décomposition de mélanges de liquides ioniques. Concernant la spécificité des liquides ioniques (constitués d'un anion et d'un cation), certains fragments sont proposés pour prendre en compte l'interaction anion-cation au sein d'un liquide ionique, mais pas entre liquides ioniques.

Enfin, contrairement aux travaux précédemment cités, Toropova et al. [22, 28] n'appliquent pas de formule de mélange dans leur approche basée sur des quasi-SMILES. Ces derniers sont définis comme des extensions du code SMILES [29] des molécules en ajoutant des symboles complémentaires pour des informations particulières aux systèmes étudiés. Dans le cas de mélanges, le quasi-SMILES représente l'intégralité du mélange et prend la forme suivante :

[SMILES#1][%X1][SMILES#2][%X2] (pour un mélange binaire) (9)

[SMILES#1][%X1][SMILES#2][%X2][SMILES#3][%X3] (pour un mélange ternaire) (10)

Où *[SMILES#i]* est le code SMILES du constituant i et *[%Xi]* est une chaîne de caractères représentant sa fraction molaire dans le mélange[8].

Dans cette approche, la concentration est considérée dans le modèle final de la même manière que les fragments identifiés dans la structure de chaque constituant (dans leur code SMILES), donc soit seul soit en combinaison avec un fragment moléculaire particulier.

⇨ Si différents types de descripteurs moléculaires sont rencontrés (comme dans le cas des produits purs), la plupart des modèles reposent sur le développement de descripteurs de mélanges à partir de formules combinant les descripteurs moléculaires des différents constituants en fonction de leurs fractions molaires respectives dans le mélange. C'est ainsi que sont en général prises en compte les spécificités de mélanges. Si seuls les modèles basés sur l'approche SiRMS introduisent des descripteurs moléculaires représentant directement l'interaction entre substances (via des simplex non-liés), certains descripteurs moléculaires plus classiques peuvent déjà traduire des potentiels d'interactions intermoléculaires.

[7] *Simplex Representation of Molecular Structure*

[8] Dans cette approche, la fraction molaire n'est pas considérée comme une valeur continue mais une chaine de caractère est attribuée pour une plage de concentration donnée.

3.4 Algorithmes utilisés

Différents types d'algorithmes, plus ou moins complexes, peuvent être employés dans les modèles QSAR/QSPR. L'état de l'art réalisé pour le cas des mélanges ne met pas en évidence de particularité dans les algorithmes utilisés par rapport aux modèles dédiés aux produits purs.

Comme montré en Figure 11, les modèles recensés sont la plupart du temps basés sur les régressions multilinéaires (MLR, pour *Multi-Linar Regression*). On retrouve également quelques modèles basés sur des approches de *machine learning* plus complexes comme des réseaux de neurones artificiels (ANN, pour *Artificial Neural Network*), des machines à vecteur support (SVM, pour *Support Vector Machine*) ou autres méthodes non-linéaires (régressions non-linéaires, *k-Nearest-Neighbors* (k-NN), *Random Forest* (RF), *Bootstrap Tree* (BT)).

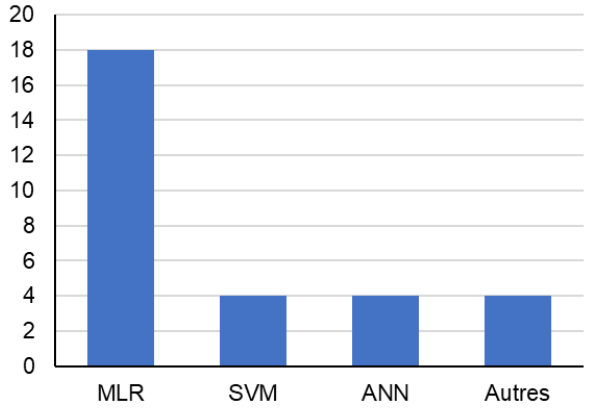

Figure 11 – Algorithmes utilisés dans les publications recensées

Certaines études comparent différentes approches et semblent montrer un certain potentiel des méthodes non-linéaires à prendre en compte la complexité des phénomènes associés aux propriétés de mélanges. Par exemple, Jiao et al. (2020) [30] ont employé 5 méthodes différentes (MLR, k-NN, RF, BT and SVM) pour la limite inférieure d'inflammabilité de mélanges binaires d'hydrocarbures et mis en évidence les performances les plus importantes pour le modèle basé sur l'approche *Random Forest*. Néanmoins, en considérant la taille limitée de la base de données utilisée (à savoir 54 données), les performances des différentes approches restent proches entre elles avec un R^2 de 0,98 observé sur le jeu de validation externe pour l'approche RF, contre par exemple 0,92 pour l'approche multilinéaire (comme montré en Table 2).

Méthode	R^2	
	Entrainement (n=43)	Validation (n=11)
MLR	0,9486	0,9186
k-NN	0,8735	0,8583
RF	0,9973	0,9831
BT	0,9926	0,9523
SVM	0,9900	0,9704

Table 2 – Performances des modèles développés par Jiao et al. (2020) [30] pour la LIE de mélanges

⇨ Globalement, les algorithmes utilisés dans les modèles recensés sont les mêmes que ceux rencontrés dans le cas des modèles QSPR pour des produits purs et aucune adaptation particulière n'a été rencontrée pour la prise en compte de spécificité des mélanges au niveau de ceux-ci. Si la régression multilinéaire est l'approche la plus couramment employée, les approches non linéaires (réseaux de neurones, *random forest*…) pourraient permettre de prendre en compte la complexité des mélanges.

3.5 Validation des modèles

Pour valider la performance d'un modèle QSAR/QSPR, différentes méthodes de validation sont employées. Si l'évaluation de la qualité des prédictions au sein du jeu d'entrainement permet d'évaluer la qualité d'ajustement du modèle, d'autres méthodes de validation interne et externe sont recommandées. Ainsi, la validation croisée (ou *cross validation*) ou le *Bootstrapping* permettent d'évaluer la robustesse du modèle alors que la méthode de *Y-Randomization* permet de vérifier que le modèle n'a pas été obtenu par chance. Pour évaluer la capacité prédictive du modèle, une validation externe est nécessaire sur des données différentes de celles utilisées pour son entrainement (jeu de validation).

Les modèles recensés pour la prédiction des dangers physiques de mélanges étant globalement récents, ils ont tous fait l'objet de validations au-delà de la simple analyse de corrélations entre les données expérimentales et prédites de leur jeu d'entrainement. Une seule étude n'a pas inclus de validation externe, les modèles de Jiao et al. (2016) [31] n'ayant été validés que par cross-validation (CV). Comme montré en Figure 12, 78 % des études ont utilisé cette méthode de validation interne alors qu'on rencontre plus rarement l'usage de validations internes par *Bootstrapping* ou *Y-randomisation*.

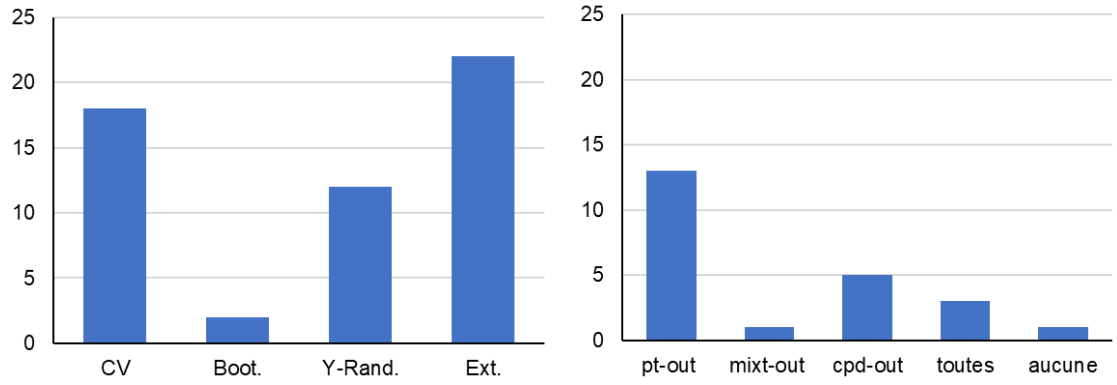

Figure 12 – Méthodes de validation (à gauche) et types de validation externe (à droite)[9] employés dans les publications recensées

Si aucune adaptation spécifique n'a été proposée jusqu'à présent dans les méthodes de validation internes, Muratov et al. (2012) [5] distinguent trois types de validation externe selon la manière dont le jeu de validation se distingue du jeu d'entrainement :

- *Points-out* : La partition entre jeu d'entrainement et de validation considère chaque donnée expérimentale individuellement sans considérer le fait qu'elles puissent être associées aux mêmes mélanges et constituants. Un mélange peut donc être présent à la fois dans le jeu d'entrainement et dans le jeu de validation (avec différentes proportions entre ses constituants). Cette méthode évalue surtout la capacité des modèles à prédire les propriétés de mélanges existants dans de nouvelles proportions ;
- *Mixture-out* : Dans ce type de partition, tous les points de données correspondant à des mélanges composés des mêmes constituants sont regroupés dans le même jeu (entrainement ou validation). L'erreur attendue pour une telle validation externe est plus élevée que pour la stratégie en *Points-out* mais elle évaluera la capacité du modèle à prédire les propriétés de nouveaux mélanges au-delà du seul effet de concentration ;
- *Compounds-out* : Les données du jeu de validation ne concernent que des mélanges pour lesquels au moins un des constituants est absent du jeu d'entraînement. Il s'agit de la méthode la plus rigoureuse de validation externe dans la modélisation QSAR/QSPR des mélanges. Si l'erreur de prédiction attendue pour cette stratégie est la plus grande, les modèles validés avec une telle partition auront démontré leur capacité à prédire la propriété étudiée pour des mélanges constitués par un ou plusieurs nouveaux produits.

[9] CV : Cross-Validation ; Boot. : Bootstrapping ; Y-Rand. : Y-randomization ; Ext. : Validation externe ; pt-out : points-out ; mixt-out : mixtures-out ; cpd-out : compounds-out

La plupart des modèles ont réalisé une validation externe sur la base d'une partition Points-out (70 %). Les modèles validés sur des partitions Compounds-out sont moins nombreux (35 %). Une telle validation externe est la plus simple à mettre en œuvre mais elle n'est pas, a priori, la plus rigoureuse manière d'évaluer le pouvoir prédictif du modèle puisque le jeu de validation n'est pas strictement indépendant du jeu d'entraînement puisque des mélanges constitués des mêmes produits et ne différant que d'une faible valeur de concentration par rapport au jeu d'entrainement peuvent se retrouver dans le jeu de validation.

Cela dit, comme indiqué en section 3.2, les bases de données disponibles présentent une diversité chimique plutôt limitée. Ce constat impacte la qualité d'ajustement du modèle mais aussi sa validation externe, surtout dans le cadre de partitions Mixtures-out ou Compounds-out puisque le jeu de validation est donc d'autant plus limité en termes de diversité structurale. Si un jeu de validation en Points-out n'est pas rigoureusement indépendant (chimiquement) du jeu d'entrainement, il présente dans de tels cas l'avantage d'offrir une validation externe sur un échantillon plus varié qu'avec des partitions en Mixtures ou Compounds-out. Aucune de ces différentes partitions ne semble donc idéale avec de telles bases de données. On notera d'ailleurs que Wang et al. (2019) [32] ont appliqué les trois types de partitions et recommandent finalement de retenir le modèle en Points-out du fait de meilleures performances observées en validation externe (avec un R^2_{ext} de 0,965 contre respectivement 0,879 et 0,923 pour les partitions Mixtures et Compounds-out).

Si aucune valeur limite absolue n'est imposée pour la validation d'un modèle QSAR/QSPR[10], il est néanmoins en général considéré qu'un modèle valide doit à minima présenter une corrélation supérieure à 0,6 en validation croisée [33] et un pouvoir prédictif supérieur à 0,7 en validation externe [34]. Tous les travaux recensés mènent au développement de modèles répondant à ces critères, même si des bases de données plus larges et diversifiées seraient nécessaires pour disposer d'évaluations plus robustes du pouvoir prédictif des modèles.

⇨ Les modèles recensés présentent des performances répondant aux critères classiques de performances attendus pour des modèles prédictifs valides. Néanmoins, les limites des bases de données expérimentales disponibles en termes de diversité chimique des mélanges concernés impactent la fiabilité des évaluations de pouvoir prédictif réalisées avec des validations externes utilisant des jeux de validation limités en termes de diversité de produits impliqués, qui sont en outre le plus souvent déjà présents dans le jeu d'entrainement (partitions de type Points-out).

3.6 Domaine d'applicabilité

Les modèles QSAR/QSPR sont des modèles empiriques dont la validité est limitée à un champ d'application délimité par les données pour lesquelles il a été entrainé. La définition de son domaine d'applicabilité est donc importante afin de déterminer si une prédiction obtenue à partir de ce modèle est fiable.

65 % des études recensées incluent une évaluation du domaine d'applicabilité des modèles, en employant, dans la grande majorité des cas (13/15), l'analyse par *Williams plot*[11], sans adaptation particulière par rapport au cas des produits purs (si ce n'est qu'ils sont basés sur les valeurs de descripteurs de mélange). Seuls les travaux de Toropova et al. [22, 28] utilisent une autre approche basée sur l'analyse de la distribution des fragments identifiés dans les quasi-SMILES.

[10] En effet, la qualité d'une prédiction dépend à la fois de la propriété prédite (et de l'incertitude associée à sa mesure expérimentale), de la substance visée et de l'usage de la prédiction.

[11] Un *Williams plot* constitue la représentation graphique de l'approche *leverage* [35] qui consiste à calculer « l'effet levier » associé à chaque molécule par rapport à la distribution globale du jeu d'entraînement.

Dans ces approches, toutes les prédictions sont prises de manière indépendante et il pourrait être intéressant d'engager une réflexion sur la possibilité de prendre en compte le caractère spécifique des mélanges en considérant non seulement la diversité chimique des molécules mises en jeu mais également les effets de concentrations et d'interactions intermoléculaires.

⇨ La plupart des modèles ont fait l'objet d'une évaluation du domaine d'applicabilité selon la même approche que celle utilisée pour les modèles dédiés aux produits purs. Une réflexion pourrait être à mener sur l'adaptation de ces approches au cas des mélanges.

4 Synthèse et perspectives

Si l'approche QSPR est initialement dédiée aux produits purs, l'utilisation de modèles QSPR pour la prédiction des dangers physiques de mélanges est en plein essor. En effet, des modèles QSPR ont été recensés dans la littérature scientifique pour la prédiction des dangers physiques de mélanges à travers 23 publications récentes (de 2013 à nos jours). Ces modèles se concentrent sur quelques propriétés (d'inflammabilité principalement tout comme les modèles QSPR existants pour les produits purs) et en majorité sur des mélanges binaires pour lesquels des données existaient en nombre relativement important (jusqu'à environ 1500 données). Mais ces données sont en général peu variées en termes de diversité chimique, puisqu'elles sont souvent limitées à quelques produits purs différents.

Pour prendre en compte les spécificités des mélanges, l'approche la plus communément employée repose sur le calcul de descripteurs de mélanges à partir de descripteurs moléculaires des différents constituants et sur leurs concentrations respectives dans le mélange. Si quelques modèles seulement introduisent des descripteurs de fragments non-liés (caractérisant explicitement des interactions intermoléculaires), certains descripteurs moléculaires plus classiques peuvent déjà traduire le potentiel des constituants individuels à présenter des interactions avec d'autres constituants.

Au-delà de cela, ces modèles sont développés à partir des mêmes algorithmes que les modèles pour les produits purs, en général par des régressions multilinéaires même si des approches non linéaires plus complexes semblent intéressantes pour prendre en compte la complexité des mélanges.

Du fait du caractère récent de ces travaux (tous publiés après 2013), tous ces modèles ont été validés et présentent des performances en accord avec les standards classiques attendus pour des modèles valides (incluant validation externe et définition d'un domaine d'applicabilité) et similaires à celles obtenues pour les produits purs. La robustesse de leurs validations reste néanmoins impactée par le manque de diversité chimique des bases de données expérimentales disponibles, puisque les jeux de validation sont eux-mêmes limités en termes de diversité de produits impliqués. De plus, ces jeux de validation sont en général basés sur des partitions de type points-out (qui induisent que certains mélanges du jeu de validation sont composés de constituants déjà présents dans le jeu d'entrainement).

Les travaux existants permettent d'ores et déjà de compléter l'approche expérimentale. Le développement de nouveaux modèles QSPR plus fiables reposera sur la disponibilité de bases de données plus importantes et plus diversifiées en termes de substances chimiques impliquées. De tels jeux de données permettront notamment de renforcer les évaluations du pouvoir prédictif des modèles en utilisant des jeux de validation à la fois indépendants des jeux d'entrainement (en favorisant une approche compounds-out) tout en disposant d'une diversité de substances représentatives chimiquement des mélanges visés par le modèle.

De nouvelles données devraient également permettre d'étendre le champ d'application des modèles à d'autres dangers physiques, au-delà de l'inflammabilité des liquides et des gaz qui concernent la quasi-totalité des modèles existants.

En termes de validation des modèles, au-delà des premières recommandations existantes en particulier pour l'évaluation du pouvoir prédictif des modèles, des réflexions mériteraient d'être engagées sur la détermination du domaine d'applicabilité des modèles qui est pour l'heure réalisée via les outils et méthodes dédiés aux produits purs qui ne prennent pas en compte les spécificités des mélanges.

De plus, les recommandations de validation des prédictions[12] (qui dépendent non seulement du modèle mais aussi de la substance et de la manière dont la prédiction va être utilisée) pourraient également être à analyser afin de vérifier si des préconisations spécifiques seraient à proposer pour le cas des mélanges.

[12] Notamment au travers des documents QPRF (*QSAR Prediction Reporting Format*) proposés par l'OCDE.

5 Références

1. OECD principles for the Validation, for Regulatory Purpose, of (Q)SAR Models, Organisation for Economic Co-operation and Development (OECD), 2004. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf.
2. G. Fayet, P. Rotureau, Note de synthèse sur les démarches de diffusion de modèles QSPR développés par l'Ineris, Ineris-201438-2304954-v1.0, Ineris, 2020.
3. G. Fayet, P. Rotureau, L. Petit, Etat des lieux des méthodes prédictives pour les dangers physiques des rubriques relatives à des substances de la nomenclature ICPE, Ineris-203823-2733314-v1.0, Ineris, 2022.
4. C. Nieto-Draghi et al., *A General Guidebook for the Theoretical Prediction of Physico-Chemical Properties of Chemicals for Regulatory Purposes.* Chemical Reviews, *115*, **2015**, 13093-13164.
5. E.N. Muratov, E.V. Varlamova, A.G. Artemenko, P.G. Polishchuk, V.E. Kuz'min, *Existing and Developing Approaches for QSAR Analysis of Mixtures.* Molecular Informatics, *31*, **2012**, 202-221.
6. E. Mombelli, État de l'art sur la modélisation QSAR de la toxicité des mélanges chimiques, Ineris-204142-2727622-v1.0, Ineris, 2022.
7. S.J. Belfield, J.W. Firman, S.J. Enoch, J.C. Madden, K. Erik Tollefsen, M.T.D. Cronin, *A review of quantitative structure-activity relationship modelling approaches to predict the toxicity of mixtures.* Computational Toxicology, *25*, **2023**, 100251.
8. G. Fayet, P. Rotureau, *New QSPR Models to Predict the Flammability of Binary Liquid Mixtures.* Molecular Informatics, *38*, **2019**, 1800122.
9. T. Gaudin, P. Rotureau, G. Fayet, *Mixture Descriptors toward the Development of Quantitative Structure-Property Relationship Models for the Flash Points of Organic Mixtures.* Industrial & Engineering Chemistry Research, *54*, **2015**, 6596-6604.
10. H.-J. Liaw, V. Gerbaud, H.-T. Wu, *Flash-Point Measurements and Modeling for Ternary Partially Miscible Aqueous−Organic Mixtures.* Journal of Chemical & Engineering Data, *55*, **2010**, 3451-3461.
11. H.-J. Liaw, V. Gerbaud, Y.-H. Li, *Prediction of miscible mixtures flash-point from UNIFAC group contribution methods.* Fluid Phase Equilibria, *300*, **2011**, 70-82.
12. H.-J. Liaw, S.-C. Lin, *Binary mixtures exhibiting maximum flash-point behavior.* Journal of Hazardous Materials, *140*, **2007**, 155-164.
13. M. Noorollahy, A.Z. Moghadam, A.A. Ghasrodashti, *Calculation of mixture equilibrium binary interaction parameters using closed cup flash point measurements.* Chemical Engineering Research and Design, *88*, **2010**, 81-86.
14. J. Gmehling, P. Rasmussen, *Flash points of flammable liquid mixtures using UNIFAC.* Industrial & Engineering Chemistry Fundamentals, *21*, **1982**, 186-188.
15. H.-J. Liaw, Y.-Y. Chiu, *A general model for predicting the flash point of miscible mixtures.* Journal of Hazardous Materials, *137*, **2006**, 38-46.
16. H.-J. Liaw, W.-H. Lu, V. Gerbaud, C.-C. Chen, *Flash-point prediction for binary partially miscible mixtures of flammable solvents.* Journal of Hazardous Materials, *153*, **2008**, 1165-1175.
17. G. Fayet, P. Rotureau, B. Tribouilloy, *Assessing the flammability of liquid mixtures by predictive approach as a complement to experimental measurements.* Chemical Engineering Transactions, *77*, **2019**, 781-786.
18. T. Gaudin, P. Rotureau, G. Fayet, *Combining mixing rules with QSPR models for pure chemicals to predict the flash points of binary organic liquid mixtures.* Fire Safety Journal, *74*, **2014**, 61-70.
19. H. He et al., *Predicting Thermal Decomposition Temperature of Binary Imidazolium Ionic Liquid Mixtures from Molecular Structures.* ACS Omega, *6*, **2021**, 13116-13123.
20. E. Torabian, M. Amin Sobati, *New structure-based models for the prediction of flash point of multi-component organic mixtures.* Thermochimica Acta, *672*, **2019**, 162-172.
21. Z. Jiao, C. Ji, S. Yuan, Z. Zhang, Q. Wang, *Development of machine learning based prediction models for hazardous properties of chemical mixtures.* Journal of Loss Prevention in the Process Industries, *67*, **2020**, 104226.
22. A.P. Toropova, A.A. Toropov, D. Leszczynska, J. Leszczynski, *The index of ideality of correlation: models of the flash points of ternary mixtures.* New Journal of Chemistry, *44*, **2020**, 4858-4868.
23. B. Aljaman, U. Ahmed, U. Zahid, V.M. Reddy, S.M. Sarathy, A.G. Abdul Jameel, *A comprehensive neural network model for predicting flash point of oxygenated fuels using a functional group approach.* Fuel, *317*, **2022**, 123428.

24. L.-T. Ye, Y. Pan, J.-C. Jiang, *Experimental Determination and Calculation of Auto- Ignition Temperature of Binary Flammable Liquid Mixtures.* Acta Petrolei Sinica(Petroleum Processing Section), *31*, **2015**, 753-759.
25. S. Shen, Y. Pan, X. Ji, Y. Ni, J. Jiang, *Prediction of the Auto-Ignition Temperatures of Binary Miscible Liquid Mixtures from Molecular Structures.* International Journal of Molecular Sciences, *20*, **2019**, 2084.
26. J. Yao et al., *Prediction of the flash points of binary biodiesel mixtures from molecular structures.* Journal of Loss Prevention in the Process Industries, *65*, **2020**, 104137.
27. W. Cao, Y. Pan, Y. Liu, J. Jiang, *A novel method for predicting the flash points of binary mixtures from molecular structures.* Safety Science, *126*, **2020**, 104680.
28. A.P. Toropova, A.A. Toropov, E. Carnesecchi, E. Benfenati, J.L. Dorne, *The index of ideality of correlation: models for flammability of binary liquid mixtures.* Chemical Papers, *74*, **2020**, 601-609.
29. D. Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* Journal of Chemical Information and Computer Sciences, *28*, **1988**, 31-36.
30. Z. Jiao, S. Yuan, Z. Zhang, Q. Wang, *Machine learning prediction of hydrocarbon mixture lower flammability limits using quantitative structure-property relationship models.* Process Safety Progress, *39*, **2020**, e12103.
31. L. Jiao, X. Zhang, Y. Qin, X. Wang, H. Li, *QSPR study on the flash point of organic binary mixtures by using electrotopological state index.* Chemometrics and Intelligent Laboratory Systems, *156*, **2016**, 211-216.
32. B. Wang, K. Xu, Q. Wang, *Prediction of upper flammability limits for fuel mixtures using quantitative structure–property relationship models.* Chemical Engineering Communications, *206*, **2019**, 247-253.
33. N. Chirico, P. Gramatica, *Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient.* Journal of Chemical Information and Modeling, *51*, **2011**, 2320-2335.
34. N. Chirico, P. Gramatica, *Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection.* Journal of Chemical Information and Modeling, *52*, **2012**, 2044-2058.
35. A.C. Atkinson, *Plots, Transformations and regression - An introduction to graphical methods of diagnostic regression analysis* (ed.). 1985: Oxford Science Publications.
36. Y. Jin, J. Jiang, Y. Pan, L. Ni, *Prediction of the auto-ignition temperature of binary liquid mixtures based on the quantitative structure–property relationship approach.* Journal of Thermal Analysis and Calorimetry, *140*, **2020**, 397-409.
37. Y. Ni, Y. Pan, J. Jiang, Y. Liu, C.-M. Shu, *Predicting both lower and upper flammability limits for fuel mixtures from molecular structures with same descriptors.* Process Safety and Environmental Protection, *155*, **2021**, 177-183.
38. Y. Pan, X. Ji, L. Ding, J. Jiang, *Prediction of Lower Flammability Limits for Binary Hydrocarbon Gases by Quantitative Structure—Property Relationship Approach.* Molecules, *24*, **2019**, 748.
39. D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, B. Creton, *Prediction of Flash Points for Fuel Mixtures Using Machine Learning and a Novel Equation.* Energy & Fuels, *27*, **2013**, 3811-3820.
40. S. Shen, X. Ji, Y. Pan, R. Qi, J. Jiang, *A new method for predicting the upper flammability limits of fuel mixtures.* Journal of Loss Prevention in the Process Industries, *64*, **2020**, 104074.
41. B. Wang, H. Park, K. Xu, Q. Wang, *Prediction of lower flammability limits of blended gases based on quantitative structure–property relationship.* Journal of Thermal Analysis and Calorimetry, *132*, **2018**, 1125-1130.
42. Y. Wang, F. Yan, Q. Jia, Q. Wang, *Distributive structure-properties relationship for flash point of multiple components mixture.* Fluid Phase Equilibria, *474*, **2018**, 1-5.
43. L. Zhou, B. Wang, J. Jiang, G. Reniers, L. Liu, *A mathematical method for predicting flammability limits of gas mixtures.* Process Safety and Environmental Protection, *136*, **2020**, 280-287.

6 Annexes

Liste des annexes :

- Annexe 1 : Publications relatives au développement de modèles QSPR dédiés à la prédiction de dangers physiques recensées dans cette étude – 2 pages.

Annexe 1 : Publications relatives au développement de modèles QSPR dédiés à la prédiction de dangers physiques recensées dans cette étude

Publication	Propriétés étudiées	Types de mélanges	Nature des constituants	Algorithmes utilisés *	Descripteurs moléculaires	Descripteurs de mélanges	Validations internes	Validation externe (partition)
Aljaman-2022 [23]	PE	Jusqu'à 5 constituants	Composés organiques	ANN	Descripteurs intégraux (1D, 2D)	Fractions massiques prises en compte dans les descripteurs molaires		Points-out
Cao-2020 [27]	PE	Binaires	Composés organiques	MLR	Fragments intra et inter-moléculaires	Pondération linéaire en fonction des fractions molaires et fragments intermoléculaires	Cross-Validation Y-Randomization	Points-out
Fayet-2019 [8]	PE	Binaires #	Composés organiques	MLR	Descripteurs intégraux (1D, 2D, 3D)	Différentes formes de pondérations en fonction des fractions molaires	Cross-Validation Y-Randomization	Compounds-out
Gaudin-2015 [9]	PE	Binaires	Composés organiques	MLR	Descripteurs intégraux (1D, 2D, 3D)	Différentes formes de pondérations en fonction des fractions molaires	Cross-Validation	Compounds-out
He-2021 [19]	$T_{d,5\%}$	Binaires	Liquides ioniques	MLR	Fragments	Différentes formes de pondérations en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out
Jiao-2016 [31]	PE	Binaires	Composés organiques	MLR ANN	Descripteurs intégraux (1D, 2D)	Pondération linéaire en fonction des fractions volumiques	Cross-Validation	Mixture-out
Jiao-2020a [21]	PE ; TAI ; TIE ; LSE	Binaires et ternaires	Composés organiques	ANN	Descripteurs intégraux (1D, 2D, 3D)	Pondération linéaire en fonction des fractions molaires	Cross-Validation	Non
Jiao-2020b [30]	LIE	Binaires	Composés organiques	MLR ; k-NN ; SVM; RF ; BT	Descripteurs intégraux (2D, 3D)	Pondération linéaire en fonction des fractions molaires	Cross-Validation	Points-out
Jin-2020 [36]	TAI	Binaires	Composés organiques	MLR SVM	Descripteurs intégraux (1D, 2D, 3D)	Différentes formes de pondérations en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out
Ni-2021 [37]	LIE ; LSE	Binaires et ternaires	Composés organiques	MLR	Descripteurs intégraux (3D)	Pondération linéaire en fonction des fractions volumiques	Cross-Validation	Compounds-out
Pan-2019 [38]	LIE	Binaires	Composés organiques	MLR	Descripteurs intégraux (1D, 2D, 3D)	Différentes formes de pondérations en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out
Saldana-2013 [39]	PE	Binaire et ternaire	Composés organiques	MLR SVM	Descripteurs intégraux (1D, 2D, 3D)	Différentes formes de pondérations en fonction des fractions molaires	Bootstrap	Compounds-out

Publication	Propriétés étudiées	Types de mélanges	Nature des constituants	Algorithmes utilisés *	Descripteurs moléculaires	Descripteurs de mélanges	Validations internes	Validation externe (partition)
Shen-2019 [25]	TAI	Binaires	Composés organiques	MLR	Fragments intra et inter-moléculaires	Pondération linéaire en fonction des fractions molaires et fragments intermoléculaires	Cross-Validation Y-Randomization	Compounds-out
Shen-2020 [40]	LSE	Binaires et ternaires	Composés organiques	MLR	Descripteurs intégraux (1D, 2D, 3D)	Pondération linéaire en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out
Torabian-2019 [20]	PE	Binaires #	Composés organiques	MLR ANN	Descripteurs intégraux (1D, 2D, 3D)	Différentes formes de pondérations en fonction des fractions molaires	Cross-Validation Bootstrap Y-Randomization	Points-out
Toropova-2020a [28]	PE	Binaires	Composés organiques	MLR	Quasi-SMILES	Quasi-SMILE combiné incluant fraction molaire		Points-out
Toropova-2020b [22]	PE	Ternaires	Composés organiques	MLR	Quasi-SMILES	Quasi-SMILE combiné incluant fraction molaire		Points-out
Wang-2018a [41]	LIE	Binaires et ternaires	Composés organiques	MLR	Descripteurs intégraux (3D)	Pondération linéaire en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out mixture-out compounds-out
Wang-2018b [42]	PE	Binaires et ternaires	Composés organiques	MLR	Descripteurs intégraux (3D)	Pondération non-linéaire en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out
Wang-2019 [32]	LSE	Binaires et ternaires	Composés organiques	SVM	Descripteurs intégraux (3D)	Pondération linéaire en fonction des fractions molaires	Cross-Validation Y-Randomization	Points-out mixture-out compounds-out
Yao-2020 [26]	PE	Binaires	Composés organiques	MLR	Fragments intra et inter-moléculaires	Pondération linéaire en fonction des fractions molaires et fragments intermoléculaires	Cross-Validation Y-Randomization	Points-out
Ye-2015 [24]	TAI	Binaires	Composés organiques	NLR	Contributions de Groupes	Pondération linéaire en fonction des fractions molaires		Points-out
Zhou-2020 [43]	LIE ; LSE	Binaires et ternaires	Composés organiques	MLR	Descripteurs intégraux (1D, 2D, 3D)	Pondération linéaire en fonction des fractions molaires	Cross-Validation	Points-out mixture-out compounds-out

* MLR : *Multi-Linear Regression* ; NLR : *Non Linear Regression*; SVM : *Support Vector Machine* ; ANN : *Artificial Neural Network* ; k-NN : *k-Nearest Neighbours*; RF : *Random Forest* ; BT : *Bootstrap Tree*

avec test d'application sur quelques mélanges ternaires

Institut national de l'environnement industriel et des risques
Parc technologique Alata • BP 2 • F-60550 Verneuil-en-Halatte
03 44 55 66 77 • ineris@ineris.fr • www.ineris.fr