

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

UMR 7177

THÈSE

présentée par

Ioana OPRISIU

soutenue le : 28 mars 2012

pour obtenir le grade de

Docteur de l'université de Strasbourg

Discipline : Chimie

Spécialité : Chemoinformatique

**Modélisation QSPR de mélanges binaires
non-additifs. Application au comportement
azéotropique**

THÈSE dirigée par :

M. VARNEK Alexandre

Professeur, Université de Strasbourg

RAPPORTEURS :

M. MORIN-ALLORY Luc

Professeur, Université d'Orléans

M. TETKO Igor

Docteur, Helmholtz Zentrum München

MEMBRES DU JURY :

M. ROGNAN Didier

Docteur, Université de Strasbourg

M. ROUSSEAUX Pascal

Docteur, Processium

Présentation de la société Processium

PROCESSIUM est une société indépendante, créée en 2002 par des ingénieurs issus de grands groupes industriels. Elle compte une trentaine de collaborateurs à ce jour. PROCESSIUM est basée à Lyon, sur le domaine scientifique LyonTech de la Doua au plus près des équipes de recherche avec lesquelles elle collabore. PROCESSIUM dispose de laboratoires de pointe d'analyses, de mesures (appui pour certaines demandes REACH) et d'essais, développé pour répondre aux projets industriels.

Domaines d'intervention :

- **Procédés Industriels & Unités de production**

PROCESSIUM conçoit et développe les procédés industriels pour la chimie, la pharmacie, le pétrole, l'environnement. Grâce à sa méthodologie originale et innovante, les meilleures solutions techniques sont identifiées et mises en œuvre jusqu'à la production des premiers lots.

- **Aide à l'Innovation Technologique et Commerciale**

PROCESSIUM accompagne ses clients dans leurs projets d'innovation depuis les études de veille technologique jusqu'au transfert de technologies sur leurs sujets techniques et stratégiques.

- **Propriétés Physiques**

PROCESSIUM met en œuvre des moyens humains, expérimentaux et logiciels pour les mesures, expertises et modélisations des propriétés physiques.

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude au Professeur Alexandre Varnek, pour m'avoir accueillie dans son laboratoire. Merci de m'avoir donné de nombreux conseils scientifiques et humains pendant toute ma thèse.

Ensuite, je voudrais adresser ma sincère reconnaissance au Professeur Luc Morin-Allory, au Docteur Igor Tetko, au Docteur Didier Rognan ainsi qu'au Docteur Pascal Rousseaux pour avoir accepté de juger mon travail et de faire partie du jury de thèse.

Je remercie aussi Docteur Pascal Rousseaux, le directeur de la société Processium, de m'avoir confié le travail d'une thèse CIFRE financée par sa société.

J'adresse également mes remerciements au Docteur Fabien Rivollet pour son encadrement, son expérience et sa rigueur, qui m'ont permis de mieux avancer dans mon travail.

Un grand merci également aux Docteurs Gilles Marcou et Dragos Horvath pour leur disponibilité au quotidien, pour m'avoir conseillée et guidée de nombreuses fois.

Je tiens également à remercier tous les collaborateurs et collègues membres du laboratoire: Docteur Igor Baskin, Docteur Vitaly Solov'yev, Docteur Natalia Kireeva, Docteur Vladimir Chupakhin (Bonne chance au États-Unis), Docteur Olga Klimchuk (merci pour ta bonne humeur permanente) et Docteur Fanny Bonachera (merci pour ta gentillesse).

J'adresse des remerciements particuliers à mes collègues doctorants Aurélie De Luca, Christophe Muller, Laurent Hoffer, Fiorella Ruggiu et Tetiana Khristova pour leur sympathie, leur bonne humeur et leur intérêt scientifique. Bonne chances à tous pour la soutenance.

Merci aussi à Sandrine Garcin et Danièle Ludwig pour leur aide permanente au niveau administratif.

Pour terminer, mes pensées vont tout naturellement à ma famille (surtout à mes parents qui mon toujours soutenu et à ma sœur), à mon compagnon et à mes amis.

ASNN	:	ASsociativ Neural Networks / Réseau de neurones associatif
BA	:	Balanced accuracy / Précision Balancée
COSMO-RS	:	COnductor like Screening MOdel for Realistic Solvents
DA	:	Applicability domain / Domaine d'Applicabilité
ISIDA	:	In Silico design and Data Analysis
MLR	:	Multi Linear Regression / Régression multilinéaire
PLS	:	Partial Least Square/Régression des moindres carrés partiels
QSPR	:	Quantitative Structure Property Relationship / Relation quantitative de structure propriété
ROC	:	Receiving Operating Characteristics
ROC AUC	:	Area Under the Curve ROC/ L'aire sous la courbe ROC
UNIFAC	:	UNiversal Fonctional group Activity Coefficient
VLE	:	Vapor Liquid Equilibrium / Équilibre Liquide Vapeur
VP	:	Voting Perceptron/ Perceptron Votant

Table de matières

Présentation de la société Processium.....	1
Remerciements	2
Table de matières.....	4
Table de figures.....	8
INTRODUCTION.....	10
PREMIERE PARTIE: Méthodologies de prédiction de propriétés physiques de corps purs et de mélanges.....	12
1 Méthodologie QSPR.....	14
1.1 Descripteurs	14
1.1.1 Descripteurs fragmentaux ISIDA.....	15
1.1.2 Représentation des descripteurs MOE.....	16
1.2 Méthodes fouille de données	17
1.2.1 Régression multilinéaire (MLR).....	17
1.2.2 Réseaux de neurones et ensembles de neurones.....	21
1.2.3 Le Perceptron.....	23
1.2.4 Les Machines à Vecteurs Supports (SVM).....	24
1.2.5 La Forêt Aléatoire (RF).....	26
1.2.6 Les modèles multiples.....	27
1.3 Critères de performances des modèles QSPR.....	28
1.3.1 Modèles de régression quantitatif.....	28
1.3.2 Modèles de classification.....	29
1.4 Validation des modèles QSPR	31
1.4.1 Leave One Out.....	31
1.4.2 Validation croisée par n-paquets (N-fold Cross Validation).....	32
1.4.3 Procédure de Y-randomisation	32
1.5 Domaine d'Applicabilité (DA) des modèles.....	33
1.5.1 Contrôle des fragments	33
1.5.2 La méthode Min-Max ou « Bounding Box ».....	33
1.5.3 z-kNN.....	34
1.5.4 1-SVM.....	34
1.6 Conclusion	34
1.7 Références.....	36
2 Prédiction des propriétés de mélanges de liquides binaires.....	38
2.1 Les mélanges: définitions	39

2.1.1	<i>Changement d'état</i>	39
2.1.2	<i>Mélange binaire liquide – vapeur</i>	39
2.1.3	<i>Pression partielle d'un gaz</i>	39
2.1.4	<i>Fraction molaire</i>	39
2.1.5	<i>Variables intensives et extensives</i>	40
2.1.6	<i>Équilibre d'un système</i>	40
2.1.7	<i>Phases, corps pur, mélange</i>	40
2.1.8	<i>Variance</i>	40
2.1.9	<i>Règle des phases</i>	41
2.1.10	<i>Potentiel chimique d'un constituant dans une phase</i>	41
2.2	Équilibre vapeur liquide d'un mélange binaire	41
2.2.1	<i>Mélange binaire et solution idéale</i>	42
2.2.2	<i>Mélange binaire réel</i>	45
2.2.3	<i>Coefficient d'activité et enthalpie libre d'excès</i>	48
2.2.4	<i>Expressions classiques de l'énergie libre d'excès</i>	49
2.2.5	<i>Le concept de composition locale</i>	51
2.2.6	<i>Notion de contributions de groupes</i>	52
2.2.7	<i>COSMO-RS (Conductor-like Screening Model for Real Solvents)</i>	54
2.2.8	<i>UNIFAC vs COSMO-RS</i>	56
2.3	Approches QSPR pour mélanges, développés antérieurement	58
2.3.1	<i>Descripteurs</i>	58
2.3.2	<i>Modèles QSPR pour des mélanges</i>	60
2.3.3	<i>Conclusion</i>	63
2.4	Références	65
DEUXIEME PARTIE: Développement d'une approche QSPR pour la prédiction des propriétés physico-chimiques des corps purs et de leurs mélanges binaires		68
2.5	Références	72
3	Courbe d'ébullition $T_b=f(X)$	73
4	Courbe d'équilibre vapeur-liquide $Y=f(X)$	86
4.1	Introduction	86
4.2	Méthodologie	86
4.2.1	<i>Donnés</i>	86
4.2.2	<i>Modélisation directe</i>	86
4.2.3	<i>Modélisation indirecte</i>	87
4.2.4	<i>Matrice de descripteurs</i>	88
4.2.5	<i>Machines d'apprentissage</i>	89
4.2.6	<i>Validation croisée</i>	89
4.2.7	<i>Validation externe</i>	91
4.2.8	<i>Benchmark</i>	92

4.3	Résultats	92
4.3.1	<i>N-Cross-validation</i>	92
4.3.2	<i>Validation externe</i>	92
4.3.3	<i>Domaine d'applicabilité</i>	95
4.3.4	<i>Benchmark</i>	95
4.4	Conclusion	97
4.5	Références	99
5	Classification azéotrope/zéotrope	100
5.1	Revue	100
5.2	Méthodologie	102
5.2.1	<i>Donnés</i>	102
5.2.2	<i>Matrice de descripteurs</i>	103
5.2.3	<i>Machines d'apprentissage</i>	103
5.2.4	<i>Validation croisée</i>	104
5.2.5	<i>Benchmark</i>	104
5.3	Résultats	105
5.3.1	<i>Validation croisée</i>	105
5.3.2	<i>Y-Randomization</i>	107
5.3.3	<i>Validation externe</i>	107
5.3.4	<i>Domaine d'applicabilité</i>	108
5.3.5	<i>Benchmark</i>	109
5.3.6	<i>Classification à partir des courbes VLE</i>	110
5.4	Conclusion	112
5.5	Références	114
6	Prédiction du point azéotrope (Taz Xw%)	115
7	Température d'ébullition des corps purs	122
8	Solubilité aqueuse	148
8.1	Introduction	148
8.2	Revue	148
8.3	Méthodologie	150
8.3.1	<i>Donnés</i>	150
8.3.2	<i>Machines d'apprentissage et descripteurs</i>	151
8.3.3	<i>Résultats</i>	151
8.3.4	<i>Analyse des points aberrants (outliers)</i>	152
8.3.5	<i>Domaine d'applicabilité</i>	154
8.4	"Solubility challenge"	154
8.4.1	<i>Présentation</i>	154
8.4.2	<i>Prédictions ISIDA-MLR</i>	154

8.5	Prédiction des molécules de la Chimiothèque Nationale Essentielle (CNE)	156
8.5.1	Données.....	156
8.5.2	Résultats.....	156
8.6	Conclusion	157
8.7	Références.....	159
TROISIEME PARTIE: Développement d'un outil de prédiction		160
9	Développement	161
9.1	"Multifragmentor"	161
9.1.1	Présentation	161
9.1.2	Fonctionnement.....	161
9.2	"MixturesPredictor"	162
9.2.1	Présentation	162
9.2.2	Fonctionnement.....	163
9.2.3	Fichiers XML.....	164
9.2.4	Résultats.....	165
9.2.5	Interface WEB.....	166
9.3	Conclusion	167
CONCLUSION GENERALE.....		168
10	Communications	171
10.1	Publications	171
10.2	Communications orales.....	171
10.3	Communications par affiche	172
11	Annexes	173
11.1	Descripteurs moléculaires calculés avec MOE	173
11.2	Prédiction du point azéotropique. Matériel supplémentaire	178
11.3	Composés formant les mélanges du jeu de modélisation pour la classification.....	183
11.4	Composés formant les mélanges du TS1 pour la classification.....	185
11.5	Composés formant les mélanges du TS2 pour la classification.....	186
11.6	Composés formant les mélanges du jeu de modélisation pour modéliser la courbe de bulle ...	188
11.7	Composés formant les mélanges du jeu de test pour la pour modéliser la courbe de bulle.....	189
11.8	Composés formant les mélanges du jeu d'entraînement pour modéliser $y=f(x)$	190
11.9	Composés formant les mélanges du jeu de test pour la modéliser $y=f(x)$	191

Table de figures

Figure 1-1 Descripteurs fragmentaux: Compte d'atomes, séquences et atomes unis.	16
Figure 1-2 Principe de mise en place de deux composantes principales PC_1 et PC_2 dans un espace à 3 dimensions (x_1, x_2, x_3) lors d'une analyse PCA	20
Figure 1-3 Schéma de l'architecture classique d'un réseau de neurones artificiels pour la modélisation structure (descripteurs x) – propriété (sortie y).	22
Figure 1-4 Schéma d'un perceptron.....	23
Figure 1-5 Classification binaire dans le cas d'une SVM	25
Figure 1-6 Régression par séparateurs à vastes marges.	26
Figure 1-7 Exemple d'arbre de décision.....	26
Figure 1-8 matrice de confusion pour deux classes	29
Figure 1-9 Courbe ROC. La diagonale correspond aux individus classés au hasard.....	31
Figure 1-10 Procédure validation croisée (5-paquets).....	32
Figure 1-11 Exemplification de la méthode 1-SVM.....	34
Figure 2-1 Les différents changements d'état.....	39
Figure 2-2 Diagrammes isotherme et isobare pour un mélange liquide binaire idéal.....	44
Figure 2-3 Obtention d'un diagramme isobare d'équilibre liquide-vapeur $y=f(x)$	45
Figure 2-4 Diagrammes isobares des mélanges homoazéotropiques	47
Figure 2-5 Mélanges hétéroazéotropiques.....	48
Figure 2-6 Illustration du concept de composition locale.....	51
Figure 2-7 Exemplification du concept de contribution de groupes. Image reprise sur le site d'Unifac Consortium	52
Figure 2-8 Profil σ pour la molécule d'eau. Image reprise du cours d'Andreas Klamt	54
Figure 2-9 Interactions entre segments de surface. Image reprise du cours d'Andreas Klamt	55
Figure 2-10 Schéma pour les calculs COSMO-RS. Image reprise du cours d'Andreas Klamt	56
Figure 2-11 Exemple de système binaire difficile à classer.....	70
Figure 4-1 Modélisation directe de la courbe VLE, $Y=f(X)$	87
Figure 4-2 Ajustement des valeurs (X,Y) avec un polynôme de 4ème degré.....	87
Figure 4-3 Construction de la matrice des descripteurs pour la modélisation directe de la courbe VLE $Y=f(X)$...	88
Figure 4-4 Construction de la matrice des descripteurs pour la modélisation indirecte de la courbe VLE $Y=f(X)$	89
Figure 4-5 Stratégie de cross-validation "Points out" pour la modélisation de la courbe VLE $Y=f(X)$	90
Figure 4-6 Stratégie de cross-validation "Mixtures Out" pour la modélisation de la courbe VLE $Y=f(X)$	90
Figure 4-7 Stratégie de cross-validation "Compounds out" pour la modélisation de la courbe VLE $Y=f(X)$	91
Figure 4-8 Exemple de prédictions de la courbe VLE $Y=f(X)$. Comparaison des prédictions de nos modèles avec ceux de COSMO-RS.	96
Figure 4-9 Prédictions de la courbe VLE $Y=f(X)$ pour le jeu de test avec DA=FrgCtrl IVAB (7 mélanges).	97
Figure 5-1 Calcul de la matrice de descripteurs pour des mélanges pour la classification.....	103

Figure 5-2 Représentation des courbes ROC pour SVM (gauche) et VP (droit) en 5-CV..... 106

Figure 5-3 Comparaison entre les modèles obtenus par Y-randomization et le jeu de données original. 107

Figure 5-4 Courbes ROC pour la VS1 (gauche) et VS2 contenant respectivement un ou deux composés purs connus (milieu et droite)..... 108

Figure 5-5 Mélanges dans le DA IVAB(2-4), contenant un composé nouveau..... 109

Figure 8-1 Distribution des valeurs de logS prédites en fonction des valeurs expérimentaux pour les modèles ASNN, ASVM, SQS et MLR en 5-CV..... 152

Figure 8-2 Structure des outliers..... 152

Figure 8-3 Performances de nos modèles par rapport aux ceux d'autres équipes 155

Figure 8-4 Prédiction de la solubilité par les modèles SQS, ISIDA MLR, MOE et ALOGPS..... 157

Figure 9-1 Capture d'écran de l'application graphique MixturesPredictor..... 163

Figure 9-2 Fichier XML utilisé par le MixturesPredictor..... 164

Figure 9-3 Exemple de fichier des résultats .csv 165

Figure 9-4 Exemple de courbe Tbp vs. X% 166

Figure 9-5 Capture d'écran de l'interface WEB de Predictor 167

INTRODUCTION

La connaissance du comportement azéotropique d'un mélange binaire est une information capitale pour l'industrie chimique. Un azéotrope est un mélange liquide qui se comporte comme un corps pur : il conserve une température d'ébullition fixe tout au long de la vaporisation du mélange. De même les phases liquides et vapeur en présence ont la même composition.

Lors de la conception de procédés de purification par voie thermique, une des méthodes les plus utilisées est la distillation. Cependant, dans ce cas, un azéotrope est une barrière de séparation infranchissable ne permettant pas la purification de tous les composés. Il est alors important d'identifier les azéotropes pour pouvoir identifier ces difficultés et orienter vers une autre méthode de purification.

Dans le domaine de la réfrigération, de nombreuses recherches sont menées pour trouver de nouveaux fluides écologiques, c'est-à-dire dont l'impact sur le réchauffement climatique et sur la couche d'ozone est faible. Cependant, il est très difficile de trouver une substance pure avec de bonnes performances de refroidissement et qui réponde aux exigences environnementales. Des études sont alors menées en mélanges, mais des problèmes se posent dès que la température et la composition des phases liquide/vapeur change lors des phases d'évaporation/condensation. Par conséquent, les réfrigérants qui forment des azéotropes, sont de plus en plus recherchés pour leur comportement, qui est similaire aux substances pures.

Les mesures expérimentales pour la courbe d'équilibre vapeur-liquide (VLE) peuvent indiquer avec précision l'existence d'un azéotrope, sa composition, pression et température. Néanmoins, ces mesures sont très coûteuses, chronophages, et pas toujours réalisables (produit qui se dégrade, ...). Pour ces raisons, des méthodes théoriques permettant d'aider à la détermination de l'existence des azéotropes et d'estimer des propriétés (compositions, température, pression) azéotropiques fiables sont très recherchées. Ce travail financé par la société Processium vise à identifier si des méthodes de type QSPR permettent de développer des modèles capables répondre à ces deux questions majeures :

- Le mélange de deux composés donnés forme-t-il un azéotrope ?
- Dans le cas d'une réponse positive à la précédente question, qu'elle sera la température et la composition d'un tel mélange ?

Des méthodes avec des fondements théoriques comme UNIFAC et COSMO-RS sont souvent utilisées pour décrire des mélanges.

UNIFAC utilise la notion de contributions de groupes qui considère que les propriétés d'une molécule se déduisent de façon additive de celles des groupes fonctionnels qui la composent. L'utilisation d'UNIFAC dépend de la disponibilité des paramètres d'interaction entre ces groupes, obtenus par régression d'un grand nombre de données d'équilibre vapeur-liquide (VLE) expérimentales. De plus les modèles UNIFAC ne peuvent pas prédire des molécules très grandes et complexes pour lesquelles le schéma de décomposition en groupes devient très difficile.

COSMO-RS (Conductor-like Screening Model for Real Solvents), une autre méthode pour la prédiction de comportement des mélanges, combine des calculs DFT avec des équations thermodynamiques. Malgré l'utilisation d'un faible nombre de paramètres obtenus par régression de données expérimentales, cette méthode est laborieuse à cause des longues et délicats calculs quantiques qu'elle nécessite.

Les limitations de ces modèles, montrent le besoin d'utilisation des méthodes QSPR. Pour cette raison, le travail de cette thèse a été centré sur le développement des méthodes QSPR pour des mélanges binaires.

Ce développement est original pour deux raisons. Premièrement, peu de travaux théoriques ont été publiés sur des mélanges dont les propriétés sont non-additives. Deuxièmement, plusieurs nouveaux aspects méthodologiques ont été introduits dans ce travail. Tout d'abord des descripteurs "spéciaux", capable de décrire des mélanges ont été proposés. Ensuite, une méthode robuste de validation a été utilisée. Enfin, un domaine d'applicabilité des modèles fiable a été proposé.

Cette thèse comporte trois parties : la première partie est une étude bibliographique des méthodologies de prédictions pour les corps purs et pour les mélanges; la deuxième partie décrit le développement d'une approche QSPR pour la prédiction des propriétés physico-chimiques des corps purs et de leurs mélanges binaires, tandis que la troisième partie est dédiée à l'implémentation d'un outil de prédiction qui englobe tous les modèles développés pendant ce travail.

PREMIERE PARTIE:
Méthodologies de prédiction de
propriétés physiques de corps
purs et de mélanges

Ce chapitre est dédié à l'étude bibliographique des méthodologies QSPR. Dans une première partie la méthodologie QSPR classique et les éléments indispensables au développement de modèles seront discutés. Dans la deuxième partie les méthodes les plus utilisées aujourd'hui pour la prédiction des propriétés des mélanges (UNIFAC et COSMO-RS), ainsi que des modèles QSPR seront analysés.

1 Méthodologie QSPR

Le **QSPR** (**Quantitative Structure-Property Relationships**) est le procédé par lequel des liens quantitatifs sont établis entre la structure moléculaire d'un ensemble de composés avec une propriété physico-chimique.

Les grandes phases de développement d'un modèle QSPR peuvent être décrites comme suit :

- Choisir des descripteurs adaptés au problème structure-propriété,
- Exploiter les valeurs des descripteurs comme variables, afin de définir une relation qui les corrèle à la propriété en question, à l'aide de machines d'apprentissage. C'est la fouille de données (1.2).
- Établir des critères de performance et de validation qui aideront au choix des meilleurs modèles pour le problème posé et estimer des incertitudes de prédiction.

1.1 Descripteurs

Un descripteur est un paramètre numérique propre à une structure chimique donnée. Ces descripteurs représentent un moyen de corréliser une structure chimique et une valeur de propriété physique. Les descripteurs sont finalement des intermédiaires numériques se substituant à la molécule elle-même.

Il existe à ce jour plus de 6000 descripteurs répertoriés [1] qui sont classés en 4 types.

- Les **descripteurs 1-D** sont calculés à partir de la formule brute de la molécule et ils représentent des propriétés très générales, comme le poids moléculaire, nombre d'atomes, etc.
- Les **descripteurs 2-D** sont obtenus à partir de la structure 2D de la molécule. Dans cette catégorie rentrent les indices (topologiques, constitutionnels), fragmentaux ou les propriétés physico-chimiques (donneur de liaison H, accepteur de liaison H, cation, anion, etc.)
- Les **descripteurs 3-D** sont calculés à partir la structure 3D de la molécule. Ils peuvent être quantiques, de surfaces moléculaires ou de volume moléculaire.
- Les **descripteurs 4-D** sont obtenus par le calcul des champs d'interactions moléculaire (COMFA, GRID[2, 3]) entre une molécule et une sonde représentée par une autre molécule (eau, amide, etc.).

Dans cette thèse, les descripteurs fragmentaux ISIDA (1D ou 2D) et les descripteurs MOE (2D) ont été particulièrement étudiés.

1.1.1 Descripteurs fragmentaux ISIDA

Développés au laboratoire d'Infochimie, les descripteurs fragmentaux [4-6] ont été très largement employés durant cette thèse. Les descripteurs ISIDA peuvent être classés en deux catégories: les SMF (Substructural Molecular File) et les IPLF[7] (ISIDA Property-Labelled Fragment). Parmi ces deux classes on a utilisé constamment les descripteurs SMF.

Les SMF sont des sous-structures catégorisées en deux types:

Les séquences correspondent à des chaînes d'atomes et/ou de liaisons. Une séquence correspond au chemin le plus court à parcourir pour relier 2 atomes dans le graphe moléculaire.

Ils sont classés en trois catégories :

- IA : Chaîne d'atomes uniquement
- IB : Chaîne de liaisons uniquement
- IAB : Chaîne d'atomes et de liaisons

Pour chaque type de séquence, un nombre minimal ($n_{\min} \geq 2$) et un nombre maximal ($n_{\max} \leq 15$) d'atomes inclus sont définis. Dans la codification de la séquence, la nature A, B ou AB est suivie par le nombre minimum et maximum d'atomes. Par exemple : la fragmentation I(A, 2-8) correspond à des séquences d'atomes contenant de 2 à 8 atomes, tandis que I(AB, 2-8) correspond à des séquences d'atomes et de liaisons impliquant de 2 à 8 atomes.

Un atome uni est représenté par un atome central avec sa première sphère de coordination, incluant les atomes (A) et/ou les liaisons voisins (B) et se notent IIA, IIB ou IIAB respectivement. L'état d'hybridation et l'environnement des atomes (**IIHy**) peuvent également être considérés et codés par des types atomiques consacrés (CD : Csp² ; CB : C aromatique ; CT : Csp ; CO : carbonyle ; CN : nitrile ; NI : Nsp²).

Les atomes unis peuvent être agrandis en ajoutant un nombre n_{\min} et n_{\max} de sphères de coordination autour de l'atome concerné. Dans ce cas, la codification « III » remplace « II » (IIIA, IIIB ou IIIAB).

Le type IV correspond à des atomes unis constitués de chemins de longueur fixe entre l'atome central et chaque atome périphérique. Les longueurs n_{\min} et n_{\max} des

chemins sont des paramètres de ces descripteurs. Cela forme des fragments de type IVA, IVB et IVAB.

Le plus simple des fragments est le type AC (Atome Count - Compte d'Atomes) contenant qu'un seul atome.

La Figure 1-1 montre un exemple de différentes fragmentations.

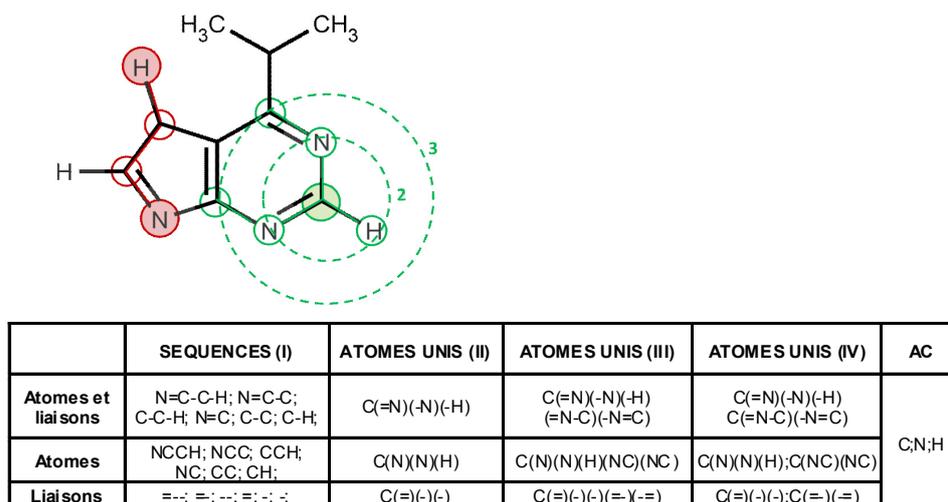


Figure 1-1 Descripteurs fragmentaux: Compte d'atomes, séquences et atomes unis.

Le descripteur est le nombre d'occurrence de fragment dans la molécule.

1.1.2 Représentation des descripteurs MOE

MOE (Molecular Operating Environment)[8] est un logiciel qui permet de calculer des descripteurs 2. Ces descripteurs ont été utilisés pour réaliser des comparaisons avec les descripteurs fragmentaux ISIDA.

Les descripteurs moléculaires 2D sont basés sur des tables de connectivité, codant la nature et le type des liaisons interatomiques.

Parmi les descripteurs 2D proposés, nous n'avons sélectionné que ceux plus pertinents et non redondants. En total, 62 sur un total de 184 descripteurs ont été choisis pour la modélisation : différentes propriétés physico-chimiques calculées à partir de la table de connectivité (densité, somme de charges formelles, poids moléculaire, volume Van der Waals, etc.) ; compte d'atomes et de liaisons ; indices de connectivité Kier&Hall ; matrice de descripteurs d'adjacences et de distances ; descripteurs pharmacophoriques; charges partielles.

La liste entière de descripteurs MOE utilisés est donnée dans l'Annexe 11.1.

1.2 Méthodes fouille de données

Le but d'une modélisation QSPR est de trouver une fonction permettant de relier de manière quantitative ou qualitative une propriété d'une molécule i , y_i avec ses descripteurs x_i^j , de la manière suivante:

$$y_i = f(x_i^j) \quad (1-1)$$

Plusieurs méthodes pour définir cette fonction « f » ont été utilisées : la *MLR* (Régression Multi-linéaire), *ASNN* (ASsociative Neural Networks, ensemble de réseau de neurones), la régression *PLS* (Partial Least Square), la *SVM* (Séparateurs à Vastes Marges), *RF* (Random Forest, Forêt Aléatoire) et le *VP* (Perceptron Votant). Les quatre premières méthodes d'apprentissage ont été employées pour construire des modèles quantitatifs. De plus, la *SVM* et le *VP* ont été utilisés pour construire des modèles de classification.

1.2.1 Régression multilinéaire (MLR)

1.2.1.1 Principes généraux

La **régression linéaire multiple** (MLR) est une généralisation, à p variables explicatives, de la régression linéaire simple, qui, pour une série de couples de points (x, y) , détermine les coefficients **a** et **b** de l'équation $y = ax + b$ de la droite passant le plus près de l'ensemble de ces points.

La méthode MLR fait l'hypothèse que la propriété y dépend linéairement des différentes variables x_1, x_2, \dots, x_n (dans notre cas les descripteurs), selon la relation :

$$y = a_0 + \sum_{j=1}^n a_j x_j \quad (1-2)$$

Dans notre cas, pour m molécules décrites par n descripteurs chacune, l'équation de régression s'écrit :

$$y_i = a_0 + \sum_{j=1}^n a_j x_i^j + \varepsilon_i \quad (1-3)$$

où y_i est la propriété de la molécule i , x_i^j est le descripteur j de la molécule i , ε_i est l'erreur du modèle résumant les informations manquantes qui permettrait d'expliquer linéairement les valeurs de y à l'aide des n variables x_i^j .

En adoptant une écriture matricielle, l'équation (1-3) devient:

$$Y = Xa + \varepsilon \quad (1-4)$$

où Y , X , a et ε représentent respectivement le vecteur de propriétés, la matrice des descripteurs, le vecteur des coefficients et le vecteur des erreurs de régression.

1.2.1.2 ISIDA-MLR

Ce logiciel est inclus dans le programme ISIDA [4, 9] développé au laboratoire d'Infochimie y comprend de nombreuses méthodes et outils de chimoinformatique. A partir de la matrice des descripteurs, remplie par les occurrences des fragments pour les molécules de jeu de données considéré, et de leurs propriétés expérimentales respectives, ISIDA/MLR construit des équations multilinéaires du type :

$$Propriété = a_0 + \sum_{i=1}^n a_i N_i \quad (1-5)$$

Où N_i représente l'occurrence du $i^{\text{ème}}$ fragment impliqué dans le modèle, et a_i est son coefficient associé.

Des méthodes de *stepwise* [10-12] ont été utilisés afin de sélectionner les variables N_i les plus pertinentes. La méthode *stepwise* est une méthode de sélection de variables pas à pas. Partant d'un modèle sans variable, on introduit d'abord celle étant la plus corrélée avec la propriété. A chaque étape, une variable est ajoutée, la plus corrélée avec le résidu. Le modèle résultat est accepté s'il est plus significatif que le précédent et que tous les paramètres sont significatifs. Ensuite, la significativité des paramètres est contrôlée de façon plus stricte et les variables du modèle sont ôtées une à une en choisissant la moins significative à chaque fois. Quand tous les paramètres sont significatifs la procédure s'arrête.

En variant la longueur minimale et maximale des fragments, des centaines d'ensembles de descripteurs sont générés. Ensuite, des modèles QSPR sont développés pour chaque ensemble de descripteurs.

Ces tests statistiques permettant d'évaluer la qualité du modèle sont le test de Fischer le test de Student. Les modèles sont ensuite évalués par les méthodes statistiques Leave One Out (LOO) et de validation croisée à n-paquets (n-fold cross validation) internes ou externes (voir le chapitre 1.4). Ces méthodes sont utilisées pour discriminer les modèles ayant les meilleures capacités prédictives. Les modèles sont caractérisés par les critères statistiques R^2 (coefficient de détermination), Q^2 (coefficient de détermination issu de la validation croisée LOO), le RMSE (erreur-

type) et MAE (moyenne des erreurs absolues). Les formules sont données dans le chapitre 1.3.

Les modèles choisis comme étant les plus pertinents sont regroupés dans un modèle consensus (**CM**) en moyennant les résultats de ces modèles individuels. (voir 1.2.5).

1.2.1.3 SQS (*Stochastic QSAR Sample*)

Une autre approche MLR utilisée dans ce travail, utilise une méthode stochastique SQS [10] (*Stochastic QSAR Sampler*), basée sur un algorithme génétique.

L'algorithme génétique fonctionne selon le principe d'évolution des populations de taille N , égale au nombre de molécules ou d'individus présents dans le jeu de données. Chaque individu est représenté par un chromosome, qui correspond à une chaîne de bits de longueur égale au nombre de descripteurs de la molécule.

L'algorithme génétique est constitué de deux principales étapes : initiation et évolution.

Dans la phase d'initiation, plusieurs sous-ensembles de descripteurs sont choisis aléatoirement. Pour chaque sous-ensemble de descripteurs des modèles sont développés.

Dans la phase d'évolution, des nouvelles générations « filles » de taille N , sont créées. Les individus « fils » sont obtenus à partir des opérations de croisement entre deux « parents » ou par la mutation d'un « parent ».

A chaque génération, les meilleurs individus sont sélectionnés à partir d'un critère d'optimisation. Le critère correspond au coefficient de détermination des modèles calculés lors d'une validation interne des modèles.

1.2.1.4 Régression des moindres carrés partiels (*PLS - Partial Least Square*)

L'approche PLS [13] est une méthode statistique postulant l'existence de relations entre des variables observées et des variables latentes. Ce type de modèles est généralement appelé modèle d'équations structurelles à variables latentes. Une variable latente est une variable construite à l'aide des données ou variables initiales du problème. Dans le cas de la PLS, c'est une combinaison linéaire des variables du problème.

Afin de pouvoir expliquer la PLS il faut commencer par expliquer l'analyse et la régression en composants principales. La régression basée sur PCA est un cas particulier de la PLS.

L'analyse par composantes principales (PCA, pour *Principal Component Analysis*) consiste à transformer un jeu de variables corrélées entre elles en un nouveau jeu de variables, appelées composantes principales. De plus, un sous-ensemble de ces nouvelles variables est retenu afin de réduire la dimensionnalité du système en perdant un minimum d'information. La Figure 1-2 représente graphiquement le principe de la méthodologie pour 3 variables x_1 , x_2 et x_3 et deux composantes principales PC_1 et PC_2 .

Considérons l'ensemble des données comme un espace à n dimensions, chaque dimension représentant une variable de la base de données. Chaque échantillon de la base de données est donc un point dans cet espace à n dimensions $[x_1 \dots x_n]$.

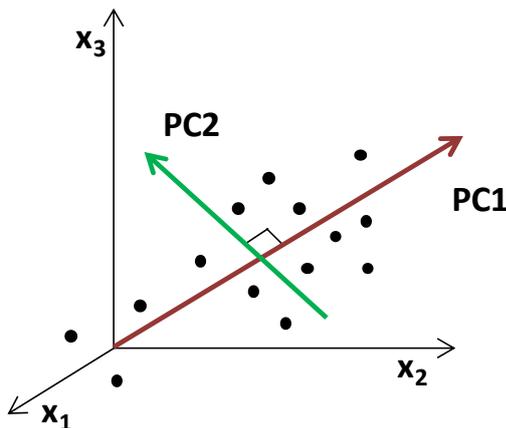


Figure 1-2 Principe de mise en place de deux composantes principales PC_1 et PC_2 dans un espace à 3 dimensions (x_1 , x_2 , x_3) lors d'une analyse PCA

Dans un premier temps, il s'agit de construire une première composante principale PC_1 dont la coordonnée t_1 correspond à la combinaison linéaire des n variables passant au plus proche de tous les points du système.

$$t_1 = c_{1,1}x_1 + c_{2,1}x_2 + \dots + c_{n,1}x_n \tag{1-6}$$

où les constantes $c_{i,1}$ sont choisies pour maximiser la variance sur t_1

Une deuxième composante principale PC_2 est ensuite construite orthogonale à la première. Ainsi, PC_1 et PC_2 sont indépendantes.

$$t_2 = c_{1,2}x_1 + c_{2,2}x_2 + \dots + c_{n,2}x_n \tag{1-7}$$

où les constantes $c_{i,2}$ sont choisies de sorte que la variance sur t_2 soit maximale tout en imposant l'orthogonalité des vecteurs $[c_{i,1}]$ et $[c_{i,2}]$. L'ajout de composantes principales supplémentaires se fait de la même manière.

La régression aux moindres carrés partiels (PLS, pour *Partial Least Squares* ou *Projection on Latent Structures*) est en quelque sorte une version supervisée de la PCA. Les variables latentes t_i sont des combinaisons linéaires de x_i mais les coefficients sont choisis pour optimiser la corrélation avec la propriété. Chaque nouvelle variable latente est corrélée avec le bruit résiduel des précédents modèles et orthogonale aux précédentes variables.

Ainsi:

$$Y = a_1 t_1 + a_2 t_2 + \dots + a_n t_n \quad (1-8)$$

Les variables latentes t_i sont elles-mêmes des combinaisons linéaires des variables indépendantes (x_i):

$$\begin{aligned} t_1 &= b_{1,1} x_1 + b_{1,2} x_2 + \dots + b_{1,p} x_p \\ t_2 &= b_{2,1} x_1 + b_{2,2} x_2 + \dots + b_{2,p} x_p \\ t_i &= b_{i,1} x_1 + b_{i,2} x_2 + \dots + b_{i,p} x_p \end{aligned} \quad (1-9)$$

Le nombre de variables latentes qui peuvent être générés est le plus petit nombre de variables ou d'observations.

1.2.2 Réseaux de neurones et ensembles de neurones

L'approche par réseaux de neurones est analogue aux systèmes de neurones biologiques. Les neurones biologiques permettent de transmettre et de traiter des informations en faisant circuler des signaux électriques dans un réseau constitué d'axones. L'information est propagée d'un neurone à un d'autres neurones qui y sont connectés via les synapses.

En modélisation, un neurone possède des entrées par lesquels lui arrivent des données. A chacune de ces entrées est associé un poids w , qui est ajusté au cours de l'apprentissage. Le neurone renvoie un signal de sortie si la somme pondérée des entrées dépasse un certain seuil.

Un réseau de neurones est constitué de multiples couches: une couche d'entrée représentée par les descripteurs, une ou plusieurs couches cachées et une couche de sortie représentés pas les propriétés à modéliser. Les neurones d'une couche sont interconnectés avec les neurones d'une couche voisine.

La Figure 1-3 illustre un réseau de neurones classique utilisé pour la modélisation structure (descripteurs x) – propriété (sortie y) à 3 couches.

La **couche d'entrée** compte autant de neurones que de descripteurs pour le jeu d'apprentissage.

Chaque neurone de la couche cachée réalise des opérations de sommations pondérées, à l'issue desquelles le neurone peut être activé ou non. Chaque neurone de la couche d'entrée est relié par des synapses à chacun des neurones de la **couche cachée**, et au niveau de ces synapses virtuelles, se trouvent des poids w_i permettant de moduler l'importance relative de chacun des descripteurs.

La **couche de sortie** compte autant de neurones que de propriétés modélisées. Dans notre cas une seule propriété à été modélisée.

Pendant la phase d'apprentissage du modèle par un réseau de neurones, les molécules sont présentées une par une aux neurones de la couche d'entrée. Les poids w_i associées aux neurones d'entrée sont ajustés itérativement, afin de minimiser l'erreur entre la propriété calculée et la propriété expérimentale.

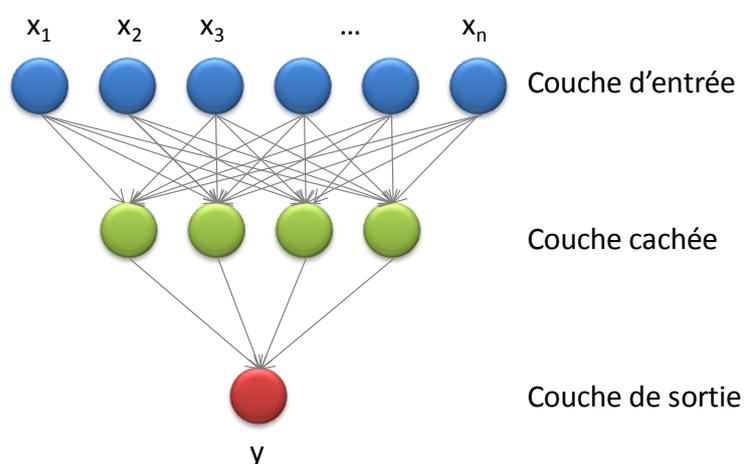


Figure 1-3 Schéma de l'architecture classique d'un réseau de neurones artificiels pour la modélisation structure (descripteurs x) – propriété (sortie y).

Beaucoup d'études QSPR employant les réseaux de neurones sont publiées [14] et conduisent à des performances élevées. Pendant le travail de cette thèse nous utiliserons à plusieurs reprises le logiciel ASNN qui implémente des réseaux de neurones.

Développé par le Dr. I. Tetko, le programme ASNN (Associative neural networks)[15] est un programme de modélisation structure-propriété couramment utilisé au laboratoire d'Infochimie: la propriété d'une molécule (y_i) est obtenue en moyennant les prédictions des 100 réseaux de neurones entraînés sur un même jeu d'apprentissage. De plus, le résultat de ce modèle consensus est corrigé grâce à un facteur de correction calculé avec la technique des k-plus proches voisins kNN dans l'espace de modèles.

$$\bar{y}'_i = \bar{y}_i + \frac{1}{k} \sum_{j \in N_k(i)} (y_{exp,j} - \bar{y}_j) \quad (1-10)$$

Dans l'équation (1-10), la prédiction du composé i est corrigée par les prédictions de l'ensemble des molécules j représentant les k voisins les plus proches de la molécule i , $N_k(i)$. Le coefficient de corrélation de Spearman entre les deux vecteurs de prédiction \bar{y}_i et \bar{y}_j est utilisé, afin de déterminer les molécules voisines de la molécule i .

Pour chaque réseau de neurones, le jeu de données est découpé aléatoirement à part égale en un jeu d'apprentissage et de test interne. Chaque modèle est donc testé sur le jeu de test interne.

1.2.3 Le Perceptron

Le perceptron peut être vu comme le type de réseau de neurones le plus simple. Il a pour objectif d'imiter la stimulation d'un neurone (sortie) par des neurones voisins (entrée). Les neurones sont binaires, un neurone a la valeur 1 s'il est actif et la valeur -1 sinon. L'état du neurone de sortie correspond à la réponse du perceptron tandis que les p neurones d'entrée sont les variables fonctionnelles sur lesquelles opère le perceptron. Le perceptron peut être alors comme une fonction

$$f_{\{-1,1\}} = \begin{cases} \{-1,1\}^p \rightarrow \{-1,1\} \\ (x_1, \dots, x_p) \mapsto f(x_1, \dots, x_p) \end{cases} \quad (1-11)$$

que l'on peut schématiser sous la forme suivante:

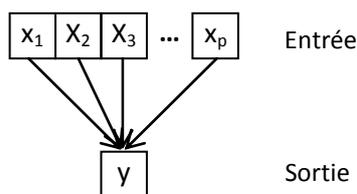


Figure 1-4 Schéma d'un perceptron

Pour définir l'action de la fonction f , on considère que les neurones d'entrée peuvent stimuler le neurone de sortie par le biais de poids synaptiques notés $\omega_1, \dots, \omega_p$. On somme alors les poids synaptiques des neurones actifs pour calculer le stimuli généré par une entrée (x_1, \dots, x_p) .

$$\sum_{j=1}^p \omega_j x_j \quad (1-12)$$

Si cette quantité est supérieure à un certain seuil d'activation ϕ , on considère que le neurone de sortie est activé. Ainsi, nous pouvons préciser l'action de la fonction f :

$$f_{\{-1,1\}} = \begin{cases} 1, si \sum_{j=1}^p \omega_j x_j > 0 \\ -1, sinon \end{cases} \quad (1-13)$$

1.2.4 Les Machines à Vecteurs Supports (SVM)

Cette approche est également connue sous le nom de *Séparateurs à Vastes Marges* et a été introduite par Vapnik[16]. La méthode peut être utilisée pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon, ou de régression, c'est-à-dire prédire la valeur numérique d'une variable. La méthode SVM a été décrite dans des nombreux ouvrages comme par exemple dans [17].

Dans le cadre d'un problème non linéairement séparable, SVM reconsidère le problème dans un espace de dimension supérieure. Dans ce nouvel espace, il est alors probable qu'il existe un séparateur linéaire. Plus précisément, une transformation non-linéaire ϕ est appliquée aux vecteurs de descripteurs, à l'aide d'une fonction noyau (appelée **kernel function**). Les fonctions noyaux les plus courantes sont polynomiales, gaussiennes (RBF) ou sigmoïdes. L'espace d'arrivée $\phi(X)$ est appelé **espace de redescription**.

Le développement des modèles SVM de classification et de régression a été faite à l'aide du logiciel libSVM [18], une librairie programmée en C.

1.2.4.1 Classification par SVM (SVC)

Le SVC peut être utilisés pour résoudre des problèmes de classification binaire. La résolution de ce problème passe par la construction d'un hyperplan, dans l'espace de descripteurs, capable de séparer les composés appartenant à ces deux classes.

L'hyperplan séparateur optimal est choisi de manière à maximiser la marge entre les plus proches voisins (**vecteurs support**) et cet hyperplan.

Une erreur de classification est traitée par une pénalité d'autant plus importante que celle-ci est loin du plan de l'hyperplan séparateur. Le facteur de proportionnalité est appelé le **coût**. La Figure 1-5 représente un exemple de problème linéairement séparable, pour deux classes données.

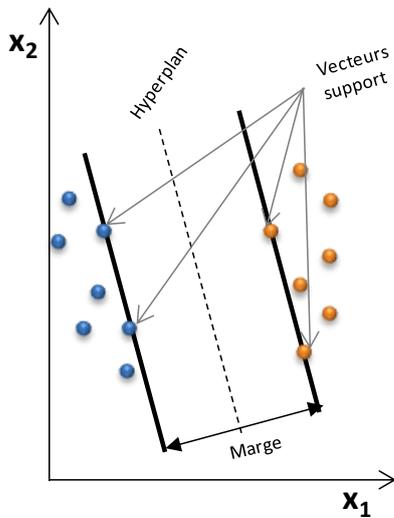


Figure 1-5 Classification binaire dans le cas d'une SVM

1.2.4.2 Régression par SVM (SVR)

Dans le cas du SVR, au lieu d'identifier un hyperplan optimal pour séparer les données comme dans le SVC, le but est d'optimiser une fonction de régression $f(x)$ sur les données définie dans l'espace de redescription. Cette fonction est recherchée sous la contrainte que :

$$|y - f(x)| < \varepsilon \quad (1-14)$$

Où ε représente le **seuil d'erreur**, c'est-à-dire la précision attendue sur les valeurs de la propriété (y).

L'espace de redescription est recherché de manière que le modèle linéaire ne produise plus d'erreurs. Cependant pour pouvoir traiter des jeux de données dans lesquelles subsistent des erreurs, on introduit un coût proportionnel à :

$$\xi = |y - f(x) - \varepsilon| \quad (1-15)$$

La Figure 1-6 représente la SVR pour des problèmes de régression.

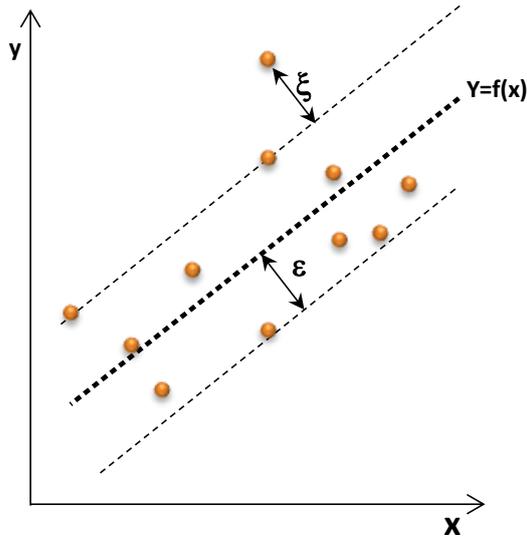


Figure 1-6 Régression par séparateurs à vastes marges.

1.2.5 La Forêt Aléatoire (RF)

La **forêt aléatoire** est un classificateur constitué d'un ensemble de modèles de classification individuels ; des arbres de décision. La sortie est la classe majoritaire parmi les sorties de chaque arbre individuel. L'algorithme pour induire une forêt aléatoire a été développé par Leo Breiman [19].

Un **arbre de décision** est un outil d'aide à la décision qui représente une problème plus ou moins complexe sous la forme d'un arbre, de façon à faire apparaître à l'extrémité de chaque branche les différents résultats possibles, en fonction des décisions prises à chaque étape. Un arbre de décision est lisible, s'exécute facilement et nécessite peu d'hypothèses a priori, d'où son utilisation fréquent pour des problèmes de classification.

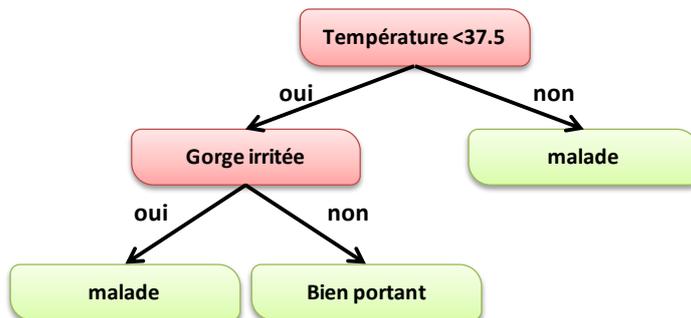


Figure 1-7 Exemple d'arbre de décision

Dans une structure d'arbre, les feuilles représentent les étiquettes de classe, les nœuds représentent les attributs et les branches représentent leurs valeurs.

La construction d'un arbre de décision se compose des étapes suivantes : (à partir de la racine de l'arbre)

1. Choix d'une variable de partitionnement parmi les attributs qui décrivent les données d'apprentissage.

2. Choix d'une ou de plusieurs valeurs de coupure de cette variable pour définir la partition, si la validation est numérique.

3. Décider si le nœud courant est terminal, c'est-à-dire décider si un nœud doit être étiqueté comme une feuille. Par exemple : tous les exemples sont dans la même classe, il y a moins d'un certain nombre d'erreurs, etc.

Recommencer les étapes 1 et 2 avec chacun des nœuds qui ne remplissent pas les critères pour devenir des feuilles.

4. Élaguer, si besoin, l'arbre.

5. Affecter à chaque feuille une classe.

Dans le cas des forêts aléatoires les arbres de décision sont construits sans élagage à partir d'un sous-ensemble aléatoire des variables du problème et le critère utilisé à chaque nœud est le gain d'information.

1.2.6 Les modèles multiples

Afin de réduire les erreurs de prédiction de modèles plusieurs modèles sont développés sur le même jeu de données et réunis dans un **modèle multiple** ou **consensus**. Ces modèles sont obtenus par différentes méthodes:

- En combinant de modèles individuels développés à partir du même jeu d'apprentissage et du même ensemble initial de descripteurs. Tous les modèles sont obtenus avec la même méthode d'apprentissage. C'est le cas de l'ASNN où le réseau de neurones est entraîné 100 fois avec des poids initiaux différents, générant 100 modèles différents. Par conséquent, la propriété de chaque molécule est calculée moyennant les valeurs prédites par les 100 modèles.
- En combinant de modèles individuels développés à partir du même jeu d'apprentissage avec un ensemble de descripteurs différents pour chaque modèle individuel. Tous les modèles sont obtenus avec la même méthode d'apprentissage. La propriété de chaque molécule est calculée comme moyenne arithmétique de prédictions données par les modèles individuels. Cette méthode a été utilisée pour la majorité des modèles développés au cours de cette thèse, que ce soit des modèles de régression ou de classification. Dans le cas d'une classification la valeur de la classe est établie par un vote à la majorité.

- En combinant modèles individuels développés à partir du même jeu d'apprentissage avec un ensemble de descripteurs différents pour chaque modèle individuel. Tous les modèles sont issus de plusieurs algorithmes d'apprentissage différents. La prediction finale est obtenue soit en moyennant les predictions données par chaque modèles, soit par une méthode appelé « stacking »[20]. Par cette méthode une prediction consensus est obtenue en développant une régression linéaire ayant comme variables indépendantes les predictions données par chaque algorithmes d'apprentissage individuel. A notre connaissance, cette approche n'a jamais été utilisée en chemoinformatique.

1.3 Critères de performances des modèles QSPR

Afin de comparer des modèles obtenus par différents méthodes, des critères de performances sont ont été définis:

1.3.1 Modèles de régression quantitatif

Les performances des modèles quantitatifs sont mesurées à travers le coefficient de détermination (R^2), l'erreur quadratique moyenne ($RMSE$) et l'erreur absolue moyenne (MAE).

Coefficient de détermination

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{pred,i} - y_{exp,i})^2}{\sum_{i=1}^n (y_{exp,i} - \bar{y}_{exp})^2} \quad (1-16)$$

Erreur quadratique moyenne:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{pred,i} - y_{exp,i})^2} \quad (1-17)$$

Erreur absolue moyenne

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{pred,i} - y_{exp,i}| \quad (1-18)$$

Avec i l'indice de la molécule, n le nombre de molécules prédites, $y_{pred,i}$ la propriété prédite par le modèle de régression, $y_{exp,i}$ la valeur expérimentale de la propriété et \bar{y}_{exp} la moyenne des valeurs expérimentale. Les paramètres $RMSE$ et MAE estiment des critères de dispersion des propriétés calculées par rapport aux propriétés expérimentales

Dans l'étape de validation des modèles, les paramètres sont calculés avec l'équation (1-16), le coefficient de détermination étant nommé coefficient de validation croisée (Q^2). Ce coefficient peut être utilisé pour comparer deux modèles sur un même jeu de données et ayant une même complexité (par exemple le même nombre de degrés de liberté).

1.3.2 Modèles de classification

Les performances des modèles de classification doivent être établies en comparant les données qualitatives prédites et expérimentales des molécules. Les données qualitatives sont appelées classes. Afin d'aider à définir des critères de performance numériques basés sur les classes, l'outil de base est la matrice de confusion (cf. Figure 1-8).

		Prédit	
		Positif (1)	Négatif (0)
Expérimental	Positif (1)	Vrais Positifs	Faux Négatifs
	Négatif (0)	Faux Positifs	Vrais Négatifs

Figure 1-8 matrice de confusion pour deux classes

Les classes expérimentales et prédites par le modèle sont représentées respectivement en lignes et en colonnes.

Les vrais positifs et les vrais négatifs correspondent aux molécules correctement classées comme positives (1) et négatives (0), respectivement. Les faux négatifs et les faux positifs, représentent les molécules classées comme étant négatives et positives, respectivement, tout en appartenant à la classe positive et négative, respectivement.

A partir de cette matrice, différents paramètres sont calculés :

- Le rappel
- La précision
- La courbe ROC

1.3.2.1 Le rappel

Le rappel explique la capacité du modèle à prédire correctement les exemples d'une certaine classe.

$$\begin{aligned} \mathbf{Rappel(0)} &= \frac{\mathbf{Tn}}{\mathbf{Tn + Fp}} \\ \mathbf{Rappel(1)} &= \frac{\mathbf{Tp}}{\mathbf{Tp + Fn}} \end{aligned} \quad (1-19)$$

1.3.2.2 La Précision

La précision représente la capacité du modèle de classer dans une classe que les exemples appartenant à celle-ci.

$$\begin{aligned} \mathbf{Précision(0)} &= \frac{\mathbf{Tn}}{\mathbf{Tn + Fn}} \\ \mathbf{Précision(1)} &= \frac{\mathbf{Tp}}{\mathbf{Tp + Fp}} \end{aligned} \quad (1-20)$$

1.3.2.2.1 La précision balancée (Balanced accuracy – BA)

La précision balancée est la capacité du modèle à prédire correctement les deux classes. C'est aussi la moyenne des rappels des deux classes.

$$\mathbf{BA} = \frac{\mathbf{Rappel(0)} + \mathbf{Rappel(1)}}{2} \quad (1-21)$$

Il faut noter que dans ce travail nous avons utilisé que deux classes, mais ces paramètres statistiques peuvent être étendus à plus de deux classes.

1.3.2.3 La Courbe ROC (Receiving Operating Characteristics)

La courbe ROC est un outil d'évaluation et de comparaison des modèles. Elle est définie pour les problèmes à deux classes (les positifs et les négatifs), quand un modèle permet de prioriser un exemple sur un autre – dire qu'une molécule a plus des chances d'être active qu'une autre. Dans ce cas, à chaque valeur possible pour la fonction de décision du modèle correspond un classeur caractéristique pour une valeur de rappel et une valeur de précision.

La courbe ROC met en relation dans un graphique les taux de faux positifs (en abscisse) et les taux de vrais positifs (en ordonnée). Elle indique la capacité du classeur à placer les positifs devant les négatifs.

Sa construction s'appuie donc sur les probabilités d'être positif fournies par les classificateurs.

Il est possible de calculer un indicateur à partir de la courbe ROC, il s'agit de l'AUC (Area Under Curve – Aire Sous la Courbe) qui indique la probabilité d'un individu positif d'être classé devant un individu négatif (dans le meilleur des cas AUC = 1). Si les individus sont classés au hasard, l'AUC sera égal à 0.5, symbolisée par la diagonale principale dans la Figure 1-9.

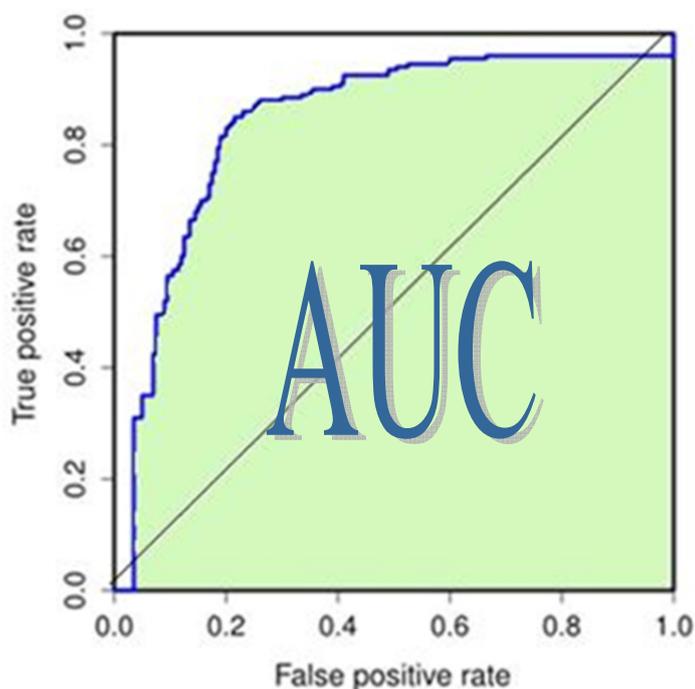


Figure 1-9 Courbe ROC. La diagonale correspond aux individus classés au hasard.

1.4 Validation des modèles QSPR

Afin de déterminer la capacité prédictive d'un modèle des procédures de validation croisée sont couramment utilisées[21]. A partir d'un jeu initial de molécules est divisé plusieurs fois en deux sous-parties : un jeu d'apprentissage et un jeu de test externe. Uniquement les molécules du jeu d'apprentissage sont utilisées pour le développement du modèle qui est testé ensuite sur le jeu de test: on parle de validation externe et on retrouve 3 méthodes principales décrite ci-après.

1.4.1 Leave One Out

La procédure « Leave-One-Out » retire successivement une molécule du jeu d'apprentissage contenant m molécules. Un modèle QSPR est construit sur un ensemble $m-1$ de composés et la molécule retirée est prédite par le modèle. Cette procédure est répétée m fois afin de prédire les propriétés de toutes les molécules.

1.4.2 Validation croisée par n-paquets (N-fold Cross Validation)

La procédure « N-Fold Cross Validation » correspond à un découpage en n parties du jeu de données. A tour de rôle, une partie du jeu de données est attribuée pour un jeu de test externe, les autres constituant le jeu d'entraînement (cf. Figure 1-10).

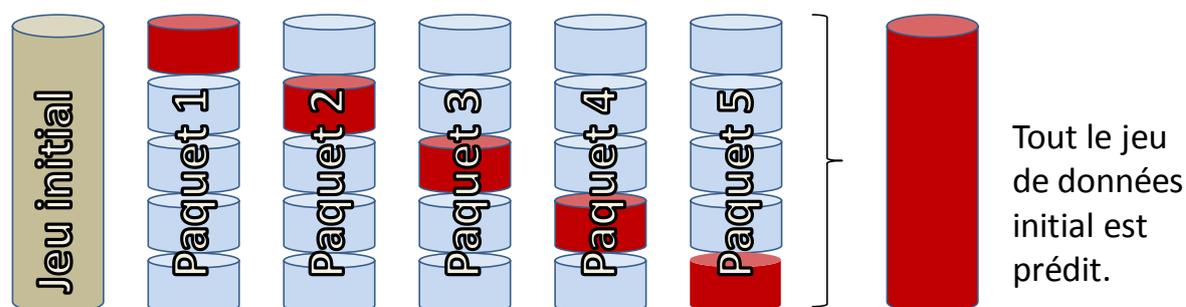


Figure 1-10 Procédure validation croisée (5-paquets)

Les critères statistiques (R^2 , MAE et RMSE) calculés à partir des résultats de prédictions sont généralement moins bons, comparés à ceux obtenus avec une LOO, ces dernières ayant tendance à surestimer les capacités prédictives du modèle. En fait la LOO est un cas particulier de la validation croisée. Il est recommandé d'utiliser $n=2$ pour avoir le moins biais et de réitérer le partage en paquets.

1.4.3 Procédure de Y-randomisation

Afin de s'assurer qu'un modèle QSPR est fiable, les tests de Y-randomization[21] sont une des techniques les plus employées. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « *chance correlation* »), c'est-à-dire un modèle affichant des bons résultats statistiques (R^2 , MAE) pour l'apprentissage, mais impliquant des descripteurs qui dans la réalité ne sont pas reliés à la propriété modélisée. Ces modèles aléatoires peuvent être détectés par la procédure Y-randomization. Elle consiste à mélanger aléatoirement les propriétés expérimentales pour le jeu d'apprentissage et en, en utilisant les mêmes descripteurs, à entraîner à nouveau l'algorithme d'apprentissage pour tenter d'obtenir un modèle. Normalement, les modèles obtenus doivent avoir des performances très faibles. La distribution des modèles obtenus permettent de fixer un seuil heuristique de signification des modèles. Ainsi, on peut choisir les modèles qui ont au plus de 1% de chances d'être confondus avec un modèle fortuit.

1.5 Domaine d'Applicabilité (DA) des modèles

Un modèle QSPR ne peut pas être considéré comme un modèle universel, parce qu'il est développé sur un nombre limité de composés qui ne couvrent pas tout l'espace chimique. Par conséquent la propriété prédite d'un composé, chimiquement dissimilaire au jeu d'apprentissage, ne pourra pas être considérée fiable.

Le domaine d'applicabilité (DA) va nous permettre de définir la zone dans laquelle un composé pourra être prédit avec confiance. Le DA correspond donc à la région de l'espace chimique incluant les composés du jeu d'apprentissage et les composés similaires, proches dans ce même espace

Au cours ce travail, plusieurs domaines d'applicabilité ont été utilisés, dont les quatre plus importants sont décrits.

1.5.1 Contrôle des fragments

Le contrôle des fragments permet de tester si une molécule se trouve dans le DA d'un modèle par la présence de fragments. Si une nouvelle molécule contient des nouveaux fragments, elle sera considérée en dehors du DA du modèle. Si la nouvelle molécule ne présente aucun nouveau fragment, la molécule est considérée dans le DA du modèle et la prédiction sera considérée comme fiable.

Une dérivation de cet DA utilisé au cours de cette thèse est le **contrôle de fragments IVAB (FrgCtrl IVAB)**. Un mélange est considéré comme faisant partie du cet DA si chacun de ces composés n'a pas de nouveaux fragments de type IVAB(n_{\min} - n_{\max}) par rapport au composés de jeu d'entraînement, qui sont fragmentés de la même façon. Indifféremment de modèle, donc de fragmentation, seule la fragmentation IVAB va dicter l'appartenance au DA.

1.5.2 La méthode Min-Max ou « Bounding Box »

Cette méthode utilise les valeurs minimales et maximales $D=[d_{\min}, d_{\max}]$ pour chaque descripteur. Pour une nouvelle molécule les valeurs de ses descripteurs x_i sont comparées à cet intervalle D .

- Si $x_i \notin [d_{\min}-d_{\max}]$ le composé est considéré hors de DA du modèle et la prédiction du nouveau composé sera considérée comme peu fiable.
- Si $x_i \in [d_{\min}-d_{\max}]$ le composé est considéré dans le DA du modèle et la prédiction du nouveau composé sera considérée comme fiable.

1.5.3 z-kNN

Cette méthode calcule, pour un nouveau composé, la distance moyenne D entre les k plus proches molécules du modèle. Si l'inégalité (1-22) est respectée, la molécule est considérée comme étant dans le DA du modèle.

$$D \leq d + Z \cdot s_d \quad (1-22)$$

où Z est un paramètre empirique, d correspond à la distance euclidienne moyenne des molécules dans l'espace des descripteurs du modèle, s_d est la déviation standard des distances euclidiennes séparant chaque composé du jeu d'entraînement avec ses k plus proches voisins. En général $k=2$ et $Z=0.5$.

1.5.4 1-SVM

La méthode 1-SVM associe le domaine d'applicabilité des modèles QSPR avec la zone de l'espace de descripteurs d'entrée où la densité de données d'entraînement dépasse un certain seuil. La principale hypothèse de cette procédure est que la performance de prédiction des modèles tend à être plus élevée pour les composés à tester à l'intérieur des zones à forte densité que pour ceux qui sont à l'extérieur. Ceci peut être vrai, car dans la zone de faible densité tous les composés à tester sont loin des composés d'entraînement ce qui rend l'interpolation des propriétés de jeu d'entraînement au jeu de test peu fiables.

Au lieu de chercher une surface de décision de séparation des zones de haute (C2, Figure 1-11) et faible densité (C1, Figure 1-11) dans l'espace d'entrée, la méthode de classification à une classe (1-SVM) recherche un hyperplan dans l'espace caractéristique qui est associé au noyau RBF.

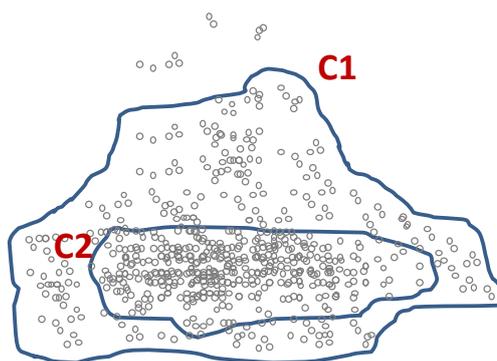


Figure 1-11 Exemplification de la méthode 1-SVM

L'utilisation de la méthode 1-SVM pour évaluer le domaine d'applicabilité des modèles a été récemment suggéré par Baskin [22].

1.6 Conclusion

Dans cette partie, le processus de construction d'un modèle QSPR a été détaillé. A partir d'un jeu de données de descripteurs, un modèle QSPR peut être développés

utilisant différentes méthode d'apprentissage. Ces méthodes permettent de faire le lien entre les descripteurs moléculaires et la propriété étudiée. Plusieurs méthodes d'apprentissage ont été décrites, comme la MLR, le réseau de neurones, la SVM, la PLS ou le Perceptron. Une fois le modèle obtenu, il est validé par validation externe (LOO ou n-Fold) en calculant des paramètres statistiques. Afin de prédire des nouvelles molécules, leur appartenance au domaine d'applicabilité du modèle est testée afin de déterminer la fiabilité des prédictions fournis.

1.7 Références

1. Todeschini, R. and V. Consonni, in *Molecular Descriptors for Chemoinformatics*. 2009, Wiley-VCH Verlag GmbH & Co. KGaA.
2. Navajas, C., et al., *Comparative Molecular Field Analysis (CoMFA) of MX Compounds using different Semi-empirical Methods: LUMO Field and its Correlation with Mutagenic Activity*. Quantitative Structure-Activity Relationships, 1996. **15**(3): p. 189-193.
3. von Itzstein, M., et al., *Rational design of potent sialidase-based inhibitors of influenza virus replication*. Nature, 1993. **363**(6428): p. 418-423.
4. Varnek, A. and V.P. Solov'ev, *"In Silico" Design of Potential Anti-HIV Actives Using Fragment Descriptors*. Combinat. Chem. High Throughput Screening, 2005. **8**(5): p. 403-416.
5. Varnek, A., et al., *ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors*. Current Computer-Aided Drug Design, 2008. **4**(3): p. 191-198.
6. Solov'ev, V.P., A. Varnek, and G. Wipff, *Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments*. J. Chem. Inf. Comput. Sci., 2000. **40**(3): p. 847-858.
7. Ruggiu, F., et al., *ISIDA Property-Labelled Fragment Descriptors*. Molecular Informatics, 2010. **29**(12): p. 855-868.
8. *Molecular Operating Environment (MOE)*. 2009, Chemical Computing Group Inc: Montreal, Quebec, Canada.
9. Varnek, A., et al., *ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors*. Cur. Computer-Aided Drug Design, 2008. **4**(3): p. 191-198.
10. Horvath, D., et al., *Stochastic versus stepwise strategies for quantitative structure - Activity relationship generation - How much effort may the mining for successful QSAR models take?* Journal of Chemical Information and Modeling, 2007. **47**(3): p. 927-939.
11. Solov'ev, V.P. and A. Varnek, *Structure-Property Modeling of Metal Binders Using Molecular Fragments*. Rus. Chem. Bull., 2004. **53**(7): p. 1434-1445.
12. Varnek, A., et al., *Exhaustive QSPR studies of a large diverse set of ionic liquids: How accurately can we predict melting points?* Journal of Chemical Information and Modeling, 2007. **47**(3): p. 1111-1122.
13. Boulesteix, A.L. and K. Strimmer, *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data*. Briefings in Bioinformatics, 2007. **8**(1): p. 32-44.
14. Artemenko, N.V., et al., *Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds*. Russian Chemical Bulletin, 2003. **52**(1): p. 20-29.
15. Tetko, I.V., *Neural Network Studies. 4. Introduction to Associative Neural Networks*. J. Chem. Inf. Comp. Sci., 2002. **42**(3): p. 717-728.
16. Vapnik, V.N., *Statistical Learning Theory*. 1998, New York: John Wiley & Sons.
17. Muller, K.R., et al., *An introduction to kernel-based learning algorithms*. IEEE Trans Neural Netw, 2001. **12**(2): p. 181-201.
18. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines*. ACM Transactions on Intelligent Systems and Technology, 2011. **2**(3): p. 1-27.
19. Breiman, L., *Random forests*. Machine Learning, 2001. **45**(1): p. 5-32.

20. Seewald, A.K., *How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness*, in *Proceedings of the Nineteenth International Conference on Machine Learning*. 2002, Morgan Kaufmann Publishers Inc. p. 554-561.
21. Tropsha, A., P. Gramatica, and V.K. Gombar, *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*. *QSAR Comb. Sci.*, 2003. **22**(1): p. 69-77.
22. Baskin, I.I., N. Kireeva, and A. Varnek, *The One-Class Classification Approach to Data Description and to Models Applicability Domain*. *Molecular Informatics*, 2010. **29**(8-9): p. 581-587.

2 Prédiction des propriétés de mélanges de liquides binaires

Tout l'enjeu de ce travail est l'application des méthodologies de type QSPR à la prédiction de propriétés physiques de mélanges de liquides binaires. Il convient dans un premier temps de présenter les méthodes classiquement utilisées par les ingénieurs dans le domaine du génie des procédés.

Les propriétés d'intérêts pour ce travail concernent les équilibres de phases et plus particulièrement les Équilibres Liquide-Vapeur (VLE). Lorsqu'un ingénieur a des données VLE spécifiques pour le système binaire qu'il étudie, il peut utiliser des modèles de régression basés soit sur une approche d'équation d'état, soit sur une approche de modèle de solution. Lorsqu'aucune donnée expérimentale n'est disponible, des outils prédictifs existent. Deux modèles basés sur des fondements théoriques sont présentés dans ce travail : UNIFAC et COSMO-RS. Le lecteur pourra se référer à l'ouvrage de Prausnitz[1] et le cours de thermodynamique de Schwartzentruber [2] pour plus de détail.

Enfin, la littérature présente déjà des travaux sur l'utilisation de modèles QSPR pour des mélanges. Une revue de ces travaux est présentée en fin de chapitre.

2.1 Les mélanges: définitions

2.1.1 Changement d'état

Les différents termes sont donnés sur le diagramme ci-joint (Figure 2-1).

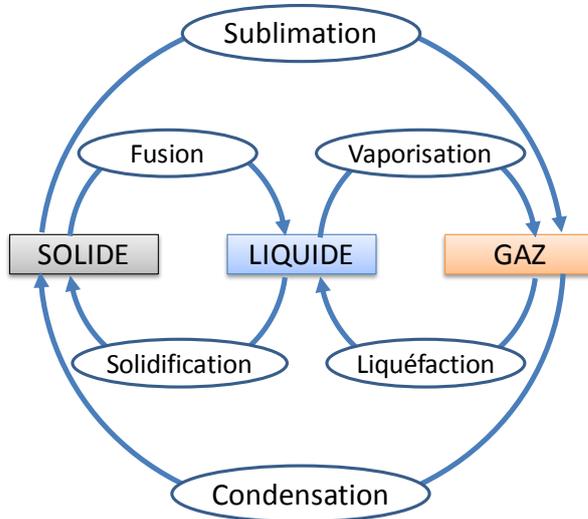


Figure 2-1 Les différents changements d'état.

2.1.2 Mélange binaire liquide – vapeur

Un mélange binaire 1-2 est constitué de deux corps purs 1 et 2. Cette notion n'implique aucun état de la matière pour 1 et 2 qui peuvent exister:

- seulement à l'état de liquide ou de vapeur
- simultanément à l'état de liquide et de vapeur

2.1.3 Pression partielle d'un gaz

La pression partielle du gaz i est la contribution du gaz i à la pression d'équilibre du mélange. On la note P_i . Ainsi, la pression totale d'un mélange binaire s'exprime comme suit :

$$P = P_1 + P_2 \quad (2-1)$$

2.1.4 Fraction molaire

Considérons un système composée de deux entités, 1 et 2, avec n_1 et n_2 les quantités de matière de 1 et 2 en phase liquide, n'_1 et n'_2 les quantités de matière de 1 et 2 en phases gazeuse. Les fractions molaires s'expriment comme suit :

$$x_1 = \frac{n_1}{n_1 + n_2} \text{ et } x_2 = 1 - x_1$$

$$y_1 = \frac{n'_1}{n'_1 + n'_2} \text{ et } y_2 = 1 - y_1 \quad (2-2)$$

2.1.5 Variables intensives et extensives

Une variable est intensive lorsqu'elle ne dépend pas de la taille des parties choisies dans le système considéré. Une variable est extensive lorsqu'elle est proportionnelle au volume de matière considéré. Quelques exemples:

- variables extensives: volume, masse, nombre de moles
- variables intensives: pression (totale, partielle), température, fraction (molaire, massique), concentration (molaire, massique)

2.1.6 Équilibre d'un système

A l'équilibre, toutes les variables intensives décrivant le système sont constantes dans le temps.

2.1.7 Phases, corps pur, mélange

Une phase est une région de l'espace où les variables intensives ont des valeurs indépendantes des points considérés (identité des propriétés physiques et chimiques).

Une phase peut être constituée d'un seul corps pur ou être un mélange homogène de plusieurs corps.

Un seul corps pur peut se trouver simultanément sous la forme de plusieurs phases qui coexistent : de l'eau liquide peut coexister avec de l'eau vapeur ou avec de la glace.

Deux corps purs peuvent sembler "mélangés", mais constituer en réalité deux phases distinctes : c'est le cas par exemple lorsqu'on mélange de la poudre de zinc et de la poudre d'aluminium (on a deux types de grains de solide, distincts). On peut obtenir une seule phase homogène par fusion de l'ensemble (un liquide).

2.1.8 Variance

La variance d'un système à l'équilibre est le nombre minimum de variables intensives indépendantes que l'expérimentateur utilise pour décrire totalement un

état d'équilibre de ce système. Les autres variables intensives s'obtiennent à partir de celles fixées par l'expérimentateur.

2.1.9 Règle des phases

Elle fixe la variance v d'un système.

$$v = c + 2 - \varphi - k - r$$

c : nombre de corps purs; φ : nombre de phases; k : nombre de relations imposées par les équilibres chimiques; r : nombre de relations imposées par l'expérimentateur.

Dans le cas d'un mélange binaire en équilibre liquide-vapeur à T et P fixés, les paramètres prennent les valeurs suivantes :

$$c = 2 ; \varphi = 2 \text{ (liquide+vapeur)} ; k = 0 \text{ (pas de réaction chimique).}$$

$v = 2$, ce qui implique qu'il faut fixer deux paramètres intensifs pour définir le système.

2.1.10 Potentiel chimique d'un constituant dans une phase

Le potentiel chimique d'un constituant physico-chimique i , dans une phase est égal à la dérivée partielle de l'enthalpie libre du système par rapport à la quantité de matière N_i de ce constituant, les autres variables du système étant constantes:

$$\mu_i = \left. \frac{\partial G}{\partial N_i} \right|_{T,P,N_j,j \neq i} \quad (2-3)$$

2.2 Équilibre vapeur liquide d'un mélange binaire

Considérons un mélange binaire, c'est-à-dire formé de deux constituants, qui peuvent tous les deux se partager entre les phases liquide ou vapeur.

Habituellement, les constituants sont numérotés par ordre de volatilité décroissante. Le composé le plus volatil, à l'état de corps pur, est celui qui, à température donnée, a la pression de saturation la plus élevée ou celui qui, à pression donnée, a la température d'ébullition la plus basse.

L'équilibre liquide-vapeur d'un tel mélange peut être représenté par 4 variables intensives : la température T , la pression P , la fraction molaire du constituant 1 en phase liquide x_1 , et la fraction molaire du constituant 1 en phase vapeur y_1 .

Ces variables sont reliées par les deux équations d'équilibre, qui décrivent l'égalité des potentiels chimiques des deux constituants entre les deux phases :

$$\begin{aligned} \mu_1^L(T, P, x_1) &= \mu_1^V(T, P, y_1) \\ \mu_2^L(T, P, x_2) &= \mu_2^V(T, P, y_2) \end{aligned} \quad (2-4)$$

On remarque que le potentiel chimique du constituant 2 s'exprime en fonction de la fraction molaire du constituant 1 car les fractions molaires ne sont pas indépendantes, $x_1+x_2=1$ et $y_1+y_2=1$. Il faut alors choisir l'une des fractions molaires, par exemple celle du constituant 1, pour décrire la composition de chaque phase.

2.2.1 Mélange binaire et solution idéale

La phase liquide d'un mélange binaire est considérée comme une solution idéale lorsqu'elle suit la **loi de Raoult**, c'est-à-dire lorsque, à température fixée, **T**, on peut écrire:

$$P_1 = x_1 P_1^s(T) \quad (2-5)$$

$$P_2 = x_2 P_2^s(T) = (1 - x_1) P_2^s(T) \quad (2-6)$$

Où P_1 et P_2 sont les pressions partielles de composé 1 et 2, respectivement. $P_1^s(T)$ et $P_2^s(T)$ sont les pressions de vapeur saturante du composé 1 et 2, respectivement.

La pression de vapeur saturante est la pression à laquelle la phase gazeuse d'une substance est en équilibre avec sa phase liquide ou solide. Elle dépend exclusivement de la température.

De l'équation (2-5) et (2-6) on obtient donc :

$$P = P_1 + P_2 = P_2^s(T) + x_1(P_1^s(T) - P_2^s(T)) \quad (2-7)$$

La loi de Raoult permet de faire le lien entre les compositions de la phase liquide et de la phase vapeur.

Le modèle de la solution idéale s'applique d'autant mieux que les constituants sont des structures chimiques voisines.

Dans la phase vapeur la **loi de Dalton** peut s'appliquer si les gaz ont des comportements proches des gaz parfaits. Cela signifie que les pressions partielles des gaz sont directement proportionnelles aux fractions molaires des gaz, à la pression d'étude. Les relations qui en découlent s'écrivent:

$$P = P_1 + P_2 \text{ donc } y_1 = \frac{P_1}{P} \text{ et } y_2 = \frac{P_2}{P} \quad (2-8)$$

En combinant les lois de Dalton (2-1) et de Raoult (2-5), on peut alors écrire:

$$y_1 = x_1 \frac{P_1^s}{P} \text{ et } y_2 = x_2 \frac{P_2^s}{P} \quad (2-9)$$

La variance d'un tel système est égale à 2 comme décrit au paragraphe 2.1.8. Pour décrire parfaitement le système, il est donc nécessaire de fixer deux variables indépendantes. Les variables indépendantes possibles sont la pression totale, la température, les pressions partielles et les fractions molaires en 1 et 2 dans les deux phases. On montre facilement que si on fixe par exemple la température et la pression totale, toutes les autres variables en découlent car il existe une relation entre P et y . Ainsi, fixer la température revient à connaître P_1^s et P_2^s . Les bilans matière sur les phases conduisent à trouver la composition du composé 2 à partir de celle du composé 1. De même, la relation entre x_1 et y_1 permet d'avoir seulement à déterminer y_1 pour en déduire x_1 , P et P_1^s étant déjà connus.

De façon classiques, trois diagrammes différents sont tracés pour permettre la représentation de ces systèmes:

- **Lentille isotherme:** T fixée, on représente les variations de P en fonction de x_1 et y_1 .
- **Lentille isobare:** P fixée, on représente les variations de T en fonction de x_1 et y_1 .
- **Diagramme d'équilibre:** on représente les variations de y_1 en fonction de x_1 . Cette représentation peut être effectuée à T ou P fixée. On parlera alors de diagramme isotherme ou isobare.

2.2.1.1 Diagramme isotherme

A partir des relations plus haut on aboutit facilement aux deux relations suivantes qui permettent de tracer le diagramme :

$$P = P_2^s + x_1(P_1^s - P_2^s) \text{ et } P = \frac{P_1^s \cdot P_2^s}{P_1^s - y_1(P_1^s - P_2^s)} \quad (2-10)$$

On obtient le tracé d'une droite et d'un arc d'hyperbole (Figure 2-2). L'expérience montre que 3 zones peuvent être différenciées :

- partie supérieure: phase liquide homogène
- partie comprise entre les deux courbes: 2 phases (liquide et vapeur) présents simultanément
- partie inférieure: phase vapeur homogène

La droite est nommée **courbe de bulle** et l'arc d'hyperbole **courbe de rosée**.

On remarque que pour x_2 et y_2 égaux à 1, la pression totale est égale à la pression de vapeur saturante de 2 car on se trouve en présence d'un seul composé. Le raisonnement est le même pour 1 si x_1 et y_1 sont égaux à 1 (c'est à dire x_2 et y_2 égaux à 0).

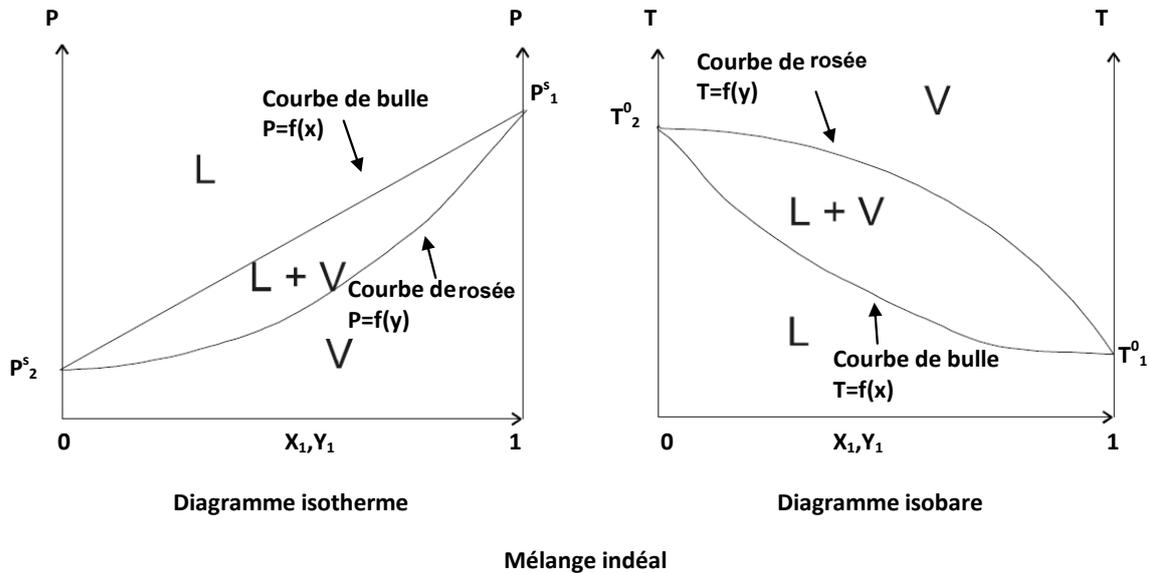


Figure 2-2 Diagrammes isotherme et isobare pour un mélange liquide binaire idéal

2.2.1.2 Diagramme isobare

Un calcul complexe permet d'aboutir à des relations entre la température et x_1 ou y_1 . On construit ainsi le diagramme isobare.

Le diagramme permet de distinguer trois zones (inversées par rapport au diagramme isotherme)(Figure 2-2):

- partie supérieure: phase vapeur homogène
- partie comprise entre les deux courbes: 2 phases (liquide et vapeur) présentes simultanément
- partie inférieure: phase liquide homogène

La courbe inférieure est nommée **courbe de bulle** (T en fonction de x_1) et la courbe supérieure est la **courbe de rosée** (T en fonction de y_1). Le terme courbe de rosée s'explique par l'expérience consistant à refroidir un mélange de vapeur de 1 et 2: la température de rosée est la température pour laquelle la première goutte de liquide apparaît.

2.2.1.3 Diagramme isobare d'équilibre liquide-vapeur ($y=f(x)$)

Ce diagramme est le plus utilisé dans le domaine de la distillation. Il passe par les points de coordonnées (0,0) et (1,1) correspondant respectivement à 1 pur et à 2 pur.

Ce diagramme s'obtient à partir du diagramme isobare. Il suffit de se placer à une température donnée et on reporte les compositions simultanées des deux phases en le composé le plus volatil (1). Chaque point correspond à une température précise.

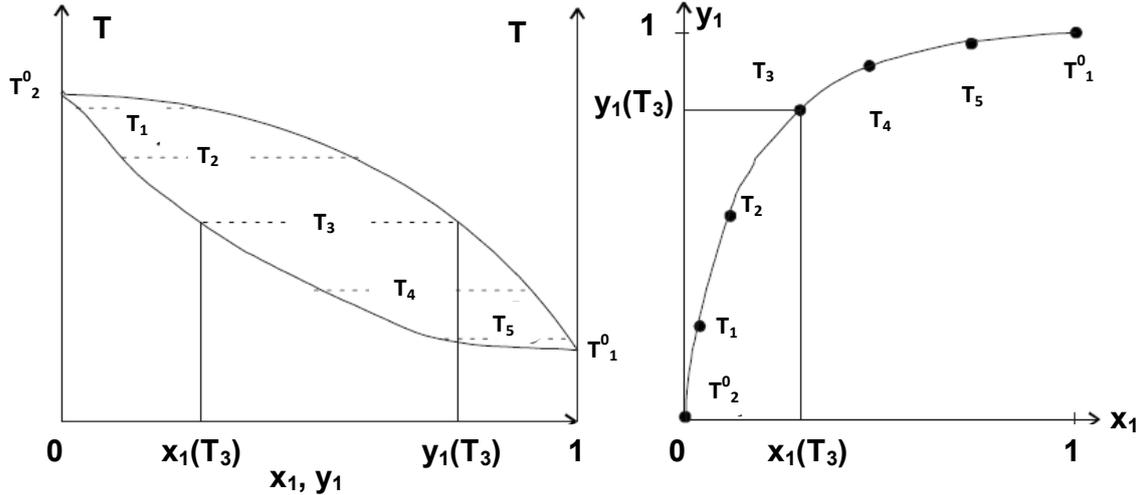


Figure 2-3 Obtention d'un diagramme isobare d'équilibre liquide-vapeur $y=f(x)$

On peut déterminer par le calcul une relation entre y_1 et x_1

Si on pose $\alpha = \frac{P_1^s}{P_2^s}$, on obtient la relation suivante:

$$y_1 = \frac{\alpha \cdot x_1}{1 + x_1 \cdot (\alpha - 1)} \tag{2-11}$$

Le tracé du diagramme est alors aussi possible à l'aide de cette relation si on suppose α constant pour toutes les températures.

2.2.2 Mélange binaire réel

Le mélange idéal n'est qu'un modèle simplifié contenant des hypothèses qui visent à minimiser les interactions entre les composés dans chacune des phases. Cela peut s'appliquer à quelques composés, souvent de structure proche et sous des pressions modérées (< 10 bar). Dans la majorité des cas les mélanges binaires s'écartent de ce modèle et ce d'autant plus que les structures chimiques sont différentes. Pour prendre en compte cette déviation à l'idéalité on introduit un **coefficient d'activité** γ_1 tel que la pression partielle du composé 1 s'écrit :

$$P_1 = \gamma_1 \cdot x_1 \cdot P_1^s \quad (2-12)$$

γ_1 dépend de x_1 et de la température et n'a pas d'unité. De même par définition, une solution idéale est celle où $\gamma_i=1$.

Suivant l'importance de l'écart existant par rapport au modèle idéal, trois types de mélanges vont exister:

- **mélanges zéotropiques:** les déviations sont faibles. Le mélange liquide est toujours miscible en toutes proportions et les compositions des phases gaz et liquides sont différentes en tout point des lentilles d'équilibres isothermes et isobares (sauf pour les corps purs).
- **mélanges homoazéotropiques:** les déviations sont fortes mais le mélange liquide reste toujours miscible en toutes proportions. Par contre, il existe au moins un point d'équilibre de la lentille où les phases liquides et vapeurs ont la même composition.
- **mélanges hétéroazéotropiques:** les déviations sont très fortes. Les deux composés n'ont qu'une faible affinité (constituants liquides partiellement miscibles) ou aucune affinité (constituants liquides toujours non miscibles). De plus, il existe au moins un point d'équilibre de la lentille où les phases liquides et vapeurs ont la même composition.

2.2.2.1 Mélanges zéotropiques

Ce sont les mélanges réels les plus proches du mélange idéal. Les allures des courbes des deux sortes de diagrammes isobares sont comparables à celles du modèle idéal. Sur le diagramme isotherme la courbe de bulle n'est plus une droite.

2.2.2.2 Mélanges homoazéotropiques

Les trois types de diagrammes sont fortement modifiés. Il existe deux types de mélanges homoazéotropiques qui sont faciles à visualiser sur les diagrammes isobares simples:

- mélanges à azéotropie positive : point d'ébullition minimum (exemple: eau/éthanol)
- mélanges à azéotropie négative : point d'ébullition maximum (exemple: toluène/éthanol)

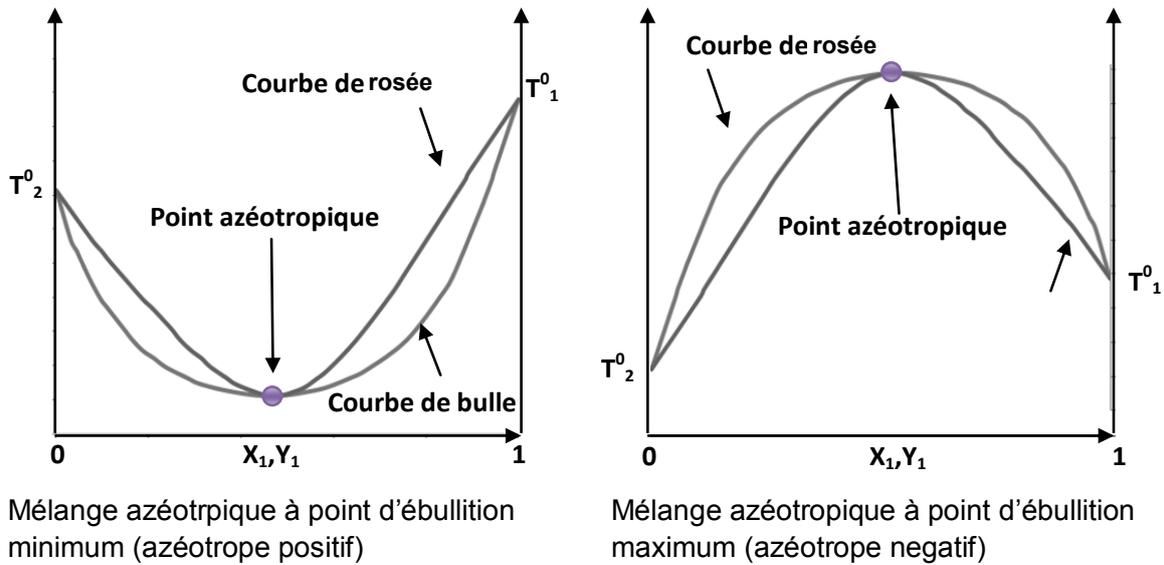


Figure 2-4 Diagrammes isobares des mélanges homoazéotropiques

Sur ces diagrammes (Figure 2-4) on note la présence d'une composition particulière nommée **point azéotrope**. Pour cette composition le mélange se vaporise à une température fixe pour une pression donnée, comme un corps pur. La vapeur émise a la même composition que le liquide. La différence apparaît par contre si on modifie la pression totale P : dans ce cas la composition de l'azéotrope est elle aussi modifiée. Le point azéotrope peut ne pas exister à certaines pressions.

2.2.2.3 Mélanges hétéroazéotropiques

Les trois types de diagrammes sont très fortement déformés par rapport à ceux du mélange idéal car les deux constituants n'ont que très peu d'affinité en phase liquide. La solubilité ne devient que partielle ou nulle. Il existe deux types de mélanges hétéroazéotropiques qui sont faciles à visualiser sur les diagrammes isobares simples:

- **Mélanges à immiscibilité totale** (exemple: eau - toluène): Un tel mélange en dessous de la température d'ébullition se présente en deux couches superposées de liquides purs (deux phases) dont l'ordre est donné par les densités respectives. Si on fixe une pression totale, la température de vaporisation d'un mélange de composition quelconque est fixe. La vapeur émise a même composition quelle que soit la composition initiale du mélange liquide: on nomme hétéroazéotrope la composition de ce mélange. Chacun des constituants se comporte comme s'il était seul dans le mélange. La pression totale de la phase vapeur est donc donnée par la relation suivante à une température T fixée:

$$P = P_{1,T}^s + P_{2,T}^s \quad (2-13)$$

- Mélanges à immiscibilité partielle** (exemple: eau – 1-butanol): Un tel mélange présente un cas intermédiaire entre un mélange homoazéotropique et un mélange hétéroazéotropique à non miscibilité totale. Dans le domaine d'immiscibilité (la courbe de bulle est une droite horizontale) les remarques faites pour le cas d'immiscibilité totale s'appliquent. En dessous de la température d'ébullition, le mélange liquide se présente comme la superposition de deux couches de liquides (deux phases) dont la composition est donnée par les limites du palier d'immiscibilité (ce ne sont pas deux couches de liquides purs). Aux extrémités des diagrammes on se trouve dans le cas où un constituant est largement majoritaire. On comprend dans ce cas que la miscibilité soit possible. Dans ces deux domaines les températures de vaporisation ne sont pas fixes.

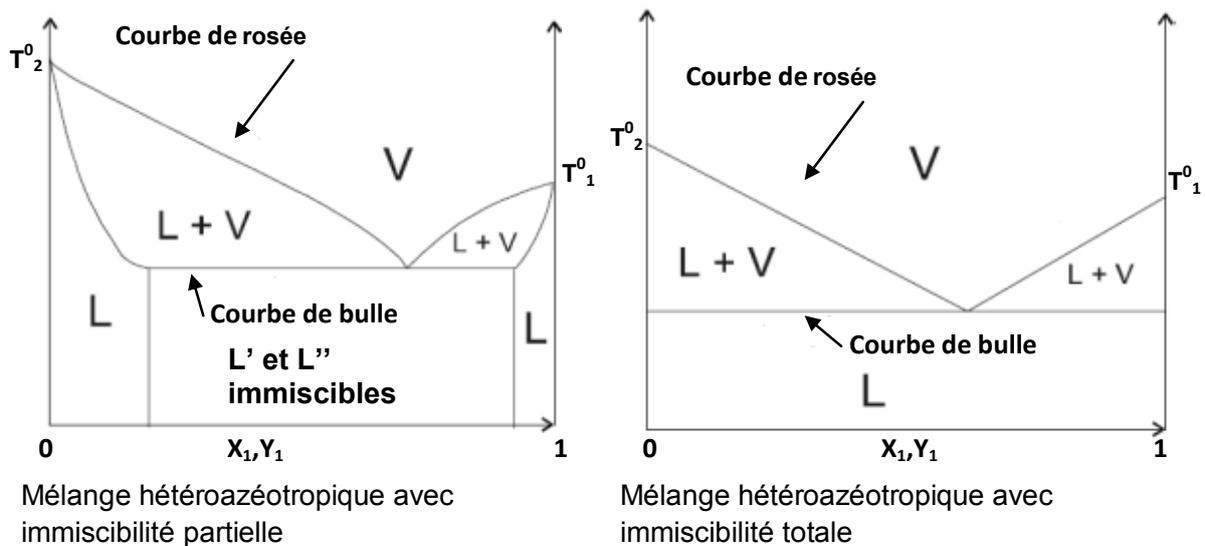


Figure 2-5 Mélanges hétéroazéotropiques

2.2.3 Coefficient d'activité et enthalpie libre d'excès

Dans les mélanges réels (non-idéal) la loi de Raoult ne s'applique plus en l'état, mais il faut introduire un **coefficient d'activité** pour expliquer les interactions entre les composés dans la phase liquide.

Il existe une notion thermodynamique qui s'avère un outil pratique pour corrélérer les coefficients d'activités entre eux : **l'énergie libre d'excès de Gibbs, G^E** . Le mot "excès" est introduit pour traduire les déviations par rapport à un modèle de solution

idéale. L'énergie d'excès de Gibbs totale (G^E) pour une solution binaire, contenant n_1 moles de composé 1 et n_2 moles de composé 2, est définie par:

$$G^E = RT(n_1 \ln \gamma_1 + n_2 \ln \gamma_2) \quad (2-14)$$

Ainsi, dans le cas d'une solution idéale, $\gamma_1 = \gamma_2 = 1$ et $G^E=0$.

L'équation (2-14) exprime G^E en fonction de γ_1 et γ_2 simultanément. Pour pouvoir les exprimer indépendamment, il faut introduire **l'équation de Gibbs-Duhem** :

$$x_1 \frac{\partial \ln \gamma_1}{\partial x_1} \Big|_{T,P} = x_2 \frac{\partial \ln \gamma_2}{\partial x_2} \Big|_{T,P} \quad (2-15)$$

En combinant l'équation (2-14) et l'équation (2-15), les coefficients d'activité peuvent être exprimés indépendamment :

$$\begin{aligned} RT \ln \gamma_1 &= \frac{\partial G^E}{\partial n_1} \Big|_{T,P,n_2} \\ RT \ln \gamma_2 &= \frac{\partial G^E}{\partial n_2} \Big|_{T,P,n_1} \end{aligned} \quad (2-16)$$

Afin d'obtenir une expression dépendante de la composition plutôt que du nombre de mole, on utilise **l'énergie libre molaire de Gibbs** : $g^E = G^E/(n_1+n_2)$.

L'équation (2-14) devient, alors:

$$g^E = RT(x_1 \ln \gamma_1 + x_2 \ln \gamma_2) \quad (2-17)$$

2.2.4 Expressions classiques de l'énergie libre d'excès

Pour un mélange binaire, l'énergie libre d'excès est, donc, une fonction de x_1 , définie sur $[0,1]$ et nulle aux deux extrémités de ce domaine. La fonction la plus simple qui réponde à cette condition est :

$$\frac{g^E}{RT} = Ax_1x_2 \quad (2-18)$$

A est un paramètre ajustable sur des données expérimentales (en particulier d'équilibre liquide-vapeur) et représente l'existence d'interactions spécifiques entre les molécules qui forment la solution. Cette expression a été proposée par **Margules**[3].

Par différentiation des équations (2-16) on obtient:

$$\ln \gamma_1 = Ax_2^2$$

$$\ln \gamma_2 = Ax_1^2 \tag{2-19}$$

Plus souvent des expressions à deux paramètres sont utilisées comme par exemple l'équation de Margules à deux paramètres ou l'équation de Van Laar (voir Tableau 2-1). Ces expressions sont totalement empiriques, et elles ne s'appliquent qu'à des mélanges binaires : leur généralisation aux mélanges multi-constituants est très malaisée.

Tableau 2-1 Expressions de l'énergie libre d'excès. Tableau adapté de l'ouvrage de Prausnitz [1]

Nom	$\frac{g^E}{RT}$	Paramètres	$\ln \gamma_1$ et $\ln \gamma_2$
Margules à un paramètre	$\frac{g^E}{RT} = Ax_1x_2$	A	$\ln \gamma_1 = Ax_2^2$ $\ln \gamma_2 = Ax_1^2$
Margules à deux paramètres	$\frac{g^E}{RT} = x_1x_2[A + B(x_1 - x_2)]$	A, B	$\ln \gamma_1 = (A + 3B)x_2^2 - 4Bx_2^3$ $\ln \gamma_2 = (A - 3B)x_1^2 + 4Bx_1^3$
van Laar	$\frac{g^E}{RT} = \frac{Ax_1x_2}{x_1\left(\frac{A}{B}\right) + x_2}$	A, B	$\ln \gamma_1 = A \left(1 + \frac{A x_1}{B x_2}\right)^{-2}$ $\ln \gamma_2 = B \left(1 + \frac{B x_2}{A x_1}\right)^{-2}$
Wilson	$\frac{g^E}{RT} = -x_1 \ln(x_1 + \Lambda_{12}x_2) - x_2 \ln(x_2 + \Lambda_{21}x_1)$	$\Lambda_{12}, \Lambda_{21}$	$\ln \gamma_1 = -\ln(x_1 + \Lambda_{12}x_2)$ $+ x_2 \left(\frac{\Lambda_{12}}{x_1 + \Lambda_{12}x_2} - \frac{\Lambda_{21}}{\Lambda_{21}x_1 + x_2}\right)$ $\ln \gamma_2 = -\ln(x_2 + \Lambda_{21}x_1)$ $- x_1 \left(\frac{\Lambda_{12}}{x_1 + \Lambda_{12}x_2} - \frac{\Lambda_{21}}{\Lambda_{21}x_1 + x_2}\right)$
NRTL (Non Random Two Liquids)	$\frac{g^E}{RT} = x_1x_2 \left(\frac{\tau_{21}G_{21}}{x_1 + x_2G_{21}} + \frac{\tau_{12}G_{12}}{x_2 + x_1G_{12}}\right)$ where $\tau_{12} = \frac{\Delta g_{12}}{RT}$ $\tau_{21} = \frac{\Delta g_{21}}{RT}$ $\ln G_{12} = -\alpha_{12}\tau_{12}$ $\ln G_{21} = -\alpha_{12}\tau_{21}$	$\Delta g_{12} = g_{12} - g_{22}$ $\Delta g_{21} = g_{21} - g_{11}$ α_{12}	$\ln \gamma_1 = x_2^2 \left[\tau_{21} \left(\frac{G_{21}}{x_1 + x_2G_{21}}\right)^2 + \frac{\tau_{12}G_{12}}{(x_2 + x_1G_{12})^2} \right]$ $\ln \gamma_2 = x_1^2 \left[\tau_{12} \left(\frac{G_{12}}{x_2 + x_1G_{12}}\right)^2 + \frac{\tau_{21}G_{21}}{(x_1 + x_2G_{21})^2} \right]$
UNIQUAC (Universal Quasi Chemical)	$g^E = g^E(\text{combinatorial}) + g^E(\text{residual})$ $\frac{g^E(\text{combinatorial})}{RT} = x_1 \ln \frac{\Phi_1}{x_1} + x_2 \ln \frac{\Phi_2}{x_2}$ $+ \frac{z}{2} \left(q_1x_1 \ln \frac{\theta_1}{\Phi_1} + q_2x_2 \ln \frac{\theta_2}{\Phi_2} \right)$ $\frac{g^E(\text{residual})}{RT} = -q_1x_1 \ln[\theta_1 + \theta_2\tau_{21}]$ $- q_2x_2 \ln[\theta_2 + \theta_1\tau_{12}]$ $\Phi_1 = \frac{x_1r_1}{x_1r_1 + x_2r_2}$ $\theta_1 = \frac{x_1q_1}{x_1q_1 + x_2q_2}$ $\ln \tau_{21} = -\frac{\Delta u_{21}}{RT}$ $\ln \tau_{12} = -\frac{\Delta u_{12}}{RT}$ <i>r_i et q_i représentent respectivement le volume et l'aire de van der Waals et z=10 est le numéro de coordination.</i>	$\Delta u_{12} = u_{12} - u_{22}$ $\Delta u_{21} = u_{21} - u_{11}$	$\ln \gamma_i = \ln \frac{\Phi_i}{x_i} + \frac{z}{2} q_i \ln \frac{\theta_i}{\Phi_i}$ $+ \Phi_j \left(l_j - \frac{r_j}{r_i} l_j \right) - q_i \ln(\theta_i + \theta_j\tau_{ji})$ $+ \theta_j q_i \left(\frac{\tau_{ji}}{\theta_i + \theta_j\tau_{ji}} - \frac{\tau_{ij}}{\theta_j + \theta_i\tau_{ij}} \right)$ where $i = 1 \quad j = 2$ or $i = 2 \quad j = 1$ $l_i = \frac{z}{2} (r_i - q_i) - (r_i - 1)$ $l_j = \frac{z}{2} (r_j - q_j) - (r_j - 1)$

2.2.5 Le concept de composition locale

Des considérations plus physiques sur la structure des mélanges liquides ont permis de proposer des expressions plus générales. Le concept de "**composition locale**" est souvent utilisé.

Ce concept exprime le fait que les molécules, à l'échelle microscopique, s'organisent en "cellules" dans lesquelles les compositions locales peuvent être différentes de la composition moyenne de la solution.

Prenons comme exemple un mélange binaire, équimolaire, dans lequel les interactions entre les molécules 1 et 2 sont "plus répulsives" que les interactions entre molécules 1 et 1 ou les interactions entre molécules 2 et 2. Les molécules formeront alors des cellules dans lesquelles les molécules de même nature ont tendance à s'agglomérer pour exclure les molécules de nature différente. Par conséquent, la composition locale de molécules 1 autour d'une molécule 1, x_{11} est plus grande que la composition globale x_1 dans le mélange. Par contre, la composition de molécules 2 autour d'une molécule 1, x_{21} est plus petite que la composition globale de molécules 2, x_2 .

Cette répartition non aléatoire des molécules dans chaque cellule est liée aux énergies d'interaction entre molécules. Dans la cellule centrée autour d'une molécule 1, g_{11} est l'énergie d'interaction entre deux molécules 1, et g_{21} l'énergie d'interaction entre une molécule 2 et une molécule 1. De même, dans la cellule centrée autour d'une molécule 2, g_{22} est l'énergie d'interaction entre deux molécules 2, et g_{12} l'énergie d'interaction entre une molécule 1 et une molécule 2. On constate alors que l'énergie $\Delta g_{12}=g_{12}-g_{11}$ diffère de l'énergie $\Delta g_{21}=g_{21}-g_{11}$.

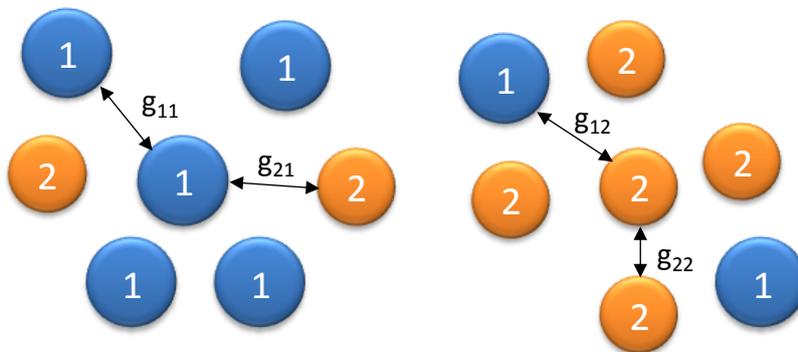


Figure 2-6 Illustration du concept de composition locale

Cette notion de "composition locale" est à la source des modèles qui se sont révélés les meilleurs pour la corrélation et la prédiction des déviations à l'idéalité : **équation de Wilson, modèles NRTL, UNIQUAC** (voir Tableau 2-1).

2.2.6 Notion de contributions de groupes

Un autre concept très utilisé pour définir le comportement de mélanges est celui de "contribution de groupes", considérant que les propriétés d'une molécule se déduisent de façon additive de celles des groupes fonctionnels qui la composent. En particulier, les interactions entre deux molécules sont supposées comme provenant des interactions deux à deux des groupes qui les composent. Ce concept est très utile parce qu'il suffit en principe de connaître les interactions entre groupes fonctionnels pour pouvoir prédire les interactions entre n'importe quelles molécules.

La plus utilisée des expressions des coefficients d'activité en contribution de groupes est l'**équation UNIFAC** (UNiversal Functional group Activity Coefficient)[4].

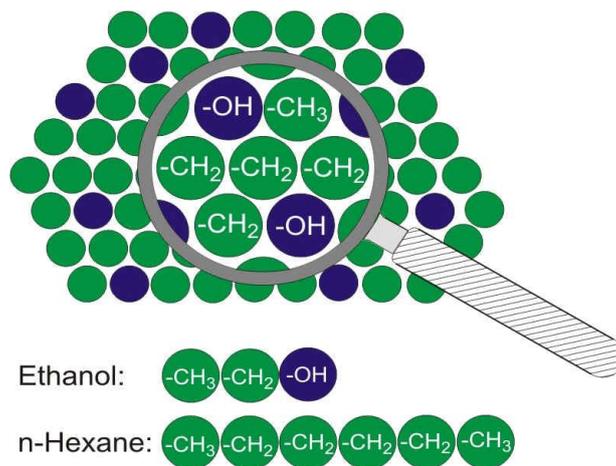


Figure 2-7 Exemplification du concept de contribution de groupes. Image reprise sur le site d'Unifac Consortium [23]

Comme UNIQUAC (cf. Tableau 2-1), l'expression des coefficients d'activité par UNIFAC présente deux termes :

- Le **terme combinatoire** a exactement la même forme que dans UNIQUAC. Simplement, les paramètres moléculaires r_i et q_i sont calculés par sommation à partir des paramètres des groupes R_k et Q_k , qui représentent le volume et l'aire d'un groupe.
- Le **terme résiduel** est calculé en considérant une solution réelle comme une « solution de groupes », obtenue en divisant chaque molécule en ses groupes fonctionnels considérés comme des espèces indépendantes en solution.

Le coefficient d'activité résiduel de l'espèce moléculaire i s'en déduit par la formule empirique :

$$\ln \gamma_i^{(res)} = \sum_{k=1}^n \nu_k^{(i)} (\ln \Gamma_k - \ln \Gamma_k^{(i)}) \quad (2-20)$$

où $\nu_k^{(i)}$ est le nombre d'occurrences du groupe k dans la molécule i , n est le nombre total de groupes et $\ln \Gamma_k^{(i)}$ le coefficient d'activité résiduel du groupe k dans la solution de groupes obtenue à partir de la molécule i pure par rapport à la $\ln \Gamma_k$ qui est le coefficient d'activité résiduel du groupe k dans la solution de groupes. Le terme est nécessaire pour assurer que le coefficient d'activité d'une espèce pure est bien égal à 1.

Les paramètres d'interaction entre groupes étant connus, on peut calculer $\ln \Gamma_k$, en utilisant l'expression:

$$\ln \Gamma_k = Q_k \left(1 - \ln \sum_m \Theta_m \Psi_{mk} - \sum_m \frac{\Theta_m \Psi_{mk}}{\sum_n \Theta_n \Psi_{nm}} \right) \quad (2-21)$$

Dans cette équation, Θ_m est la somme de la fraction surfacique du groupe m , sur tous les groupes:

$$\Theta_m = \frac{Q_m X_m}{\sum_n Q_n X_n} \quad (2-22)$$

où X_m est la fraction molaire du groupe m , c'est-à-dire le nombre de groupes m divisé par le nombre total de groupes n dans la solution. Q_m représente la surface de van der Waals du groupe m .

Ψ_{nm} est le **paramètre d'interaction de groupes** et il mesure l'énergie d'interaction entre groupes n et m . Il peut être exprimé par l'équation:

$$\Psi_{nm} = e^{\left(-\frac{U_{nm}-U_{nn}}{RT}\right)} = e^{-\frac{a_{nm}}{T}} \quad (2-23)$$

U_{nm} étant la mesure de l'énergie d'interaction entre les groupes n et m .

Le successeur d'UNIFAC appelé *UNIFAC modifié* (ou **UNIFAC Dortmund**) change l'expression du paramètre d'interaction de la forme suivante :

$$\Psi_{nm} = e^{-\frac{a_{nm}+b_{nm}T+c_{nm}T^2}{T}} \quad (2-24)$$

Les paramètres nécessaires pour décrire un mélange sont, alors, les paramètres d'interaction a_{mn} et a_{nm} pour l'UNIFAC original ou bien a_{mn} , a_{nm} , b_{nm} , b_{mn} , c_{nm} et c_{mn} pour UNIFAC Dortmund. Afin de compléter la matrice des paramètres d'interactions entre groupes principaux un grand travail de régression de données d'équilibre liquide-vapeur a été entrepris. Des mises à jour et des extensions sont publiées périodiquement [5, 6].

L'équation UNIFAC est utilisée pour représenter l'équilibre liquide vapeur de nombreux mélanges. Son avantage principal est de permettre la prédiction du comportement de systèmes pour lesquels il n'existe pas de données expérimentales, avec, en général, un assez bon degré de fiabilité.

2.2.7 COSMO-RS (Conductor-like Screening Model for Real Solvents)

Les modèles de type COSMO-RS (A. Klamt [7, 8]) sont de plus en plus utilisés ces dernières années comme une alternative aux méthodes classiques cités précédemment

Le modèle utilise le résultat de calculs COSMO (COnductor-like Screening Model) dans lequel une surface est construite autour d'une molécule et un nombre important de charges électrostatiques sont placés sur cette surface. Les charges individuelles, la structure et la distribution de charges d'une molécule sont optimisées afin de trouver l'énergie minimale du système, généralement par des calculs de chimie quantique DFT. En connaissant la distribution optimale des charges sur la surface d'une molécule, une liste de segments de surface des charges peut être générée. Chaque molécule est donc réduite à un histogramme de la surface par rapport à la densité de charge de polarisation, appelée profil σ .

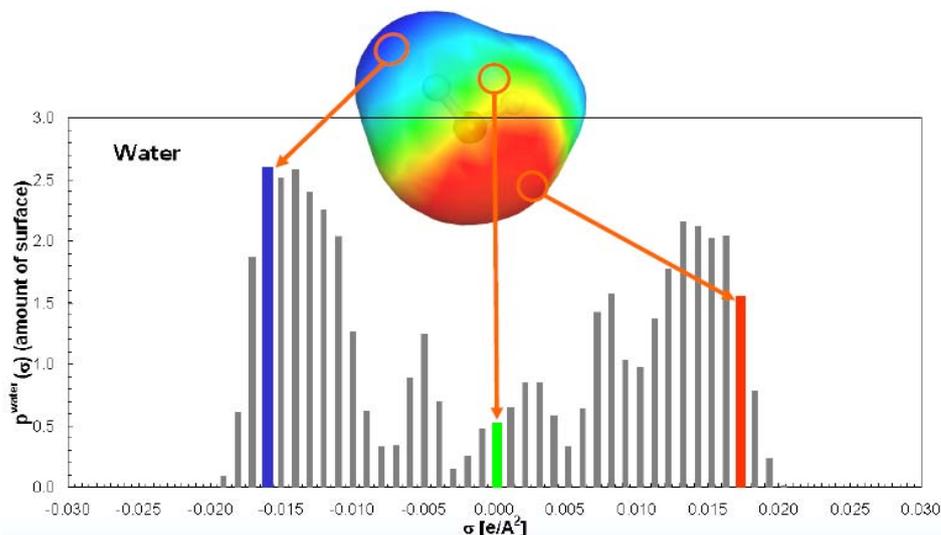


Figure 2-8 Profil σ pour la molécule d'eau. Image reprise du cours d'Andreas Klamt [8]

Après avoir réduit toutes les interactions dans le solvant aux interactions locales entre paires de segments de surface, on peut maintenant considérer l'ensemble des molécules qui interagissent comme un ensemble des interactions indépendantes entre segments de surface. Toutes les surfaces sont supposées être en contact étroit, par conséquent, seulement des interactions par paires sont possibles.

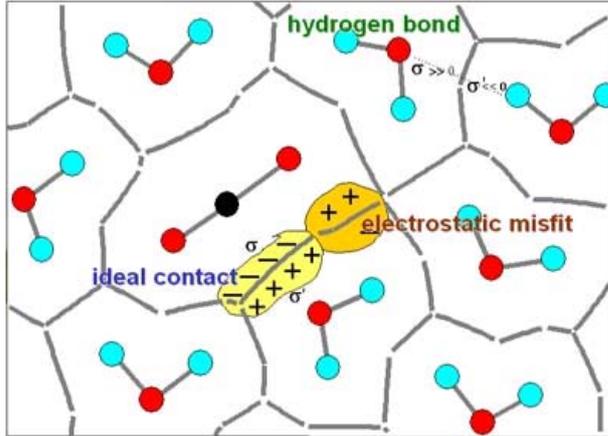


Figure 2-9 Interactions entre segments de surface. Image reprise du cours d'Andreas Klamm [8]

Si les densités de charges de polarisation σ et σ' sur les surfaces en contact ne sont pas complémentaires, une énergie d'interaction résultera de cette différence. Par conséquent, une énergie électrostatique de « disparité » (mismatch energy) décrira les interactions électrostatiques entre les segments de surface moléculaires de polarités différentes. Cette énergie avec l'énergie de liaison d'hydrogène et de van der Waals sont les composants principaux de l'énergie d'interaction.

$$E_{int} = E_{misfit} + E_{HB} + E_{vdW} \quad (2-25)$$

Le profil σ pour tout le système/mélange, $p_s(\sigma)$, est juste la somme de profils σ des composants X_i , p_{x_i} pondéré par leur fraction molaire dans le mélange (x_i).

$$p_s(\sigma) = \sum_{i \in S} x_i p_{x_i}(\sigma) \quad (2-26)$$

Ensuite, le potentiel chimique d'un segment de surface est parfaitement décrit par l'expression suivante, qui est résolue de façon itérative:

$$\mu_s(\sigma) = -kT \ln \int p_s(\sigma') e^{-\frac{E_{int}(\sigma, \sigma') - \mu_s(\sigma')}{kT}} d\sigma' \quad (2-27)$$

Ce potentiel chimique $\mu_s(\sigma)$ mesure l'affinité d'un système S pour une surface de polarité σ . D'où le potentiel chimique d'un soluté X dans un solvant S :

$$\mu_S^X = \int p_X(\sigma) \mu_S(\sigma) d\sigma + \mu_{C,S}^X \quad (2-28)$$

$\mu_{C,S}^X$ est le terme combinatoire (similaire à l'UNIFAC) qui tient compte des différences de taille et de forme des molécules dans le système.

Les potentiels chimiques μ_S^X peuvent être utilisés pour calculer toutes sortes de propriétés thermodynamiques d'équilibre: pression de vapeur, chaleur de vaporisation, énergie libre de solvation, etc. Dans le cas présent, la propriété qui nous intéresse est le coefficient d'activité qui s'exprime :

$$\gamma_S^X = e^{\frac{\mu_S^X - \mu_S^I}{kT}} \quad (2-29)$$

Le schéma ci-joint donne un aperçu de la méthode de calcul COSMO-RS.

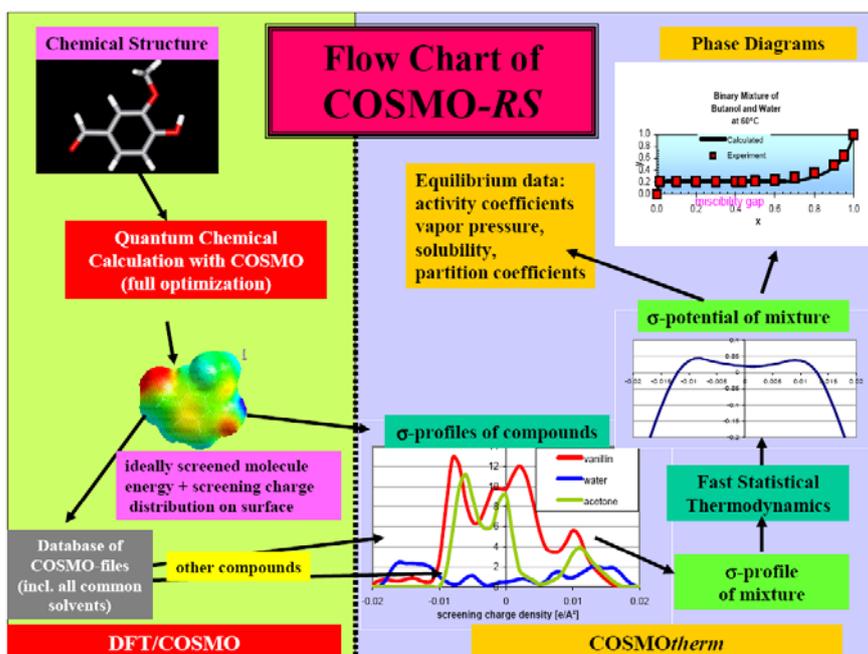


Figure 2-10 Schéma pour les calculs COSMO-RS. Image reprise du cours d'Andreas Klamt [8]

Les résultats de COSMO-RS dépendent principalement de la méthode de calcul de chimie quantique utilisé pour estimer les profils sigma. De plus, la méthode utilise une vingtaine de paramètres optimisés pour un set de 800 molécules petites et plus de 6000 points incluant des données de logP, pression de vapeur, énergie libre d'hydratation et coefficient d'activité à dilution infinie.

2.2.8 UNIFAC vs COSMO-RS

Dans les méthodes de type UNIFAC, les groupes représentent des entités uniques avec un comportement chimique spécifique alors que dans COSMO-RS la molécule se décompose en plusieurs segments qui ne diffèrent que par leur densité

molécule se décompose en plusieurs segments qui ne diffèrent que par leur densité de charge. UNIFAC et UNIFAC Dortmund nécessitent des paramètres d'interaction entre les groupes obtenus par régression, tandis que dans le cas de COSMO-RS, des interactions électrostatiques obtenus empiriquement et une contribution des liaisons hydrogène et interactions de van de Waals sont utilisées. Ces derniers ont vocation d'être plus universelle.

Les deux méthodes sont très utilisées aujourd'hui et des nombreux travaux ont été publiés concernant la prédiction de propriétés physico-chimiques des corps purs et de mélanges en utilisant des modèles UNIFAC [9-12] et COSMO-RS [13-16]. Le lecteur pourra également se référer à plusieurs publications [17-21] qui comparent ces deux méthodes.

Malgré leur utilisation habituelle, il faut souligner quelques désavantages et limitations de ces deux modèles.

Le principal inconvénient d'UNIFAC est le très grand nombre de paramètres obtenus par régression d'une quantité impressionnante de données expérimentales, ce qui implique une très grande inter-corrélation des paramètres. Ces paramètres sont recalculés annuellement avec les nouvelles données expérimentales entrées dans la base de données de Dortmund Data Bank[22] qui contient aujourd'hui environ 31600 données VLE. De plus, le nombre total des paramètres d'interactions augmente avec le carré du nombre de groupes. Aujourd'hui 69 groupes principaux sont définis pour UNIFAC et 92 groupes principaux pour UNIFAC Dortmund et la matrice de paramètres d'interaction est rempli pour un peu moins de la moitié du nombre de combinaisons du groupes[23]. Lorsque les paramètres ne sont pas remplis c'est une interaction nulle qui est prise en compte.

Un autre inconvénient d'UNIFAC est lié à la stratégie de sélection des groupes « inflexibles » qui dans la plupart des cas rend impossible à représenter un système complexe comme un ensemble de groupes. Par exemple, le modèle UNIFAC a été récemment utilisé pour les liquides ioniques[24], 12 nouveaux groupes ont été introduits et chacun d'eux représente une espèce ionique individuelle. Considérant le grand nombre de liquides ioniques existants, nous ne rendons compte que le développement de modèles UNIFAC de ces systèmes semble irréaliste. La conclusion est similaire pour le type de systèmes impliquant des molécules hétérocycliques, car chacun d'entre eux devrait être aussi représenté comme un groupe individuel.

De son côté, COSMO-RS considère que le comportement des mélanges liquides peut être calculé à partir de seulement quelques équations simples et universelles et constantes empiriques, ce qui peut conduire à des imprécisions pour des systèmes complexes. Même si les calculs de chimie quantiques sont faits une seule fois pour chaque composé individuel, ceux-ci sont chronophages et pas toujours fiables, surtout pour des molécules très flexibles, pour lesquelles trouver l'énergie minimale est difficile. Ces résultats vont toujours dépendre de la qualité de l'échantillonnage conformationnel d'un certain composé, indépendamment de la qualité des paramètres critiques.

2.3 Approches QSPR pour mélanges, développés antérieurement

Les modèles QSPR ont déjà démontré leur efficacité dans la prédiction des propriétés des composés purs.

Le développement des modèles QSPR pour des mélanges est un nouveau domaine encore peu étudié car de nombreuses difficultés se posent quant à la représentation structurale d'un mélange. Néanmoins, quelques pas ont été déjà faits dans ce sens et sont présentés par la suite.

2.3.1 Descripteurs

L'application de techniques de QSPR à des mélanges est, en principe, parfaitement possible, mais l'un des problèmes pratiques qui surgit est de caractériser le mélange en fonction de sa composition. Considérant le cas le plus simple d'un mélange de deux composants A et B dans des rapports molaires différents, une option consiste à concaténer les deux vecteurs de descripteurs de composant A et B (voir chapitre 6). Dans ce cas, les descripteurs de chaque composé individuel peuvent être pondérés par la fraction molaire de composé correspondant, dans le mélange (voir chapitre 3).

Cette représentation des mélanges a deux désavantages : premièrement, la matrice de descripteurs double en taille et deuxièmement, l'ordre des composants dans le mélange doit être définie, même si dans certains cas celui-ci n'est pas important. Par contre cette méthode est rapide et implique seulement les calculs de descripteurs pour les composés individuels. Pour des mélanges pour lesquels la distinction entre constituants est évidente, cette approche est adéquate. Prenons les exemples des complexes : ligand/récepteur pour la prédiction de la constante de dissociation[25], solvant/soluté pour l'affinité de ligand[26], l'énergie libre de

solvatation[27] ou pKa[28], ou bien les liquides ioniques anion/cation pour la prédiction du point de fusion[29] ou de la densité[30].

Dans notre publication [31] nous présentons cette approche pour des mélanges azéotropiques binaires. L'ordre de composés purs est déterminé par leur température d'ébullition : Le plus volatile est considéré comme premier composé.

Une autre option, telle que décrite par Sheridan [32] est une **approximation centroïde** d'un mélange comme la moyenne de descripteurs de toutes les molécules. L'avantage de cette approche est que chaque descripteur n'est considéré qu'une seule fois, par conséquent la matrice de descripteurs ne s'agrandit pas.

Ajmani[33] combine ces dernières deux solutions et calcule une moyenne pondérée pour chaque descripteur dans l'ensemble.

$$\mathbf{MD}=\mathbf{R}_1\mathbf{D}_1+\mathbf{R}_2\mathbf{D}_2 \quad (2-30)$$

où MD est le descripteur de mélange, R_1 et R_2 sont les fractions molaires du premier et du deuxième composant dans le mélange, respectivement, et D_1 et D_2 sont les descripteurs du premier et du deuxième composant. Dans ce travail, l'ensemble des différents descripteurs 2D et 3D décrit par Todeschini[34] a été calculée pour les composés individuels en utilisant les logiciels Dragon[35] et Cerius2[36].

Les descripteurs pour les mélanges développés dans les travaux [37, 38] reflètent diverses interactions intermoléculaires entre les composants du mélange, en tenant compte du ratio des constituants dans le mélange. Les descripteurs de mélange décrivant les liaisons hydrogène, les interactions dipôle-dipôle, et les interactions hydrophobes ont été obtenus à partir de descripteurs de composés individuels (nombre d'accepteurs et de donneurs de liaisons hydrogène, le moment dipolaire, l'énergie de désolvatation, l'aire, le volume), qui ont été combinés de façon linéaire ou non-linéaire.

L'approche utilisée pour cette thèse se base aussi sur des descripteurs additifs, comme dans les cas présentés précédemment. Les descripteurs ISIDA sont calculés pour chaque composé individuel. Ensuite les descripteurs pour le mélange sont obtenus soit par concaténation de vecteurs de descripteurs de chaque composé (cas présenté précédemment), soit par combinaison linéaire de vecteurs de descripteurs de chaque composé. Si la propriété dépend de la composition, les descripteurs sont pondérés par la fraction molaire des composés individuels correspondant. Étant donné que pour chaque propriété modélisée les descripteurs ont été obtenus différemment, la procédure de leur création est décrite en détail dans chaque

chapitre correspondant. Il faut noter que les descripteurs utilisés sont applicables aux mélanges binaires. Néanmoins, les règles de combinaison de descripteurs des corps individuels pourront être adaptées pour décrire des mélanges à plus de deux constituants (voir 9.1).

Une nouvelle approche concernant des descripteurs, les SiRMS (Simplex representation of molecular structure)[39] pour les mélanges a été utilisée par Kuz'min[40] et Muratov[41]. Cette approche inclus au niveau de la création de descripteurs l'effet de non-additivité de la propriété considérée. Les descripteurs utilisés pour décrire les mélanges contiennent des fragments formés par des atomes liés ou non-liés. Ces fragments peuvent provenir d'un seul constituant de mélange ou de plusieurs constituants de mélange. Pour les descripteurs de mélange la contribution de chaque fragment est pondérée par la fraction molaire du composé du mélange qui le contient.

Pour conclure ce qui a été présenté précédemment, deux types de descripteurs sont utilisés pour la description de mélanges : additifs et non-additifs. Jusqu'à présent aucune étude n'a montré la supériorité de l'un par rapport à l'autre. Le chapitre 3 présente une comparaison entre les descripteurs additifs (basés sur les fragments ISIDA) et non-additifs (Simplex).

2.3.2 Modèles QSPR pour des mélanges

Jusqu'à présent très peu de travail a été fait pour développer des modèles QSPR pour les mélanges. La majorité des modèles publiés ont les mêmes problèmes : la taille de jeu de données est réduite (maximum 100 mélanges), il est difficile de définir une stratégie de validation fiable, la composition n'est pas prise en compte (le ratio des deux composés est considéré 1).

Ajmani[33] développe des modèles pour prédire la déviation de la densité expérimentale d'un mélange par rapport à la densité du mélange «idéal» calculée en combinant les densités des composants individuels en fonction de leur fraction molaire dans le mélange. Le jeu de données utilisé contient 271 mélanges décrits par 4679 points. Deux manières différentes de création du jeu d'entraînement / test ont été utilisées: (i) les mêmes mélanges à différentes fractions molaires sont présents à la fois dans le jeu d'entraînement et de test (QMD-1) et (ii) certains mélanges ont été entièrement supprimés (tous les points correspondants) du jeu d'entraînement et ont

été ensuite utilisés pour tester les modèles développés (QMD-2). Une validation interne (Leave 10% Out) a été faite et des bons résultats ont été obtenus ($Q^2_{cv} > 0.9$).

Les modèles obtenus sont prédictifs ($R^2 > 0.75$), cependant, la prédictivité est limitée par les points manquants dans les mélanges de jeu d'apprentissage (QMD-1). Dans la stratégie QMD-2 le jeu de test contenant 39 mélanges sur 271 a été choisi manuellement et une seule fois, d'où la difficulté de faire confiance à ces résultats. De plus, parmi les 39 mélanges du jeu de test seulement 7 mélanges contiennent un composé pur ne se trouvant pas dans le jeu d'apprentissage ce qui surestime la performance du modèle. Malgré quelques points manquants, comme par exemple une définition du domaine d'applicabilité et une validation plus stricte, ce travail est le premier qui mérite d'être considéré. Une tentative d'interprétation mécanistique des descripteurs jugés importants qui correspond aux propriétés physico-chimiques de base responsables de la variation de la propriété à modéliser, le test de Y-randomization et la validation externe sont les atouts de cette étude.

Les auteurs ont par la suite continué leurs études des mélanges et créés leur propre méthodologie pour la description des mélanges [37, 38]. Les coefficients d'activité à dilution infinie des deux composants de mélanges binaires liquides ont été prédits à l'aide d'un ensemble de réseaux de neurones dans l'étude [38], pour un jeu de données de 411 mélanges. Ce jeu a été divisé en trois, une seule fois : jeu d'entraînement (269 mélanges), jeu de validation (45 mélanges) et jeu de test (95 mélanges).

Des modèles QSPR prédictifs et robustes pour les deux propriétés recherchées, ont été développés. Ils ont également démontré l'utilité de cinq descripteurs de mélange qui a été en outre confirmée par les résultats d'Y-randomization de jeu de test. Les résultats de la modélisation sont en total accord avec l'importance des interactions de type liaison hydrogène et dipôle-dipôle qui a confirmé la pertinence de la base mécanistique des descripteurs développés. Cependant, une analyse plus détaillée de la publication montre quelques failles : les jeux de validation et de test contiennent 6 mélanges chacun, déjà présents dans le jeu d'entraînement. Par conséquent les valeurs pour ces mélanges ne sont pas des prédictions mais de ajustements obtenues dans l'apprentissage, ce qui surestime la capacité prédictive des modèles. De plus ces jeux ont été choisis rationnellement en utilisant l'algorithme de sphère d'exclusion et un seul fois, ce qui à nouveau peut favoriser une surestimation des résultats.

La différence de performances entre le jeu de validation et celui de test peut être expliquée par la présence dans le jeu de validation de 21 sur 45 mélanges contenant au moins un composé nouveau par rapport au jeu d'entraînement ne contenant que 25 sur 97 mélanges de ce type. Les meilleurs résultats obtenus pour le jeu de prédiction par rapport au jeu d'entraînement, ce qui est rare, peuvent être expliqués par le choix biaisé des jeux de validation et de prédiction.

Malgré le manque d'une stratégie de validation rigoureuse ce travail est original et pourra servir comme point de départ pour de futures modélisations.

Les mêmes auteurs ont poursuivi leur effort en matière d'analyse QSPR de mélanges dans l'étude [37], où les descripteurs de mélange développés dans [38] ont été appliqués à l'ensemble des données décrites dans [33]. En plus de la déviation de la densité expérimentale du mélange, le volume molaire d'excès a également été modélisé.

Les auteurs corrigent les erreurs antérieures concernant la méthodologie de validation. Cette fois-ci, 50% des points sont sélectionnés pour le jeu d'apprentissage, 25% pour le jeu de validation et 25% pour le test, en fonction de la fraction molaire du premier composant. Comme dans leurs deux précédentes études [33, 38] les modèles QSPR développés pour les deux propriétés recherchées, sont prédictifs et robustes. De plus l'utilité des cinq descripteurs de mélange a été démontrée. Toutefois, la prédictivité des modèles, similaire au QMD-1 présenté dans [33] est limitée par les points manquants dans la constitution d'un mélange donné, parce que, selon le fractionnement de l'ensemble de données, des points correspondant à des ratios différents de constituants du mélange sont présents pour une part dans le jeu d'apprentissage et pour l'autre dans le jeu de test. Ainsi, chaque mélange est présent simultanément dans les deux jeux décrits par des points différents.

Pour conclure l'analyse de la série d'études QSPR rapportés dans [33, 37, 38], il faut souligner que la comparaison des résultats présentés dans [33, 37] ont montré que cinq descripteurs développés pour des mélanges étaient aussi bons que plusieurs centaines de descripteurs additifs obtenus avec Dragon et Cerius2, mais bien meilleurs que ceux-ci pour l'interprétation mécaniste. Par ailleurs, les sept descripteurs rapportés dans les études [37, 38] encodent les interactions non covalentes intermoléculaires les plus importantes et peuvent être suffisants pour la

modélisation des propriétés des mélanges binaires, où les composants sont dilués l'un dans l'autre.

Le seul travail qui modélise des d'équilibres vapeur-liquide est celui de Ravindranath et son équipe [42]. Dans ce travail Ravindranath généralise les modèles décrivant l'énergie libre d'excès de Gibbs (cf. 2.2.3) comme par exemple Margules, NRTL ou UNIQUAC (voire Tableau 2-1) en développant des modèles QSPR pour prédire leurs paramètres. Plus précisément, des modèles G^E -QSPR capables de prédire *a priori* l'équilibre vapeur-liquide (VLE), sont développés.

Le jeu de données utilisé contient 332 systèmes binaires VLE totalisant plus de 10000 points, mesurés en conditions isobares ou isothermes. Le jeu de données est repartis en 221 mélanges pour l'apprentissage et 111 pour le test.

Les paramètres de l'équation de Margules, NRTL et UNIQAC sont d'abord obtenus par régression et ensuite des modèles QSPR sont développés pour leur prédiction. Les résultats ont été comparés aux ceux fournis par le modèle d'une solution idéale, et le modèle UNIFAC.

Les modèles NRTL-QSPR sont rapportés comme les meilleures conduisant à une erreur de prédiction, RMSE=2.96K pour la température et RMSE=0.64 bar pour la pression, tandis que l'erreur de prédiction des modèles UNIFAC et de 11K pour la température et de 5.5 bar pour la pression comme Ravindranath l'indique.

Même si, ce travail est original et pourra être un premier pas important dans le développement des modèles fiables capables de prédire *a priori* l'équilibre, il faut mentionner quelques points faibles. Les auteurs présentent la liste de descripteurs CODESSA utilisés, mais à aucun moment ils n'expliquent la méthode qu'ils utilisent pour décrire un mélange. De plus, encore un fois, la stratégie de validation n'est pas rigoureuse: le jeu de validation est simplement choisi une seule fois parmi tous les mélanges formant le jeu initial. Il faut aussi souligner que pour la régression des paramètres de l'équation de Margules, NRTL ou UNIQUAC des valeurs de la pression de vapeur saturante pour les corps purs doivent être connues.

2.3.3 Conclusion

En conclusion, nous pouvons affirmer qu'aucune étude présentée antérieurement ne peut être recommandé comme un outil fiable pour analyse des mélanges.

Un gros problème aujourd'hui dans ce domaine est le manque de données, surtout publiques, néanmoins nous espérons que ce problème disparaîtra dans

quelques années, avec la croissance des bases de données. Un deuxième problème reste la description adéquate d'un mélange quelconque. Même si des descripteurs de mélange ont été développés ils ne s'appliquent généralement qu'à des mélanges binaires, ou des mélanges dont l'ordre de composés peut être déterminé facilement. Le troisième problème, qui ressort des travaux présentés précédemment, est l'absence d'une méthodologie de validation rigoureuse mais aussi l'absence d'un domaine d'applicabilité.

Les dernières tendances de l'analyse QSPR classique, à savoir, la collecte minutieuse et la compréhension des données, au travers du "nettoyage" des données, les validations interne et externe rigoureuses, va considérablement améliorer la qualité des modèles QSPR pour des mélanges.

Le QSPR est et sera très utile dans la modélisation et, surtout, pour prédire les propriétés des mélanges. L'ensemble du domaine est encore en développement et des approches développées dans cette thèse pourront être considéré comme un fondement pour la prochaine génération de descripteurs et d'études.

2.4 Références

1. Poling, B.E., J.M. Prausnitz, and J.P. O'Connell, *The properties of gases and liquids*. 5th ed. 2001, New York: McGraw-Hill.
2. Schwartzentruber, J. *Thermodynamique*. 2011 [cited 2011; Available from: http://nte.mines-albi.fr/Thermo/co/Thermo_web.html].
3. Gokcen, N.A., *Gibbs-Duhem-Margules laws*. Journal of Phase Equilibria, 1996. **17**(1): p. 50-51.
4. Fredenslund, A., R.L. Jones, and J.M. Prausnitz, *Group-Contribution Estimation of Activity-Coefficients in Nonideal Liquid-Mixtures*. Aiche Journal, 1975. **21**(6): p. 1086-1099.
5. Wittig, R., J. Lohmann, and J. Gmehling, *Vapor-Liquid Equilibria by UNIFAC Group Contribution. 6. Revision and Extension*. Industrial & Engineering Chemistry Research, 2002. **42**(1): p. 183-188.
6. Jakob, A., et al., *Further development of modified UNIFAC (Dortmund): Revision and extension 5*. Industrial & Engineering Chemistry Research, 2006. **45**(23): p. 7924-7933.
7. Klamt, A. and F. Eckert, *COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids*. Fluid Phase Equilibria, 2000. **172**(1): p. 43-72.
8. Klamt, A. *COSMO-RS: A novel tool for the prediction of industrially relevant thermodynamic data*. Available from: http://www.ensic.inpl-nancy.fr/SFGP/userfiles/file/Agenda/GT_Thermo_07-01-2010/GT-Thermo-0701-Klamt_COSMOLOGIC.pdf
9. Weidlich, U. and J. Gmehling, *A Modified Unifac Model .1. Prediction of Vle, He, and Gamma-Infinity*. Industrial & Engineering Chemistry Research, 1987. **26**(7): p. 1372-1381.
10. Gierycz, P. and B. Wisniewska, *Applicability of the Unifac Method for Prediction of Binary and Ternary Vapor-Liquid-Equilibrium Data in Systems Formed by Hydrocarbons and Alcohols*. Thermochimica Acta, 1990. **170**: p. 269-276.
11. Siimer, E. and L. Kudryavtseva, *Correlation and Prediction of Excess-Enthalpies of Ester Plus N-Alkane Systems Using the Unifac Model*. Thermochimica Acta, 1994. **240**: p. 207-213.
12. Geana, D. and V. Feriui, *Prediction of vapor-liquid equilibria at low and high pressures from UNIFAC activity coefficients at infinite dilution*. Industrial & Engineering Chemistry Research, 1998. **37**(3): p. 1173-1180.
13. Spuhl, O. and W. Arlt, *COSMO-RS predictions in chemical engineering - A study of the applicability to binary VLE*. Industrial & Engineering Chemistry Research, 2004. **43**(4): p. 852-861.
14. Guo, Z., et al., *Predictions of flavonoid solubility in ionic liquids by COSMO-RS: experimental verification, structural elucidation, and solvation characterization*. Green Chemistry, 2007. **9**(12): p. 1362-1373.
15. Roy, S., et al., *Predictions of thermodynamic properties of energetic materials using COSMO-RS*. Iccs 2010 - International Conference on Computational Science, Proceedings, 2010. **1**(1): p. 1197-1205.
16. Palomar, J., et al., *Density and molar volume predictions using COSMO-RS for ionic liquids. An approach to solvent design*. Industrial & Engineering Chemistry Research, 2007. **46**(18): p. 6041-6048.
17. Navas, A., et al., *Thermodynamic Analysis of Systems Formed by Alkyl Esters with alpha,omega-Alkyl Dibromides: New Experimental Information and the*

- Use of a Dense Database to Describe Their Behavior Using the UNIFAC Group Contribution Method and the COSMO-RS Methodology.* Industrial & Engineering Chemistry Research, 2010. **49**(24): p. 12726-12739.
18. Mokrushina, L., et al., *COSMO-RS and UNIFAC in prediction of micelle/water partition coefficients.* Industrial & Engineering Chemistry Research, 2007. **46**(20): p. 6501-6509.
 19. Buggert, M., et al., *Prediction of equilibrium partitioning of nonpolar organic solutes in water-surfactant systems by UNIFAC and COSMO-RS models.* Chemical Engineering & Technology, 2006. **29**(5): p. 567-573.
 20. Kato, R. and J. Gmehling, *Systems with ionic liquids: Measurement of VLE and gamma(infinity) data and prediction of their thermodynamic behavior using original UNIFAC, mod. UNIFAC(Do) and COSMO-RS(O1).* Journal of Chemical Thermodynamics, 2005. **37**(6): p. 603-619.
 21. Mu, T.C., J. Rarey, and J. Gmehling, *Performance of COSMO-RS with sigma profiles from different model chemistries.* Industrial & Engineering Chemistry Research, 2007. **46**(20): p. 6612-6629.
 22. Onken, U., J. Rareynies, and J. Gmehling, *The Dortmund Data-Bank - a Computerized System for Retrieval, Correlation, and Prediction of Thermodynamic Properties of Mixtures.* International Journal of Thermophysics, 1989. **10**(3): p. 739-747.
 23. *The UNIFAC Consortium.* Available from: <http://unifac.ddbst.de>.
 24. Lei, Z.G., et al., *UNIFAC Model for Ionic Liquids.* Industrial & Engineering Chemistry Research, 2009. **48**(5): p. 2697-2704.
 25. Bock, J.R. and D.A. Gough, *Virtual screen for ligands of orphan G protein-coupled receptors.* Journal of Chemical Information and Modeling, 2005. **45**(5): p. 1402-14.
 26. Jover, J., R. Bosque, and J. Sales, *Quantitative Structure-Property Relationship Estimation of Cation Binding Affinity of the Common Amino Acids.* Journal of Physical Chemistry A, 2009. **113**(15): p. 3703-3708.
 27. Kravtsov, A.A., et al., *"Bimolecular" QSPR: Estimation of the solvation free energy of organic molecules in different solvents.* Doklady Chemistry, 2007. **414**: p. 128-131.
 28. Jover, J., R. Bosque, and J. Sales, *Neural network based QSPR study for predicting pK(a) of phenols in different solvents.* QSAR & Combinatorial Science, 2007. **26**(3): p. 385-397.
 29. Billard, I., et al., *In Silico Design of New Ionic Liquids Based on Quantitative Structure-Property Relationship Models of Ionic Liquid Viscosity.* Journal of Physical Chemistry B, 2011. **115**(1): p. 93-98.
 30. Lazzus, J.A., *rho(T, p) model for ionic liquids based on quantitative structure-property relationship calculations.* Journal of Physical Organic Chemistry, 2009. **22**(12): p. 1193-1197.
 31. Solov'ev, V.P., et al., *Quantitative Structure-Property Relationship (QSPR) Modeling of Normal Boiling Point Temperature and Composition of Binary Azeotropes.* Industrial & Engineering Chemistry Research, 2011. **50**(24): p. 14162-14167.
 32. Sheridan, R.P., *The centroid approximation for mixtures: Calculating similarity and deriving structure-activity relationships.* Journal of Chemical Information and Computer Sciences, 2000. **40**(6): p. 1456-1469.
 33. Ajmani, S., et al., *Application of QSPR to mixtures.* Journal of Chemical Information and Modeling, 2006. **46**(5): p. 2043-2055.

34. Todeschini, R. and V. Consonni, *Handbook of Molecular Descriptors*. 2000, Weinheim, New York, Chichester, Brisbane, Singapore, Toronto: Wiley-VCH.
35. Mauri, A., et al., *Dragon software: An easy approach to molecular descriptor calculations*. Match-Communications in Mathematical and in Computer Chemistry, 2006. **56**(2): p. 237-248.
36. *Cerius2 software*. Version 4.8; Available from: <http://www.accelrys.com/>.
37. Ajmani, S., et al., *Characterization of Mixtures. Part 2: QSPR Models for Prediction of Excess Molar Volume and Liquid Density Using Neural Networks*. Molecular Informatics, 2010. **29**(8-9): p. 645-653.
38. Ajmani, S., et al., *Characterization of Mixtures Part 1: Prediction of Infinite-Dilution Activity Coefficients Using Neural Network-Based QSPR Models*. QSAR & Combinatorial Science, 2008. **27**(11-12): p. 1346-1361.
39. Muratov, E.N., V.E. Kuz'min, and A.G. Artemenko, *COMP 251-Hierarchical QSAR technology on the base of simplex representation of molecular structure*. Abstracts of Papers of the American Chemical Society, 2007. **234**.
40. Kuz'min, V.E., et al., *Consensus QSAR Modeling of Phosphor-Containing Chiral ACNE Inhibitors*. Qsar & Combinatorial Science, 2009. **28**(6-7): p. 664-677.
41. Muratov, E.N., et al., *QSAR Analysis of Poliovirus Inhibition By Dual Combinations of Antivirals*. Antiviral Research, 2010. **86**(1): p. A62-A62.
42. Ravindranath, D., et al., *QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior*. Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

DEUXIEME PARTIE:

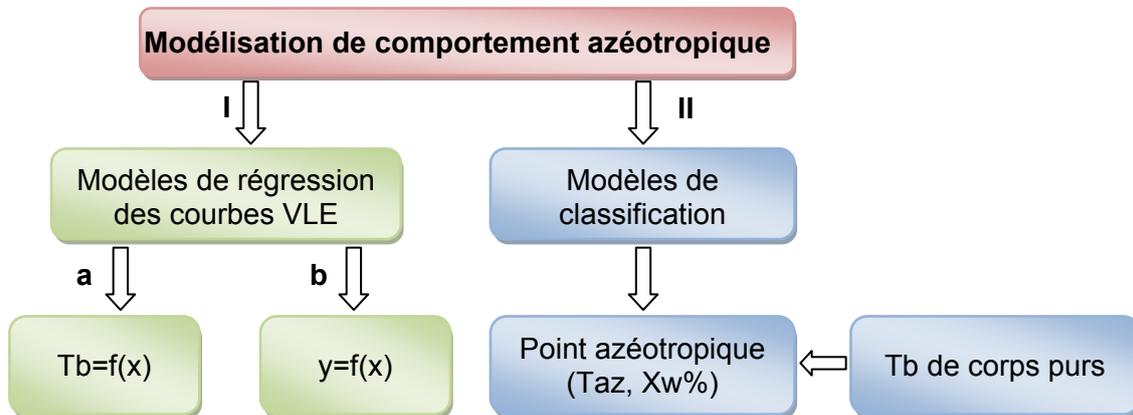
**Développement d'une approche
QSPR pour la prédiction des
propriétés physico-chimiques des
corps purs et de leurs mélanges**

Ce chapitre est dédié au développement de méthodes d'estimation du caractère azéotropique/zéotropique d'un mélange liquide binaire. Deux stratégies sont abordées dans ce travail :

- Stratégie I : Approche quantitative par régression des courbes VLE
- Stratégie II : Approches qualitative/quantitative:
 - * modélisation par classification azéotrope/zéotrope
 - * modélisation par régression du point azéotropique ($T_{az}/X_w\%$)

Le Schéma 2-1 décrit la logique du développement des deux approches.

Schéma 2-1 Différentes stratégies de modélisation du comportement azéotropiques.
a : Température d'ébullition en fonction de la composition en phase liquide
b : Composition de la phase gazeuse en fonction de la composition en phase liquide



La stratégie I permet de prédire la courbe d'équilibre en entière (la courbe de rosée et la courbe de bulle) ce qui montre en particulier le caractère azéotropique et permet de déterminer le point azéotropique (température et composition), s'il existe. De plus, la modélisation de la courbe entière permet d'avoir une information complète sur un mélange binaire pour n'importe quelle composition. Néanmoins, cette méthode n'est pas fiable pour la détermination du caractère azéotropique/zéotropique: dans certains cas les courbes VLE sont difficile à interpréter. Une courbe prédite décrivant un azéotrope peut être très proche d'une courbe expérimentale décrivant un zéotrope. La Figure 2-11 représente un exemple d'un système binaire difficile à classer à cause de la forme des courbes près du point $x=1$.

Pour cette raison la stratégie II a été développée. Le comportement azéotropique d'un mélange est prédit par des modèles de classification, ensuite, dans le cas d'un

azéotrope, sa température d'ébullition et sa composition sont prédites à l'aide des modèles de régression. Ces modèles utilisent l'information sur la température d'ébullition des corps purs, d'où le développement des modèles pour la prédiction cette propriété.

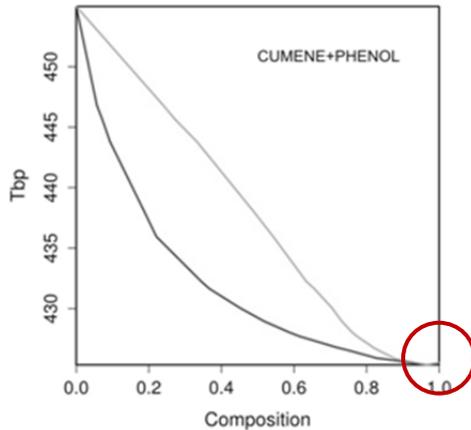


Figure 2-11 Exemple de système binaire difficile à classer.

Un projet qui n'est pas directement relié aux autres implique le développement de modèles prédictifs de la solubilité aqueuse. La modélisation de cette propriété est utile pour la société Processium, qui s'intéresse à des nombreuses propriétés physico-chimiques. La solubilité d'un liquide dans un autre est un problème important pour les mélanges, néanmoins cette question n'a pas été traitée en détail et juste une première étape, la modélisation de la solubilité dans l'eau, a été résolue.

La disponibilité des données dépende de la propriété à modéliser. Pour cette raison, des jeux de données différents ont été utilisés pour chaque approche.

Pour la modélisation de la courbe d'ébullition et la courbe d'équilibre les données ont été regroupés à partir de la base de données coréenne (KDB)[1].

Les données pour la classification ont été rassemblées une partie dans la KDB et une autre dans la compilation de Horsley[2]. Une partie de ces données a servi comme jeu de modélisation et une autre comme jeu de test. Un deuxième jeu de test a été fourni par la société Processium contenant des données provenant de sa propre base de données.

Le jeu de modélisation pour le point azéotrope a été pris dans le livre [3] tandis que le jeu de test a été choisi parmi les mélanges retrouvées dans la publication de Gmehling [4].

Les données qui ont servi pour modéliser la température d'ébullition des corps purs ont été fournies par la société Processium et ont été sélectionnés dans sa propre base de données. Un premier jeu de test provient de mêmes sources tandis qu'un deuxième jeu de test contient les données de travaux d'Artemenko [5].

Enfin, les modèles pour la solubilité ont été développés en utilisant un jeu de données compilé des travaux de Huuskonen[6], Yaffe[7] et Jurs[8]. Ce jeu a été déjà utilisé par Dr. Denis Fourches pendant sa thèse[9].

Le Tableau 2-2 synthétise le nombre de composés dans le jeu de modélisation et de test (VS), pour chaque propriété modélisée.

Tableau 2-2 Synthèse de jeu de modélisation et de test, pour chaque propriété modélisée.

Propriété modélisée	Jeu de modélisation	Jeux de test		
		VS1	VS2	VS3
Courbe d'ébullition	167	94		
Courbe d'équilibre $y=f(x)$	224	17		
Classification	400	96	499	
Point azéotropique	176	24		
Température d'ébullition	2098	516	290	
Solubilité	1635	21	77	191

2.5 Références

1. Kang, J.W., et al., *Development and current status of the Korea Thermophysical Properties Databank (KDB)*. International Journal of Thermophysics, 2001. **22**(2): p. 487-494.
2. Horsley L, H., *Table of Azeotropes and Nonazeotropes*, in *AZEOTROPIC DATA*. 1973, American Chemical Society. p. 1-314.
3. Gordon, A.J. and R.A. Ford, *The Chemist's Companion. A Handbook of Practical Data, Techniques, and References*. 1972, New York: John Wiley and Sons. 537.
4. Gmehling, J. and R. Boelts, *Azeotropic Data for Binary and Ternary Systems at Moderate Pressures*. J. Chem. Eng. Data, 1996. **41**(2): p. 202-209.
5. Artemenko, N.V., et al., *Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks within the Substructure Approach*. Doklady Chemistry, 2001. **381**(1): p. 317-320.
6. Huuskonen, J., *Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology*. Journal of Chemical Information and Computer Sciences, 2000. **40**(3): p. 773-777.
7. Cohen, Y., et al., *A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) for predicting aqueous solubility of organic compounds*. Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1177-1207.
8. Jurs, P.C. and N.R. McElroy, *Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure*. Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1237-1247.
9. Fourches, D., *Modèles multiples en QSAR/QSPR: Développement de nouvelles approches et leurs applications au design "in silico" de nouveaux extractants de métaux, aux propriétés ADMETox ainsi qu'à différentes activités biologiques de molécules organiques*. 2007.

3 Courbe d'ébullition Tb=f(X)

Ce travail a été fait en collaboration avec une équipe de l'Institut physico-chimique A.V. Bogatsky, d'Odessa. Le but a été de développer et comparer deux types différents de descripteurs "spéciaux" pour des mélanges liquides binaires basés sur les approches SiRMS (descripteurs Simplex) et ISIDA (descripteurs fragmentaux). Pour les modèles RF les descripteurs Simplex ont été utilisés, tandis que les descripteurs fragmentaux ont été employés pour des modèles ASNN et SVM.

La modélisation a été réalisée sur un jeu de données de 167 mélanges liquides binaires contenant 67 composés purs, pour lesquels 3252 points sont disponibles (cf. Annexe 11.6). La validation croisée a été faite selon trois stratégies développées dans cette thèse: «Points Out», "Mixtures Out" et "Compounds out».

Indépendamment des machines d'apprentissage ou des descripteurs utilisés, les performances obtenues sont très proches les unes des autres. L'utilité d'une méthode de combinaison de modèles, le stacking a été démontrée. Les résultats obtenus avec cette méthode sont meilleurs par rapport aux modèles individuels.

De plus, une étude de benchmark montre que, malgré le nombre réduit de données, nos modèles sont aussi performants que les modèles COSMO-RS ou UNIFAC.

Cette étude est décrite plus en détail dans l'article qui a été publié récemment dans le *Molecular Informatics*. DOI: [10.1002/minf.201200006](https://doi.org/10.1002/minf.201200006)

QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids

I. Oprisiu,^[a, b] E. Varlamova,^[c] E. Muratov,^[c, d] A. Artemenko,^[c] G. Marcou,^[a] P. Polishchuk,^[c] V. Kuz'min,^[c] and A. Varnek^{*[a]}

Abstract: This paper is devoted to the development of methodology for QSPR modeling of mixtures and its application to vapor/liquid equilibrium diagrams for bubble point temperatures of binary liquid mixtures. Two types of special mixture descriptors based on SiRMS and ISIDA approaches were developed. SiRMS-based fragment descriptors involve atoms belonging to both components of the mixture, whereas the ISIDA fragments belong only to one of these components. The models were built on the data set containing the phase diagrams for 167 mixtures represented by different combinations of 67 pure liquids. Consensus models were developed using nonlinear Support Vector Machine (SVM), Associative Neural Networks (ASNN), and Random Forest (RF) approaches. For SVM and ASNN calculations, the ISIDA fragment descriptors were used,

whereas Simplex descriptors were employed in RF models. The models have been validated using three different protocols: "Points out", "Mixtures out" and "Compounds out", based on the specific rules to form training/test sets in each fold of cross-validation. A final validation of the models has been performed on an additional set of 94 mixtures represented by combinations of novel 34 compounds and modeling set chemicals with each other. The root mean squared error of predictions for new mixtures of already known liquids does not exceed 5.7 K, which outperforms COSMO-RS models. Developed QSAR methodology can be applied to the modeling of any nonadditive property of binary mixtures (antiviral activities, drug formulation, etc.)

Keywords: QSAR/QSPR · Vapor/liquid equilibrium · Bubble point curve · Mixtures prediction

1 Introduction

Vapor-liquid equilibrium (VLE) data represent one of the most important type of information required to evaluate the phase behavior of a binary liquid mixture, which is crucial for the design of separation processes.^[1] Particular interest represents the dependence of bubble point or vapor pressure on the mixture composition (Figure 1). Theoretical assessment of these data could significantly reduce the costs of selection of proper agents for industrial processes.

Group contribution methods (GCM), such as UNIFAC,^[2] UNIFAC-Dortmund,^[3] ASOG^[4] or UNIQUAC^[5] are used worldwide to predict mixture behavior. UNIFAC is based on the thermodynamic equation for the activity coefficient (γ) of liquid 1 in the environment of liquid 2. To calculate γ , UNIFAC considers interactions of selected "structural groups" i ($i \in 1$) and j ($j \in 2$) accounted for the "energy parameters" A_{ij} and A_{ji} . The latter are fitted on available experimental data. An extensive table of UNIFAC group-interaction parameters was first published by Fredenslund et al.^[6] in 1977. Then it has been several times revised because of the growing volume of experimental data. The latest UNIFAC update was based on the Dortmund Data Bank containing more than 39 000 VLE data.

Despite its solid thermodynamics basis and excellent results of quantitative estimations of vapor-liquid equilibrium, UNIFAC has two serious drawbacks. The first one concerns the energy parameters A_{ij} and A_{ji} which cannot be assessed from the parameters of individual groups i and j but must be fitted directly on experimental VLE data for the binary

[a] I. Oprisiu, G. Marcou, A. Varnek
University of Strasbourg
Strasbourg, France
phone: +33.3.68.65.15.60
*e-mail: varnek@chimie.u-strasbg.fr

[b] I. Oprisiu
Processium
62 Boulevard Niels Bohr, BP 2132, F 69603 Villeurbanne, France

[c] E. Varlamova, E. Muratov, A. Artemenko, P. Polishchuk, V. Kuz'min
A. V. Bogatsky Physical-Chemical Institute
Odessa, Ukraine

[d] E. Muratov
University of North Carolina
Chapel Hill, USA

Supporting Information for this article is available on the WWW under <http://dx.doi.org/10.1002/minf.201200006>

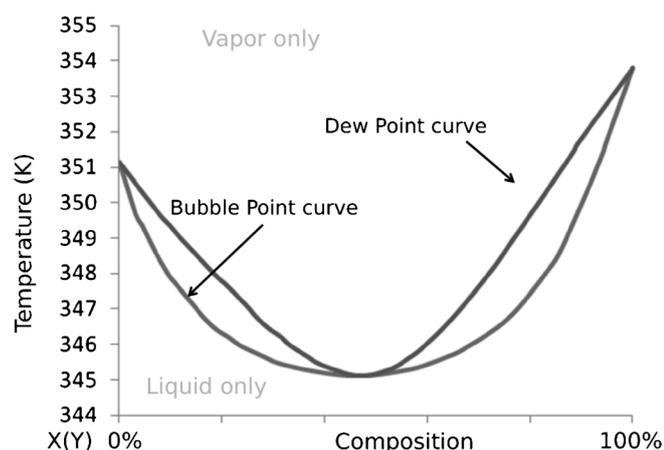


Figure 1. Vapor-liquid equilibrium curve showing the variation of equilibrium composition of the liquid mixture with the temperature at a fixed pressure. The dew-point curve represents the temperature at which the saturated vapor starts to condense whereas the bubble-point is the temperature at which the liquid starts to boil.

mixtures. Thus, the total number of the energy parameters rises as the squared number of groups. According to our estimations, for the current list of some 60 groups, less than half energy parameters were calculated. The second drawback is related to inflexible strategy of groups' selection which in most of cases makes impossible to represent a complex systems as an ensemble of groups. For instance, in recently reported UNIFAC model for ionic liquids (IL),^[7] 12 new groups have been introduced and each of them represents an individual IL. Taking into account a large number of existing ILs, further development of UNIFAC models for these systems looks unrealistic. Similar conclusion could be drawn for variety of liquid binary systems involving heterocyclic molecules because each of them should be also represented as an individual group.

As an alternative of GCM, a novel method for the prediction for thermo-physical properties of fluids is used. COSMO-RS (Conductor-like Screening Model for Real Solvents)^[8] approach is based on dielectric continuum models and statistical thermodynamics. The standard procedure of COSMO-RS calculations consists essentially in two steps: quantum chemical calculation of the local polarization charge density σ for each component of the mixture followed by COSMO-RS statistical calculations.^[9] As a prerequisite, the 3D distribution of the polarization charge on the surface of each molecule is converted into a surface composition function (σ -profile).^[8]

For the calculation of the chemical potentials (activity coefficients) using this model, analogously to GCM, the interactions between the molecules are taken into account. The surface of each molecule is subdivided into segments with equal area, and the chemical potential is derived from the interactions between the surface segments. Unlike GCM, in COSMO-RS model only a few element-based parameters must be fitted. Notice that dispersive interactions and hy-

drogen bonding are poorly accounted for in COSMO-RS.^[10,11]

Quantitative Structure-Property Relationship (QSPR) approach could be considered as a valuable alternative to the previously described methods. A QSPR model relates a given physical property with chemical structure encoded by molecular descriptors. Although, QSPR technique is traditionally used to model individual compounds, some efforts have been recently made to model their mixtures.^[12] Thus, Kravtsov et al. developed neural networks models for solvation free energies in different solvents^[13] and rate constants of nucleophilic substitution reactions^[14] and SN1^[15] using simple concatenation of substrate and solvent descriptors. Ravindranath et al.^[16] reported a "mixed" approach which integrates thermodynamics Margules, NRTL and UNIQUAC approaches with QSPR analysis. In this study, the groups interaction parameters involved in Margules, NRTL or UNIQUAC equations for activity coefficients have been modeled by QSPR. Ajmani et al.^[17-19] reported QSPR models for infinite-dilution activity coefficients, excess molar volume and density of liquid binary mixtures using special mixture descriptors which were calculated as mole weighted average using the descriptor value and mole fraction of each pure component in the mixture. Recently, Karitzky et al.^[20] performed QSPR modeling of normal boiling point temperature of azeotropes (T_{az}) using the CODESSA PRO program. Two different strategies have been used to prepare mixture descriptors: either simple arithmetic average of those calculated for individual molecular components 1 and 2, or weighting 1 and 2 by their molar ratios in the azeotrope. Predictive performance of the obtained in^[16] linear models is rather weak: the standard deviation of about 23 K has been obtained at the fitting stage and has not even been reported for the external test set.

The most challenging problem in QSAR of mixtures is a representation of mixture by descriptors. Thus, prior to modeling, the investigators should decide which descriptors are the most suitable for the modeling of mixtures (binary liquids in the given study). Should the nonadditivity effects be included at the descriptor design level, or mixture descriptors could be simply constructed from those of individual components? Here, we examine both strategies. Another question is related to proper external validation of models for mixtures which is less obvious than in classical QSAR. Some efforts have been done by^[17-19] who reported validation procedure similar to "Points Out" and "Mixture Out" strategies suggested in this work (see Section 1.2). This, however, is not sufficient to assess prediction performance for mixtures containing new compounds. Indeed, if both training and external sets include data points of the same mixture the model's performance to predict new mixture is not truly estimated. The drawbacks of conventional n-fold external cross-validation in QSAR of mixtures are discussed elsewhere.^[21] Thus, new more rigorous protocol for external validation must be developed specially for QSAR modeling of mixtures. One more problem is a detection of

outliers for the curves which is also different from classic QSAR.

Thus, the goal of this study is the development of the solid workflow of QSPR analysis of mixtures which includes: (i) development of two types of mixture descriptors: "non-additive" SiRMS descriptors and additive ISIDA descriptors; (ii) ensemble QSPR modeling of bubble-point temperature of mixtures of organic compounds using SVM, ASNN, and RF methods; (iii) rigorous external validation of obtained models; (iv) detection of outlier mixtures for developed QSPR models; (v) benchmarking of obtained QSPR models with COSMO-RS approach.

It should be noted that the developed QSAR methodology is not limited by particular case of phase diagrams, but it could be used to model any property of binary mixtures (antiviral activities, drug formulation, etc.)

2 Materials and Methods

2.1 Dataset Descriptions

2.1.1 Modeling Set

The dataset was compiled from Korean Data Base (KDB).^[22] It consists of 67 pure liquids and 167 of their mixtures. Each mixture has been represented by several (7–57) points, thus, 167 modeling set mixtures have been described by 3185 data points. The matrix of mixtures is very sparse and consists of only 167 out of possible 2211 combinations, i.e., sparsity degree is 92.5%. One compound could be involved in different number of mixtures (from 1 to 25). The total distribution of the number of mixtures and data points per pure liquid is represented in Table 1, with an average number of 5 mixtures and 95 data points per pure compound. The bubble temperature (T_b) was expressed in Kelvin scale and has a range from 280.25 to 462.65 K for modeling set and from 315.95 to 544.26 for the external set. It creates an additional problem to predict external compounds and mixtures, because for some of them the temperature exceeds up to 80 K the maximal T_b for the modeling set.

Generally experimental measure errors reported in different publications are around 0.06 K for the bubble temperature (T_b) and 0.1% for the composition (X).^[23,24] These are the measure errors made within the same laboratory; they are smaller than the errors of the same measurement made by different labs, which could be as high as 0.5 K and 1%, respectively.^[25,26] In some cases due to imprecise control of atmospheric pressure and different approximation methods employed in each laboratory, the average error for T_b of pure compounds may reach 12–18 K.^[27]

2.1.2 External Validation Set

The models built on the entire modeling set have been additionally validated on the external validation set of 94 new

mixtures involving 66 compounds. Only 27 out 94 mixtures (632 data points) contain no new pure compounds and 67 mixtures (1386 points) contain at least one new compound. Thus, 32 external compounds are common to the modeling set, whereas other 34 are new. Four mixtures have no common compounds with the modeling set.

2.2 Strategies for Model External Validation in QSAR of Mixtures

Three different strategies of external validation were established (Figure 2): (i) "points out" – prediction of T_b for any molar ratio of the known biphasic systems, (ii) "mixtures out" – prediction of T_b for the missed data in the mixture matrix (gap-filling) formed by 67 pure liquids from the modeling set, and (iii) "compounds out" – prediction of T_b for mixtures formed by "new" pure compound(s) absent in the modeling set.

"Points out". All pure compounds were always kept in the training set and mixture data points were randomly taken to each fold of external cross-validation set. Each mixture is present both in training and test sets. Here, 2-fold external cross-validation repeated 3 times has been performed.

"Mixtures out". In each fold of external cross-validation all pure compounds were always kept in the training set but whole mixtures were randomly selected to test set. Thus, each mixture is present either in the training or in the test set, but never in both sets. Here, 5-fold external cross-validation repeated 3 times has been performed. Expected error of prediction for this models is bigger than for "points out" strategy, however, this model will not be limited by already known mixtures, but will be useful for the filling of the missed data in the mixture matrix formed from 67 training set compounds. Because this matrix is very sparse, $2211 - 167 = 2044$ new mixtures could be predicted.

"Compounds out". Pure liquid and all its mixtures were simultaneously taken to an external fold. Thus each mixture in the external set contains at least one compound which is absent in the training set. The difference with classical CV algorithm is that the folds were not created randomly, but supervised in order to keep the number of both pure liquids and the mixtures amongst the folds more or less constant. The supervision is needed because one pure liquid, for instance bromobenzene, can participate in only one mixture, while another – carbon tetrachloride – can create 25 mixtures and the classical random algorithm is unable to consider such situation during external folds creation. Moreover, despite the supervised process of folds creation, we are aware that some folds could be predicted badly because they are still anisotropic and sufficient lack of information in the training set could be observed for some external folds. That was the reason of triple repetition of 10-fold external CV. It is necessary to note that, because we are dealing with binary mixtures, every mixture will be taken to the external set twice, except the case when both

Table 1. Distribution of number of mixtures and data points per pure liquid.

Compound	Mixtures	Points	Compound	Mixtures	Points
Carbontetrachloride	25	444	<i>m</i> -Xylene	2	59
Methanol	16	313	Water	3	59
Benzene	16	310	Dibromomethane	4	54
Ethanol	12	279	2,4-Dimethylpentane	3	53
<i>n</i> -Butanol	13	260	1,1-Dichloroethane	2	44
Cyclohexane	12	250	Benzotrifluoride	3	43
Acetonitrile	10	242	Methylcyclopentane	3	43
1-Bromopropane	10	226	Cyclohexene	2	40
Toluene	11	221	1-Hexene	2	39
Ethylacetate	9	217	1-Chlorobutane	3	35
Methylacetate	10	213	2,2,5-Trimethylhexane	2	35
<i>n</i> -Propanol	12	207	1-Octene	2	34
<i>n</i> -Octane	8	159	<i>n</i> -Butylacetate	1	29
<i>n</i> -Heptane	9	148	<i>n</i> -Butylformate	1	28
Hexafluorobenzene	5	139	<i>n</i> -Butyl- <i>n</i> -butyrate	1	28
<i>sec</i> -Butanol	8	136	<i>n</i> -Butylpropionate	1	28
1,2-Dichloroethane	8	132	<i>p</i> -Cresol	1	23
Chloroform	6	132	<i>n</i> -Decane	1	22
Vinylacetate	6	117	Cumene	1	19
Isopropanol	6	114	Phenol	1	19
<i>n</i> -Hexane	6	114	2-Methylpentane	1	18
Acetone	7	107	3-Methylpentane	1	18
Ethylbenzene	6	105	Cyclopentane	1	18
2,3-Dimethylbutane	5	102	2,2-Dimethylbutane	1	18
<i>tert</i> -Butanol	6	102	1,1,2-Trichlorotrifluoroethane	1	17
Chlorobenzene	6	87	1,2-Dichlorotetrafluoroethane	1	17
<i>p</i> -Xylene	4	83	1,1,1-Trichloroethane	1	16
Isobutanol	6	75	Dibutylether	1	14
2,2,4-Trimethylpentane	4	72	<i>n</i> -Propylacetate	1	14
<i>o</i> -Xylene	5	72	Bromobenzene	1	13
Methylcyclohexane	4	69	Ethyleneglycol	1	12
Methylmethacrylate	3	69	<i>n</i> -Propylbenzene	1	10
Tetrachloroethylene	4	64	Acrylonitrile	1	7
Trichloroethylene	4	64			

compounds are belonging to the same external fold. It simulates the addition of novel component to existing matrix of mixtures. This is the most rigorous way of external validation of QSAR models for mixtures. Although the error of prediction for this strategy is expected to be the biggest, the models passed the validation will be able to predict T_b for mixtures created by a new pure compound beyond the modeling set.

2.3 Mixture Descriptors

2.3.1 Simplex Representation of Molecular Structure (SiRMS)

In the frameworks of Simplex representation of molecular structure (SiRMS)^[28–30] any molecule can be represented as an ensemble of different 2D tetratomic fragments of fixed composition, structure, chirality and symmetry simplexes. A number of identical simplexes in a molecule is a descriptor value of that simplex. The connectivity of atoms in simplex, atom type and bond nature (single, double, triple, or aromatic) have been considered at the 2D level (Figure 3). Bounded and unbounded 2D simplexes were used. Not

only atom type, but some other physical-chemical characteristics of atoms, i.e., partial charge, lipophilicity, refraction, and atom's ability for being a donor/acceptor in hydrogen-bond formation were used for atom labeling in simplexes (Figure 3). For these atom characteristics the binning procedure has been used to transform real values (charge, lipophilicity, and refraction) to four categories corresponding to their (i) partial charge $A \leq -0.05 < B \leq 0 < C \leq 0.05 < D$, (ii) lipophilicity $A \leq -0.5 < B \leq 0 < C \leq 0.5 < D$, and (iii) refraction $A \leq 1.5 < B \leq 3 < C \leq 8 < D$. Three characteristics of atom H-bond formation ability were specified A (acceptor of hydrogen in H-bond), D (donor of hydrogen in H-bond), and I (indifferent atom).

Bounded simplexes describe only single components of the mixture (Compounds 1 or 2), whereas unbounded simplexes can describe both the constituent parts and the mixture as a whole (Figure 4). With this purpose it is necessary to indicate whether the parts of unbounded simplexes are belonging to the same molecule or to different ones. A special mark is used during descriptors generation to distinguish such simplexes. Descriptors of constituent parts (Compounds 1 and 2) are weighted according to their

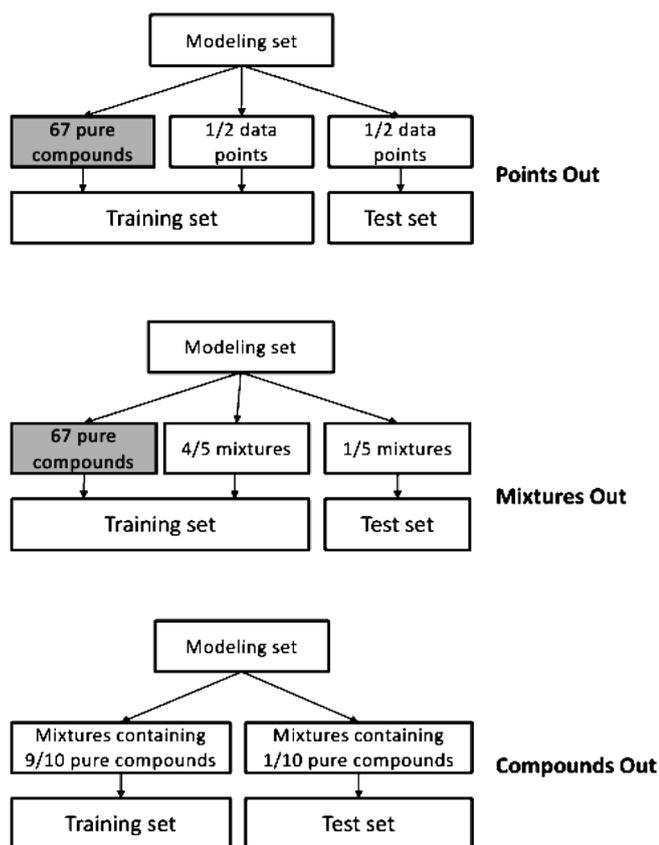


Figure 2. Different strategies used for models validation.

molar fraction and summarized, and mixture descriptors are multiplied on doubled minimal weight according to Equation 1. If both mixtures and pure liquids are considered, descriptors of pure liquid have the weight equal to 1.

$$D = \begin{cases} x_1 D_1 + x_2 D_2 \\ 2x_1 D_{1+2} \end{cases}, \quad (1)$$

where D is the descriptor value, x_1 and x_2 are molar frac-

tions of Components 1 and 2 ($x_1 < x_2$ and $x_1 + x_2 = 1$), D_1 , D_2 , and D_{1+2} are descriptor values for individual Compounds 1 and 2, and for their mixture, respectively.

2.3.2 ISIDA Fragment Descriptors

ISIDA fragment descriptors^[31] were used in combination with SVM and ASNN machine learning methods. Two different types of molecular subgraphs are considered (Figure 5): "sequences" (I) and "augmented atoms" (II).

The sequences correspond to consecutive set of atoms linked by chemical bonds, where either atom types (C, N, O, ...) or bond types (single, double, ...) or both of them are considered explicitly. In the following, we specify the number of atoms of a given sequence. Thus, I(AB, n_{\min} – n_{\max}) refers to all sequences containing from n_{\min} to n_{\max} atoms connected by bonds of specified type. For I(A, n_{\min} – n_{\max}), the definition is similar, but bond types are omitted. Only shortest paths from one atom to the other are used, as shown in Figure 5.

An "augmented atom" represents a selected atom within its nearest environment including both neighboring atoms and bonds (AB), or atoms only (A), or bonds only (B). The neighborhood is described by concatenating all sequences starting at a given atom and of a given length. For instance II(AB, n_{\min} – n_{\max}) refers to fragments concatenating sequences of length from n_{\min} to n_{\max} around the selected atom. In this work, $n_{\min} \geq 2$ and $n_{\max} \leq 10$ were used for the sequence and $n_{\min} \geq 2$ and $n_{\max} \leq 4$ for augmented atoms. In QSPR models each fragment is considered as an individual descriptor, whereas its occurrence in the given molecule is the descriptor value.

Mixture descriptors have been obtained by combining ISIDA descriptors for each component (Equation 2, Figure 6).

$$D = \begin{cases} x_1 D_1 + x_2 D_2 \\ |x_1 D_1 - x_2 D_2| \end{cases} \quad (2)$$

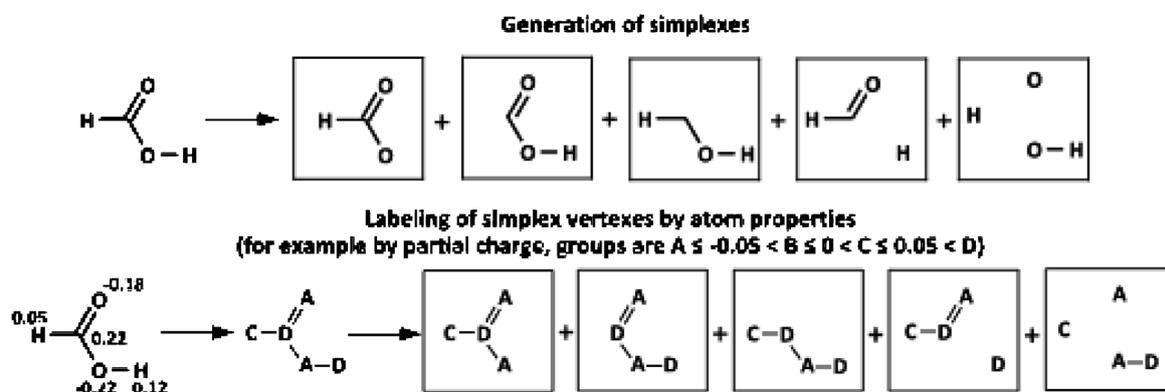


Figure 3. Simplex representation of molecular structure (SiRMS).

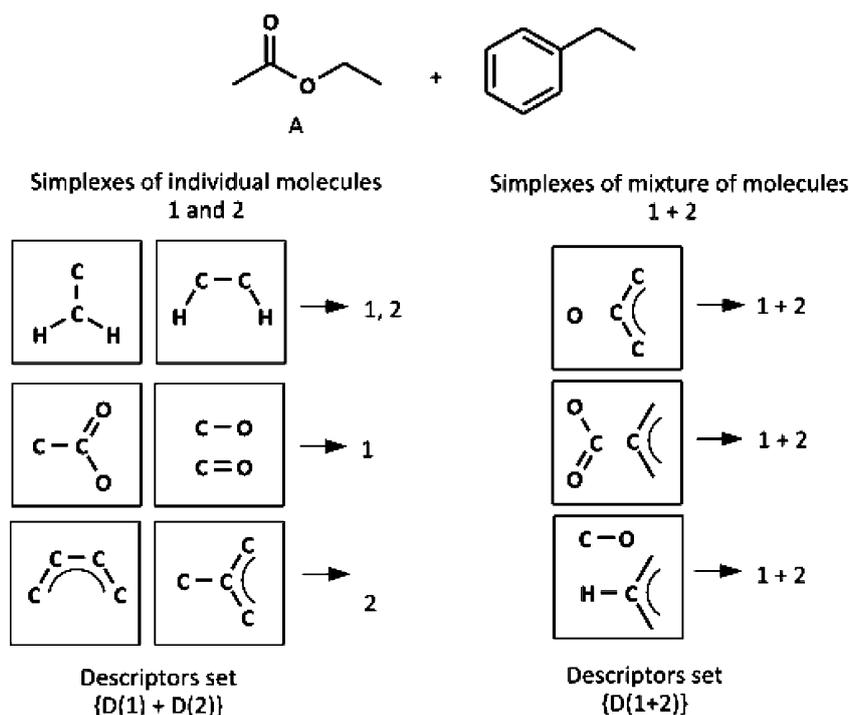


Figure 4. Simplex descriptors for a binary mixture.

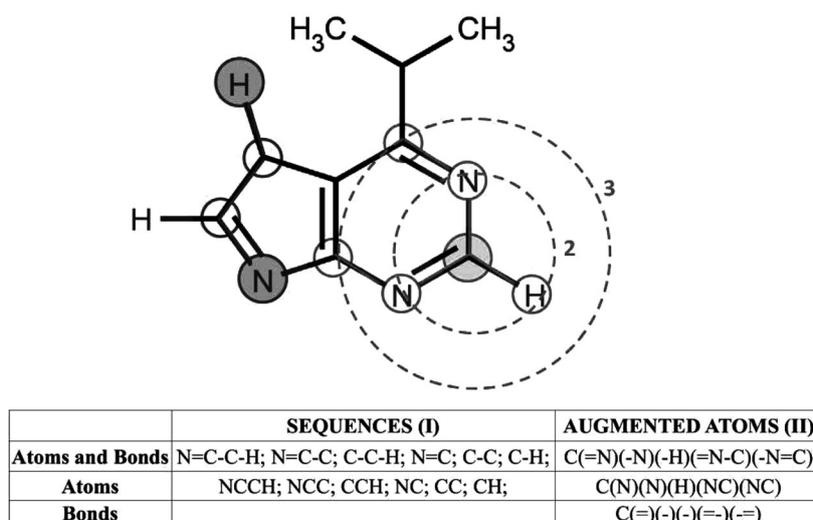


Figure 5. ISIDA Fragmentation. Two classes of substructural fragments: atom/bond sequences and augmented atoms. From top to bottom: the sequences (I) correspond to the I (AB, 2–4) and I (A, 2–4) types involving the shortest paths between each pair of atoms. Augmented atoms (II) correspond to the II (AB, 2–3), II (A, 2–3) and II (B, 2–3) types.

where D_1 and D_2 are descriptor values for individual Compounds 1 and 2, x_1 and x_2 are molar fractions of components 1 and 2.

2.4 Machine-Learning Methods

Three following statistical approaches were used: Random Forest (RF) in combination with Simplex descriptors, when

ISIDA fragment descriptors were employed in SVM and ASNN methods.

2.4.1 Random Forest

RF models were obtained according to the original RF algorithm described in^[32] and realized in.^[33] RF is an ensemble of single decision trees. This ensemble produces a corre-

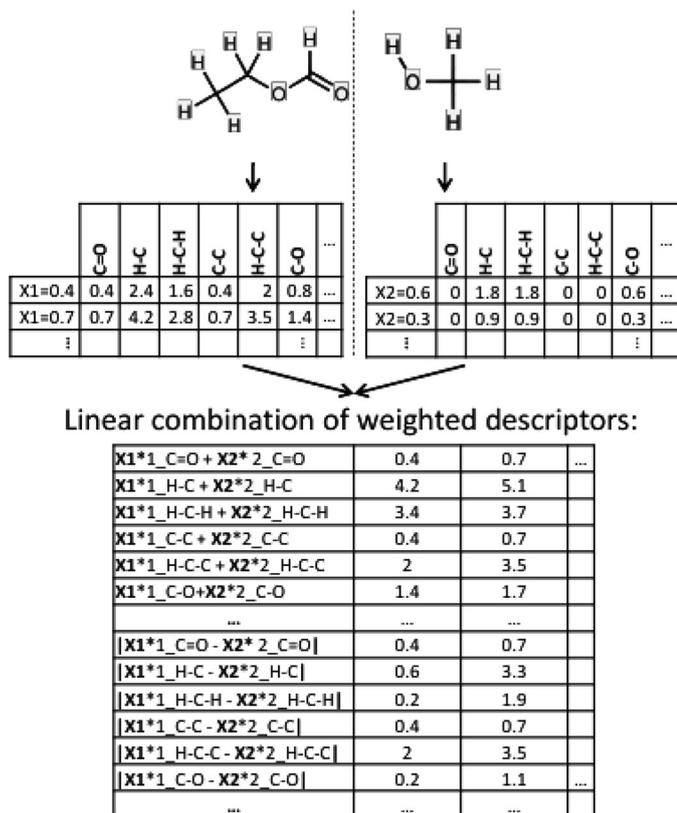


Figure 6. ISIDA mixture descriptors. X1 and X2 are the molar ratios of the first and the second component in the mixture, respectively. For each component of the liquid, the numbers correspond to the product of the fragment's occurrence and molar ratio X.

sponding number of outputs. Outputs of all trees are aggregated to obtain one final prediction as an average of the individual tree predictions. Each tree has been grown as follows: (i) A bootstrap sample, which will be a training set for the current tree, is produced from the whole training set of N compounds. Compounds which are not in the current tree training set are placed in an out-of-bag (OOB) set ($\sim N/3$ molecules). (ii) The best split among the m randomly selected parameters from the initial set of M descriptors is chosen in each node by CART algorithm.^[34] The value of m is just one tuning parameter for which RF models are sensitive. (iii) Each tree is grown to the largest possible extent without any pruning. Performance of the models has been assessed on OOB sets, which values are similar to those obtained in 5-fold external cross-validation procedure.^[35] The model selection has been performed according to R^2_{OOB} values. Each individual RF model involved 457–508 simplex descriptors.

Associative Neural Network (ASNN)

An associative neural network (ASNN) is a combination of an ensemble of 100 feed-forward neural networks and the KNN technique. Three layers architecture of neural network

has been used. The number of neurons in the input layer corresponds to the number of descriptors, whereas the output layer consists of one neuron (modeled property). The number of neurons in the hidden layer is equal to 5, as recommended by Tetko et al.^[36] ASNN uses correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbor technique. This method corrects a bias of a global model for a considered data case by analyzing the biases of its nearest neighbors determined in the space of calculated models. Early stopping technique has been used to avoid an overfit of the models.^[37] The ASNN 1.0 program provided by Dr. Igor V. Tetko has been used in this work.

Support Vector Machines (SVM)

In SVM a nonlinear function is learned by linear fitting in the feature space which is a nonlinear mapping of the initial (input) space in which the fitting problem was expressed.^[38] The libSVM, a C library was used to generate SVM models. The calculations have been performed with the RBF kernel. The γ and ϵ parameters have been optimized in grid calculations.

Using different initial pools of ISIDA descriptors corresponding to different fragmentation types, several tens of SVM and ASNN models have been trained. Only those models with determination coefficient $R^2 > 0.8$ at cross-validation have been retained and used for the predictions for SVM and ASNN consensus models.

Once the ensemble of models issued from a given machine-learning method is obtained, a consensus model can be calculated either by simple averaging of predictions or by using Stacking technique^[39] which develops a linear regression using predictions made by each model as independent variables.

2.4.2 Applicability Domain

QSPR models are obtained on a training set, which, no matter how large, may never represent a significant sample of the entire chemical space. Applicability domain (AD) of a model defines a region of the chemical space that can be adequately covered by training set compounds. Statistical models can deliver reliable predictions for the compounds belonging to this region. Presumably, the models cannot be trustfully used outside of their AD.

Minimum spanning tree AD approach^[28] was used for RF models. Minimum spanning tree has been built in the space of decision trees predictions for the given RF model using Kruskal's algorithm.^[40] Then average distance (d_{av}) and its root-mean-square deviation (σ) among all tree edges have been calculated. Substantially, such distance is the characteristic of average density of molecules distribution in the considered space. If any of external set molecules has been situated on the distance bigger than $d_{av} +$

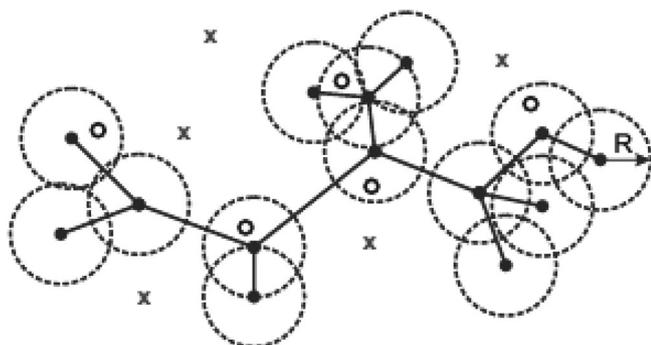


Figure 7. Scheme of Local AD approach based on minimum spanning tree. ● – training set molecules, ○ – test set compounds within AD, x – test set compounds outside AD, radius of each sphere is $R = d_{av} + 3\sigma$.

3σ from the nearest training set point, it means that this external set molecule is situated outside AD (Figure 7).

Fragment control^[41] AD approach has been used for SVM and ASNN models. Any molecules containing the fragments which do not occur in compounds of the training set are considered to be outside AD in this method.

2.5 Statistical Characteristics Used

Determination Coefficient (R^2) and Root Mean Squared Error (RMSE) for the external set were used to estimate the predictivity of the obtained models.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{\text{pred},i} - y_{\text{exp},i})^2}{\sum_{i=1}^n (y_{\text{exp},i} - \bar{y}_{\text{exp}})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{\text{pred},i} - y_{\text{exp},i})^2} \quad (4)$$

where n is the number of compounds in the external set (folds); y_{exp} is the experimental value of T_b and y_{pred} is the predicted value of T_b .

2.6 Outliers Analysis

Outliers analysis for modeled curves has been performed using z^2 parameter:

$$z^2 = \frac{\sum (Y_{\text{exp}} - Y_{\text{pred}})^2}{\frac{1}{n-1} \sum (Y_{\text{exp}} - \bar{Y}_{\text{exp}})^2}, \quad (5)$$

where Y_{exp} represents the experimental temperature, \bar{Y}_{exp} is the average of the experimental values of T_b , and Y_{pred} is T_b predicted value.

The z^2 values follow a Chi-square law^[42] with n degrees of freedom, where n is the number of points in the mixture. If z^2 for a given curve is higher than the threshold value at

99% of confidence the corresponding mixture is considered an outlier.

2.7 Benchmarking

Predictive performance of developed QSPR models has been compared with the ones of COSMO-RS model.

All COSMO-RS calculations of this work were performed using the COSMOtherm program (version C21_0108) using precomputed σ -profiles^[43] for pure compounds. Since σ -profiles for some molecules were not reported in,^[43] the calculations have been performed for 166 out of 167 mixtures of the modeling set and for 89 out of 94 mixtures of the test set.

3 Results and Discussion

3.1 External Cross-Validation

Results of n -fold external CV are shown at Table 2. As expected, the error of prediction increased in the order “points out” < “mixtures out” < “compounds out”. In “points out” strategy data points of the same mixture are simultaneously present in both training and external set. Therefore, high accuracy of predictions ($RMSE = 1.1$ K) is not surprising. In “mixtures out” strategy the data points for a given mixture are present either in training set or in test set, but not simultaneously in both. Therefore, $RMSE$ of 5.2 K is bigger than for “points out” strategy, but it remains on the acceptable level. The “compound out” is the most strict validation strategy ($RMSE = 7.0$ K) because a given pure compound and all its mixtures were simultaneously placed to external set during external cross-validation. Stacking consensus model significantly outperformed all other models for “mixtures out” and “compounds out” strategies (Table 2).

Typical examples of predicted bubble point curves are given on Figure 8. In most cases the calculations correctly reproduce the curves behavior: azeotrop predicted as azeotrop (Figure 8a) and zeotrop predicted as zeotrop (Figure 8b). However, in some cases, the “mixtures out” and “compounds out” models fail (Figure 8c). Generally, except of two mixtures (metanol-1-bromopropan and benzotri-fluoride-toluene), the “points out” curves very well reproduce the experimental ones, see Figure SM1 in Supporting Information. In the “mixture out” calculations, the experimental trend has not been reproduced only for 15% of studied mixtures. On the other hand, in most of the “failed” predictions, the difference between experimental and calculated T_b does not exceed 5 K, which is within the error bar of the model. The “compounds out” predictions led to bad predictions in about half of all curves (Figure SM1) which corresponds to bad statistical parameters given in Table 2.

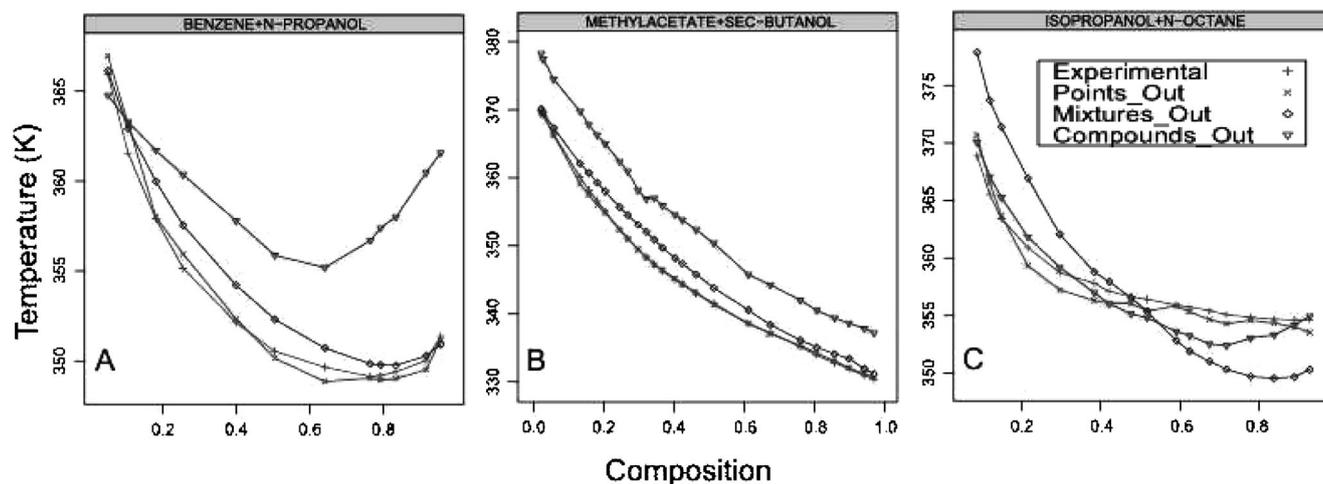


Figure 8. Typical examples of bubble point curves predictions in 5-CV for the modeling set: azeotrope predicted as azeotrope (A), zeotrope predicted as zeotrope (B) and zeotrope predicted as azeotrope in “mixture out” and “compounds out” strategies (C).

Table 2. Results of external validation of developed QSPR models.

		“Points Out”					“Mixtures Out”					“Compounds Out”				
		SVM	ASNN	RF	Av. ^[a]	Stack. ^[b]	SVM	ASNN	RF	Av. ^[a]	Stack. ^[b]	SVM	ASNN	RF	Av. ^[a]	Stack. ^[b]
Modeling set	R^2	0.98	0.99	0.98	0.99	0.99	0.86	0.92	0.90	0.93	0.95	0.73	0.89	0.79	0.85	0.90
	RMSE (K)	3.5	1.2	3.2	2.3	1.1	8.5	6.2	6.9	5.7	5.2	11.6	7.3	10.3	8.7	7.0
External set ^[c]	R^2						0.72	0.90	0.81	0.85	0.88	0.23 ^[d]	0.42 ^[d]	0.48 ^[e]	0.39	0.42
	RMSE (K)						8.8	5.2	7.2	6.6	5.9	24.3 ^[d]	21.0 ^[d]	18.5 ^[e]	22.0	21.4

[a], [b] Av.: average; Stack.: stacking. Prediction for each mixture have been calculated as a simple average of the results of SVM, ASNN and RF or using stacking approach, respectively. [c] “Mixture out” and “Compounds out” predictions were made for 27 and 67 mixtures, respectively. [d] Diethylamine + chloroform mixture was found outside of AD. [e] Formic acid + *n,n*-dimethylformamide, *p*-xylene + *n,n*-dimethylformamide and *n*-dodecane + 1-hexadecene mixtures were found outside of AD.

3.2 External Validation

The resulting models were built on the entire modeling set and then were applied to the prediction of external set compounds. The difference in range of bubble point temperatures for compounds of modeling and external sets is up to 80 K. This significantly complicates the accurate prediction of external set compounds. Calculation on 27 mixtures of external set containing no new components were considered as “mixture out” predictions, whereas those for 67 mixtures containing at least one new compound were considered as “compound out”. Performance of the “mixtures out” predictions of resulting consensus model ($R^2_{\text{ext}} = 0.88$; $RMSE = 5.7$ K) was close to 5-fold external cross-validation results ($R^2_{\text{ext}} = 0.95$; $RMSE = 5.2$ K). On the other hand, “compounds out” predictions resulted in worse statistical parameters ($R^2_{\text{ext}} = 0.44$, $RMSE = 21.0$ K) than external cross-validation ones ($R^2_{\text{ext}} = 0.90$; $RMSE = 7.0$ K). Thus, the developed models are able to fill the gaps in the matrix of mixtures formed by 67 individual compounds of modeling set, i.e., to predict T_b for 2044 missing mixtures with reasonable accuracy. But both considered methodologies (ISIDA and

SiRMS) still need some tuning to be able to predict T_b of mixtures containing at least one new compound.

At the same time, results of external cross-validation given in Table 2 and Table 3 show that consensus predictions are comparable or better than predictions obtained within one machine-learning method. The way of consensus model development is also important: stacking systematically shows better results than simple averaging.

3.3 Models Applicability Domain

An applicability domain (AD) analysis has been performed for the modeling set with “mixtures out” and “compounds out” strategies. 1,2-Dichlorotetrafluoroethane + 1,1,2-trichlorotrifluoroethane, acetonitrile + methylmethacrylate, acrylonitrile + acetonitrile and vinylacetate + methylmethacrylate were out of AD of RF models during external cross-validation. For the external set no compounds were outside AD for “mixtures out” strategy. Formic acid + *n,n*-dimethylformamide, *p*-xylene + *n,n*-dimethylformamide and *n*-dodecane + 1-hexadecene were out of AD of RF models for “compounds out” case, while ASNN and SVM

models designate only diethylamine + chloroform mixture to be out of AD.

The use of AD for RF models does not lead to a significant improvement of the model quality with the exception of external validation for “compounds out” strategy. In this case AD usage increases R_{ext}^2 value from 0.36 to 0.48 and decreases $RMSE$ from 23.2 K to 18.5 K. The AD methodology applied for ASNN and SVM did not improve predictive performance of the models.

3.4 Outliers Detection

As expected, the number of outliers increases in the order: “points out” < “mixture out” < “compounds out”. In “mixture out”, 14 mixtures have been classified as outliers for SVM, ASNN, and RF models: 2,3-dimethylbutane + acetone, 2,3-dimethylbutane + methanol, 2,4-dimethylpentane + benzene, acetonitrile + 1-bromopropane, benzene + cyclohexane, benzene + hexafluorobenzene, ethanol + 2,2,4-trimethylpentane, ethylacetate + benzene, methanol + 1-bromopropane, methanol + 1,1-dichloroethane, methylacetate + 1-hexene, methylmethacrylate + toluene, *n*-butanol + *n*-octane, *n*-propanol + *n*-octane.

Bad predictions for these mixtures could be explained either by the noise in experimental data used for model development or by specific physical phenomena observed only for few mixtures. For example, the data for isopropa-

nol + 1-bromopropane mixture were erroneously attributed to the methanol + 1-bromopropane mixture.^[44] Another example concerns the benzene + hexafluorobenzene mixture that forms both positive and negative azeotrope which is the only case regarding to 167 mixtures in the modeling set.^[45]

3.5 Benchmarking Results

Besides “1,2-dichlorotetrafluoroethane + 1,1,2-trichlorotrifluoroethane” (for which σ -profile was not reported in^[43]), the COSMOtherm program predicts all others 166 mixtures of the modeling set. Results are considerably less good ($R^2 = 0.75$, $RMSE = 11.0$ K) than QSPR stacking predictions for 5-fold external cross-validation using either “mixture out” or “compounds out” strategy (Table 2). As expected, prediction performance of the COSMO-RS model for 89 mixtures of external set ($R^2 = 0.80$, $RMSE = 11.3$ K) is similar to that for the 166 training set compounds. QSPR stacking model applied to this set led to worse statistics parameters ($R^2 = 0.52$, $RMSE = 17.6$ K). In addition, we split the external set onto “mixture out” (27 mixtures) and “compound out” (62 mixtures) test sets, predictions for which were performed separately. For the “mixture out” set the accuracy of predictions for the QSPR stacking and COSMO-RS models were found similar, but for the “compound out” set the COSMO-RS model outperforms our results (Table 3). Predicted and

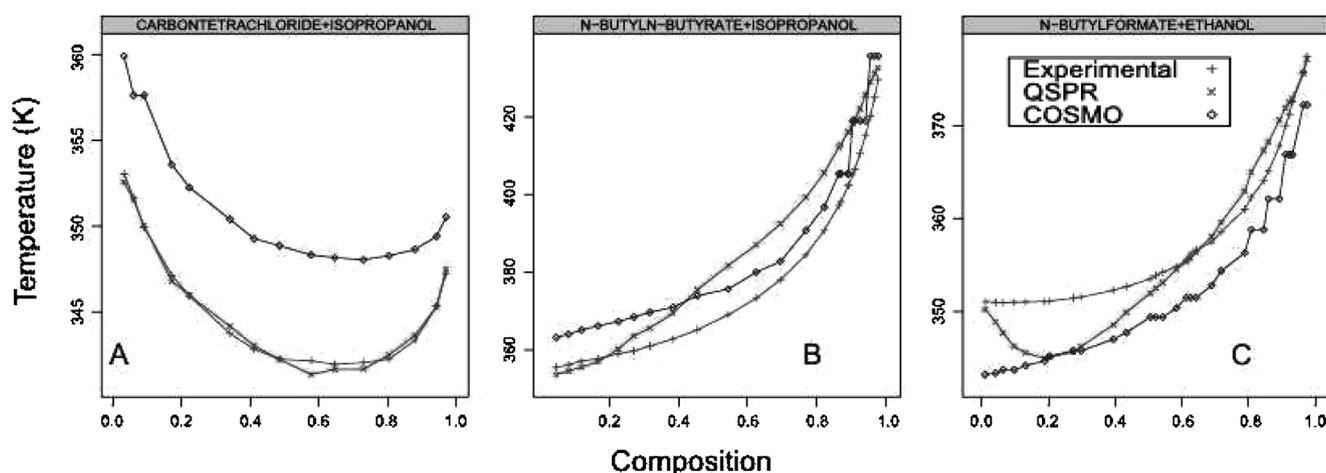


Figure 9. Typical examples of bubble point curves predicted for external “mixture out” test set of 27 compounds: azeotrope predicted as azeotrope (A), zeotrope predicted as zeotrope (B) and zeotrope predicted by QSPR as azeotrope (C).

Table 3. Statistical parameters obtained for mixtures of external test set.

	Entire set of 89 compounds ^[a]		“Mixture out” set of 27 compounds		“Compound out” set of 62 compounds	
	R^2	$RMSE$ (K)	R^2	$RMSE$ (K)	R^2	$RMSE$ (K)
Stacking	0.52	17.6	0.88	5.7	0.44	21.2
COSMO-RS	0.80	11.3	0.84	6.6	0.78	13.0

[a] Only 89 out of 94 external set compounds have been predicted with COSMO-RS model.

experimental curves for the “mixture out” and “compound out” test sets are given on Figures SM2 and SM3, respectively. As shown in Figures 9a and 9b, in most cases, both QSPR and COSMO-RS models correctly reproduce the curves behavior, although some erroneous predictions have also been observed (Figure 9c).

4 Conclusions

In this paper we described some general aspects of QSPR methodology to model the compound mixtures with non-additive response. This includes (i) design of additive or nonadditive “mixture” descriptors and (ii) cross-validation of “mixture” models which differs from that of QSPR models for individual compounds.

The developed workflow for QSPR analysis of mixtures was tested on boiling point temperatures of binary mixtures of organic compounds. Special descriptors of mixtures were developed on the basis of ISIDA (additive) and SiRMS (nonadditive) approaches. They were successfully used to predict bubble point temperatures of binary liquid mixtures from the initial data matrix. Thus, using experimental data on 167 mixtures formed by 67 individual liquids, we are able to predict with the reasonable accuracy the bubble point temperatures for $67 \times (67-1)/2 - 167 = 2044$ “missing” mixtures, i.e., to fill remaining 92.5% of the sparse data matrix. On the other hand, both SiRMS and ISIDA approaches need an additional tuning to be able to predict the mixtures containing new components absent in the modeling set.

As follows from comparing the results obtained with two different types of descriptors, nonadditivity effects taken into account at the descriptor design level do not affect the prediction performance of the models. This opens an opportunity to apply a huge variety of molecular descriptors in the modeling of nonadditive properties of mixtures.

An important achievement of this work is the establishment of the strategies of external cross-validation for QSPR of mixtures that should be different and more rigorous than the ones used in classic QSPR analysis. Developed strategies allow one to obtain realistic estimation of the model predictivity in the most rigorous way and can simulate both (i) gap-filling of initial matrix of mixtures and (ii) addition of new component (liquid) absent during the modeling.

It has been shown that consensus predictions are usually better than predictions obtained within one machine-learning method. Moreover, the way of consensus model development is critical to overall prediction performance. Thus, the stacking approach leads to better results compared to simple arithmetic averaging of the values predicted by individual models. Benchmarking studies show that QSPR models represent a good alternative to the COSMO-RS approach, especially for mixtures the individual components of which were present in the modeling set.

Developed QSAR/QSPR methodology can obviously be used for the modeling of any property of binary mixtures (antiviral activities, drug formulation, etc).

Acknowledgements

VK, PP, EM, and EV thank the ARCUS “Alsace-Russia/Ukraine” and the SupraChem Projects for support of their stay at the University of Strasbourg. EM acknowledges the support from NIH (Grant GM66940). IO thanks the Processium company for the PhD fellowship.

References

- [1] J. Gmehling, R. Böls, *J. Chem. Eng. Data* **1996**, *41*, 202–209.
- [2] J. Gmehling, P. Rasmussen, A. Fredenslund, *Chem.-Ing.-Tech.* **1980**, *52*, 724.
- [3] U. Weidlich, J. Gmehling, *Indust. Eng. Chem. Res.* **1987**, *26*, 1372–1381.
- [4] K. Tochigi, D. Tiegs, J. Gmehling, K. Kojima, *J. Chem. Eng. Jpn.* **1990**, *23*, 453–463.
- [5] T. F. Anderson, J. M. Prausnitz, *Ind. Eng. Chem. Proc. Dd.* **1978**, *17*, 552–561.
- [6] A. Fredenslund, J. Gmehling, M. L. Michelsen, P. Rasmussen, J. M. Prausnitz, *Ind. Eng. Chem. Proc. Dd.* **1977**, *16*, 450–462.
- [7] Z. Lei, J. Zhang, Q. Li, B. Chen, *Indust. Eng. Chem. Res.* **2009**, *48*, 2697–2704.
- [8] A. Klamt, *J. Phys. Chem.-Us.* **1995**, *99*, 2224–2235.
- [9] F. Eckert, *COSMOtherm User's Manual*, C2.1 Release 01.08, **2010**.
- [10] T. Mu, J. Rarey, J. Gmehling, *Indust. Eng. Chem. Res.* **2007**, *46*, 6612–6629.
- [11] A. Klamt, F. Eckert, W. Arlt, *Ann. Rev. Chem. Biomol. Eng.* **2010**, *1*, 101–122.
- [12] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, T. Khristova, V. E. Kuz'min, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *Future Med. Chem.* **2011**, *3*, 15–27.
- [13] A. Kravtsov, P. Karpov, I. Baskin, V. Palyulin, N. Zefirov, *Dokl. Chem.* **2007**, *414*, 128–131.
- [14] A. Kravtsov, P. Karpov, I. Baskin, V. Palyulin, N. Zefirov, *Dokl. Chem.* **2011**, *440*, 299–301.
- [15] A. Kravtsov, P. Karpov, I. Baskin, V. Palyulin, N. Zefirov, *Dokl. Chem.* **2011**, *441*, 314–317.
- [16] D. Ravindranath, B. J. Neely, R. L. Robinson, K. A.M. Gasem, *Fluid Phase Equilib.* **2007**, *257*, 53–62.
- [17] S. Ajmani, S. C. Rogers, M. H. Barley, D. J. Livingstone, *J. Chem. Inf. Model.* **2006**, *46*, 2043–2055.
- [18] S. Ajmani, S. C. Rogers, M. H. Barley, A. N. Burgess, D. J. Livingstone, *QSAR Comb. Sci.* **2008**, *27*, 1346–1361.
- [19] S. Ajmani, S. C. Rogers, M. H. Barley, A. N. Burgess, D. J. Livingstone, *Mol. Inf.* **2010**, *29*, 645–653.
- [20] A. R. Katritzky, I. B. Stoyanova-Slavova, K. Tamm, T. Tamm, M. Karelson, *J. Phys. Chem. A* **2011**, *115*, 3475–3479.
- [21] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'min, *Mol. Inf.* **2012**, *31*, 202–221.
- [22] J. W. Kang, K. P. Yoo, H. Y. Kim, H. Lee, D. R. Yang, C. S. Lee, *Int. J. Thermophys.* **2001**, *22*, 487–494.
- [23] N. Alpert, P. J. Elving, *Indust. Eng. Chem.* **1951**, *43*, 1174–1177.
- [24] K. J. Miller, H.-S. Huang, *J. Chem. Eng. Data* **1972**, *17*, 77–78.
- [25] T. Hiaki, K. Yamato, K. Kojima, *J. Chem. Eng. Data* **1992**, *37*, 203–206.

- [26] C. E. Kirby, M. Van Winkle, *J. Chem. Eng. Data* **1970**, *15*, 177–182.
- [27] I. V. Tetko, *The Prediction of Physicochemical Properties in Computational Toxicology*, Wiley, New York, **2006**, p. 240–275.
- [28] E. N. Muratov, A. G. Artemenko, E. V. Varlamova, P. G. Polishchuk, V. P. Lozitsky, A. S. Fedchuk, R. L. Lozitska, T. L. Gridina, L. S. Koroleva, V. N. Sil'nikov, A. S. Galabov, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, V. E. Kuz'min, *Future Med. Chem.* **2010**, *2*, 1205–1226.
- [29] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, *J. Comput. Aided Mol. Des.* **2008**, *22*, 403–421.
- [30] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, I. L. Volineckaya, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *J. Med. Chem.* **2007**, *50*, 4205–4213.
- [31] Available from: <http://infochim.u-strasbg.fr/>.
- [32] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
- [33] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, O. G. Kolumbin, N. N. Muratov, V. E. Kuz'min, *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.
- [34] L. Breiman, *The Wadsworth Statistics/Probability Series*, Wadsworth International Group, Belmont, CA, **1984**, p. 358.
- [35] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–58.
- [36] I. V. Tetko, *Neural Process Lett.* **2002**, *16*, 187–199.
- [37] I. V. Tetko, D. J. Livingstone, A. I. Luik, *J. Chem. Inf. Comp. Sci.* **1995**, *35*, 826–833.
- [38] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, *IEEE Trans. Neural Netw.* **2001**, *12*, 181–201.
- [39] A. K. Seewald, in *Proc. 19th Int. Conf. Mach. Learn.*, Morgan Kaufmann, Waltham, MA, **2002**, pp. 554–561.
- [40] J. B. Kruskal, *Proc. Am. Math. Soc.* **1956**, *7*, 48–50.
- [41] V. P. Solov'ev, I. Oprisiu, G. Marcou and A. Varnek, *Indust. Eng. Chem. Res.* **2011**, *50* (24), 14162–14167.
- [42] E. B. Wilson, M. M. Hilferty, *Proc. Natl. Acad. Sci. USA* **1931**, *17*, 684–688.
- [43] E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C.-C. Chen, M. Zwolak, K. C. Seavey, *Indust. Eng. Chem. Res.* **2006**, *45*, 4389–4415.
- [44] J. Wisniak, A. Tamir, *J. Chem. Eng. Data* **1985**, *30*, 339–344.
- [45] A. Chinikamala, *J. Chem. Eng. Data* **1973**, *18*, 322–325.

Received: January 16, 2012

Accepted: April 23, 2012

Published online: ■ ■ ■, 0000

4 Courbe d'équilibre vapeur-liquide Y=f(X)

4.1 Introduction

Comme nous l'avons présenté précédemment, la courbe d'équilibre vapeur-liquide (VLE) peut être représentée par deux courbes: la courbe d'ébullition ou de bulle (T_b en fonction de la composition en phase liquide) et la courbe de rosée (T_b en fonction de la composition en phase gazeuse). La Figure 2-3 montre l'obtention de la courbe d'équilibre $y=f(x)$ à partir d'une courbe d'équilibre $T_b=f(x)$: à une valeur donnée de température correspondent une valeur pour la composition en phase liquide (X) et une valeur pour la composition en phase gazeuse (Y), ce qui permet de tracer la courbe d'équilibre vapeur-liquide $y=f(x)$.

Le travail de ce projet a consisté à modéliser cette courbe d'équilibre vapeur-liquide $y=f(x)$ de façon directe et indirecte, avec trois méthodes d'apprentissage différents: ASNN, SVM et PLS.

4.2 Méthodologie

4.2.1 Données

Le jeu de modélisation réuni 224 mélanges représentés par 4748 points mesurés à pression atmosphérique, pris de la KDB [1]. Ces données ont été choisies de façon plus stricte par rapport au jeu de données utilisé pour la modélisation de la courbe d'ébullition: toutes les mélanges contiennent un nombre important de points (>10), les points d'une courbe qui sortaient de la tendance d'une courbe lisse ont été éliminés. De plus, chaque composé individuel doit se trouver au moins dans deux mélanges différents. Ce choix a été nécessaire pour la stratégie de validation "Mixtures Out" (cf. 4.2.6). Les mélanges sélectionnés contiennent les mêmes types de composés purs retrouvés dans la modélisation de la courbe d'ébullition: des hydrocarbures, esters, alcools, cétones, composés chlorés (cf. Annexe 11.8).

Un jeu de test contenant 17 nouveaux mélanges provenant aussi de la KDB[1] à été utilisé pour valider les modèles obtenus et le domaine d'applicabilité (cf. Annexe 11.9).

4.2.2 Modélisation directe

La modélisation directe est similaire à celle utilisée pour la modélisation de la courbe d'ébullition. Chaque point de la courbe est considéré comme un objet individuel. La différence entre plusieurs objets du même mélange se fait en multipliant les descripteurs des corps purs par la fraction molaire. (Figure 4-1).

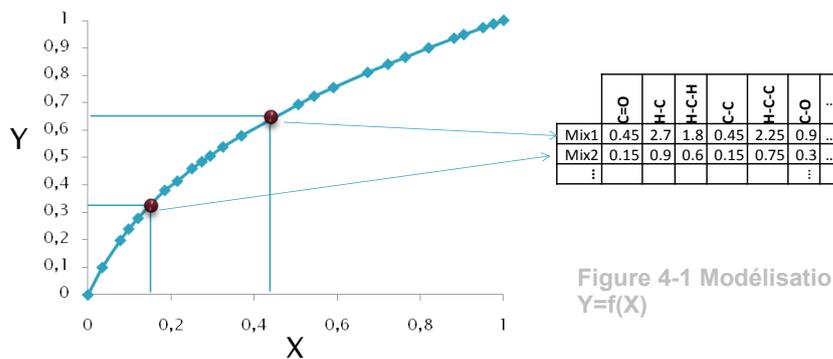


Figure 4-1 Modélisation directe de la courbe VLE, $Y=f(X)$

4.2.3 Modélisation indirecte

Cette modélisation implique plusieurs étapes. Tout d'abord, pour chaque mélange, les points (X,Y) sont ajustés avec un polynôme de 4^{ème} degré (4-1).

$$Y = a_1X + a_2X^2 + a_3X^3 + a_4X^4 \quad (4-1)$$

Les coefficients a_1 , a_2 , a_3 et a_4 **deviennent** les propriétés à modéliser. Une fois les valeurs de coefficients connues, l'équation (4-1) est appliquée afin de retrouver la valeur de Y. Cette modélisation est nommée "indirecte", parce que la valeur Y est prédite indirectement, à partir des coefficients du polynôme.

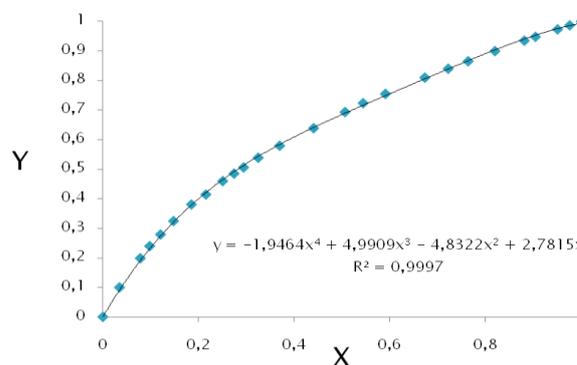


Figure 4-2 Ajustement des valeurs (X,Y) avec un polynôme de 4^{ème} degré.

L'ajustement avec un polynôme de 4^{ème} degré et l'obtention des coefficients a été fait à l'aide du logiciel R [2] pour 222 mélanges parce que pour deux d'entre eux (éthanol+hexylacétate et méthanol+hexylacétate) l'ajustement est très mauvais

indiqué par la valeur de F-statistic et le coefficient de corrélation ($F\text{-stat} < 300$ et $R^2 < 0.7$).

4.2.4 Matrice de descripteurs

Pour la modélisation d'Y l'ordre des composés est importante et par conséquent des matrices symétriques ne peuvent pas être créées. En effet il existe deux types de courbes, $y_1=f(x_1)$ et $y_2=f(x_2)$ selon que c'est le composé 1 ou 2 qui est étudié. Ces courbes ne sont pas indépendantes puisque l'une se déduit de l'autre par symétrie par rapport au point $(x,y)=(0.5,0.5)$. Toutefois, elles sont modélisées indépendamment.

Dans la matrice des descripteurs chaque mélange va être considéré deux fois: une fois A avec B et une fois B avec A, où A et B sont les constituants du mélange. Dans la Figure 4-3 on note toujours x_1 la composition en phase liquide et y_1 la composition en phase gazeuse de premier composé considéré (que ce soit A ou B).

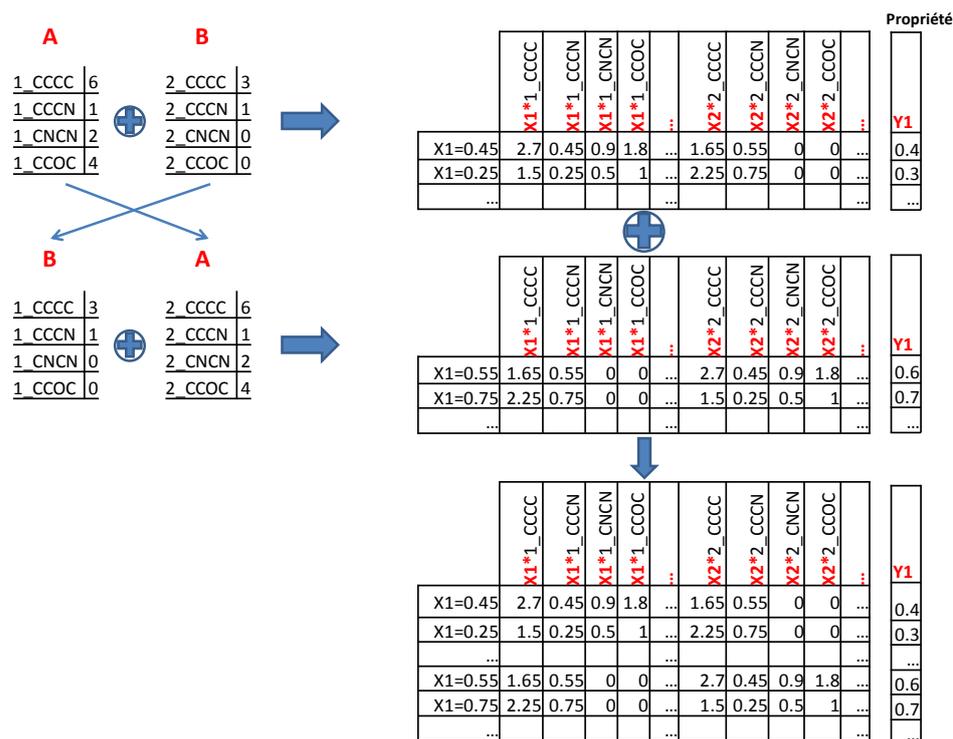


Figure 4-3 Construction de la matrice des descripteurs pour la modélisation directe de la courbe VLE $Y=f(X)$

Dans le cas de la modélisation indirecte chaque mélange est traité comme un tout, ce qui signifie que chaque mélange est représenté uniquement par deux vecteurs de descripteurs qui dépendent de l'ordre des composés mais pas de la composition (Figure 4-4).

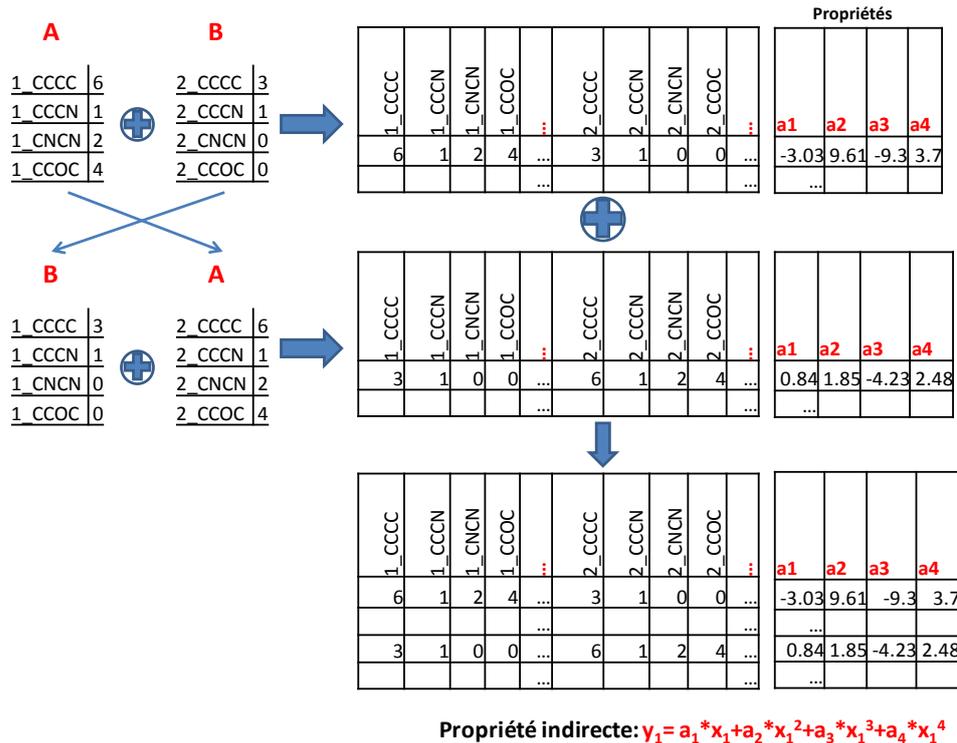


Figure 4-4 Construction de la matrice des descripteurs pour la modélisation indirecte de la courbe VLE Y=f(X)

Pour cette modélisation les 13 types de fragmentation ont été choisis avec les paramètres n_{min} égal à 2 et n_{max} égal à 10. Ce choix a été fait suite à l'observation préalable que ces paramètres donnent des bons résultats pour tous les types des fragments (Chapitre 3).

4.2.5 Machines d'apprentissage

La modélisation indirecte a été effectuée avec le paquet PLS [3] implémenté dans R. C'est une méthode qui permet d'obtenir plusieurs sorties simultanément, dans notre cas, il s'agit des valeurs des quatre coefficients. La modélisation directe, point par point, a été faite avec les méthodes SVM et ASNN. La modélisation PLS a aussi été utilisée dans ce cas pour comparer les deux stratégies, indirecte et directe.

La méthode SVM a été utilisée avec un noyau RBF pour lequel les valeurs de C, γ et ϵ ont été optimisés au préalable en fonction de la stratégie de validation.

La méthode PLS a été utilisée avec un nombre des variables latentes égal à 40, valeur qui donne un bon compromis qualité prédictive et surapprentissage

4.2.6 Validation croisée

La validation croisée est similaire à celle utilisée pour la modélisation de la courbe de bulle.

La stratégie "Points out" consiste à faire de 2-CV sur l'ensemble des points. La totalité des points est divisé aléatoirement en deux sous-ensembles de même taille (Figure 4-5). Chaque sous-ensemble est à son tour jeu de test et d'apprentissage, respectivement. Cette stratégie est applicable seulement dans la modélisation point par point, c'est-à-dire la modélisation directe.

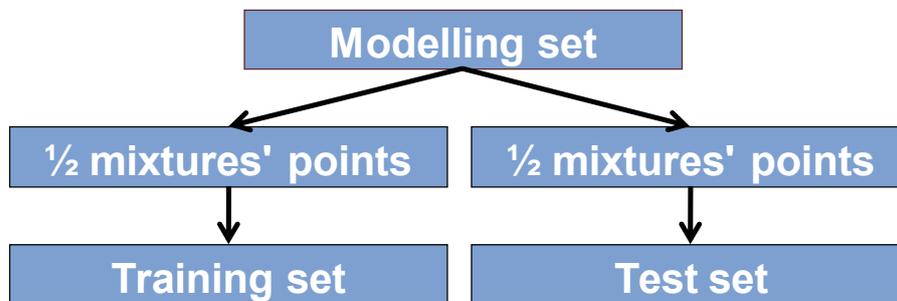


Figure 4-5 Stratégie de cross-validation "Points out" pour la modélisation de la courbe VLE $Y=f(X)$

La stratégie "Mixtures out" suit en principe la stratégie de 5-CV: l'ensemble des mélanges est divisé en 5 paquets, et chacun des paquets est utilisé à son tour comme jeu de test, pendant que les autres 4 paquets servent pour l'entraînement. De cette façon chaque mélange est prédit une fois. La différence par rapport à la 5-CV classique consiste dans la supervision de la division du jeu de données initial. Afin de garder toujours une information sur tous les composés purs de jeu de données initial, lors de choix de jeu d'apprentissage et de test, chaque composé individuel doit se trouver au moins une fois dans le jeu d'apprentissage. Par exemple, si un composé pur se trouve que dans deux mélanges, quand un de ces deux mélanges se trouve dans le jeu de test, l'autre doit se trouver impérativement dans le jeu d'apprentissage.

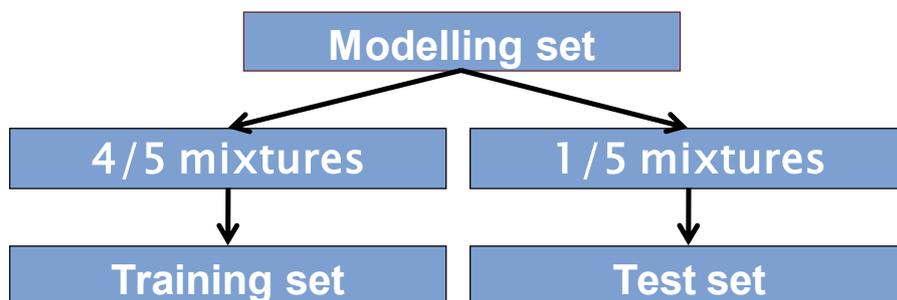


Figure 4-6 Stratégie de cross-validation "Mixtures Out" pour la modélisation de la courbe VLE $Y=f(X)$

La stratégie la plus exigeante, "Compounds out" consiste à diviser l'ensemble des composés purs du jeu de données initial en 10 paquets. Chaque paquet contient plusieurs composés purs avec tous les mélanges dont ils font partie. Plus précisément dans chaque paquet tous les mélanges contiennent au moins un des

composés purs "nouveau" (qui ne se trouve pas dans les autres paquets). La stratégie de validation croisée est toujours la même: chaque paquet sert à son tour de jeu de test, tandis que les autres 9 servent de jeu d'apprentissage.

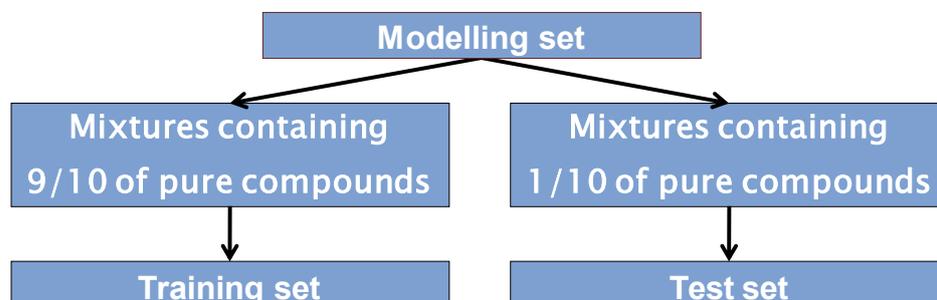


Figure 4-7 Stratégie de cross-validation "Compounds out" pour la modélisation de la courbe VLE $Y=f(X)$

La division du jeu de données initial, est supervisée de façon à assurer une distribution équitable des mélanges dans les 10 paquets : il faut toujours avoir entre 42 et 48 mélanges. Comme il a été expliqué dans le chapitre 3, concernant la modélisation de la courbe d'ébullition, le nombre total des mélanges en validation croisée est le double de nombre des mélanges dans le jeu de modélisation. Par conséquent 448 courbes divisées en 10 paquets sont considérées.

Chaque mélange est représenté par deux courbes prédites, mais ces courbes sont reliées.(cf. 4.2.4) Ces prédictions sont donc moyennées chaque fois, afin de prendre en compte cette relation.

Toutes ces stratégies sont répétées trois fois, ainsi les prédictions sont calculées en moyennant les valeurs obtenues pour chaque itération.

Toutes les courbes obtenues doivent passer par les points (0,0) et (1,1). Pour les modèles PLS (indirect) pour une valeur $X=0$ on obtient toujours $Y=0$, par construction. Dans toutes les cas les courbes prédites sont normalisées pour prendre en compte la contrainte $f(1)=1$.

4.2.7 Validation externe

Étant donné que le jeu de validation pour la modélisation de la courbe d'ébullition a été englobé dans le jeu d'apprentissage (totalisant 224 mélanges), 17 nouveaux systèmes, 5 contenant les deux composés connus (déjà existants dans le set d'apprentissage), 6 contenant un composé connu et un inconnu et 6 contenant que des composés totalement inconnus, ont été utilisés pour la validation.

4.2.8 *Benchmark*

Le benchmark a consisté en la comparaison des prédictions des modèles développés au cours de ce travail avec ceux obtenus avec COSMO-RS [4]. Ceci nous permet de regarder avec un œil plus critique les résultats obtenus.

Toutefois, les σ -profiles nécessaires pour COSMO-RS, n'étaient disponibles que pour seulement 220 mélanges du jeu d'apprentissage et 13 systèmes du jeu de test. Afin d'être cohérent, seuls ces mélanges ont été utilisés pour la comparaison.

4.3 Résultats

4.3.1 *N-Cross-validation*

Les valeurs statistiques pour les deux types de modélisation sont indiquées dans le Tableau 4-1 ci-dessous. Conforme aux attentes, les performances des modèles décroissent dans l'ordre "Points out", "Mixtures Out" et "Compounds Out". Plus la validation est stricte, moins les performances prédictives sont bonnes. Néanmoins, la méthode PLS par modélisation indirecte est légèrement plus performante que les autres méthodes. De plus, la PLS donne de meilleurs résultats en modélisation indirecte que directe où ses performances chutent considérablement par rapport aux autres méthodes. En prenant en compte les prédictions donnés par les modèles ASNN, SVM et PLS* (indirect), des courbes ont été calculées en utilisant la méthode de stacking [5]. Ces résultats sont meilleurs que ceux de chaque modèle individuel comme le montre le Tableau 4-1.

4.3.2 *Validation externe*

Le Tableau 4-2 réunit les résultats de la prédiction sur le test set contenant 17 nouveaux mélanges. Les mélanges ont été séparés en trois groupes afin de mieux comprendre les capacités prédictifs des modèles développés. Comme déjà constaté dans le chapitre précédent, les modèles développés sont très prédictifs pour des nouveaux mélanges contenant uniquement des composés connus. Si un mélange contient des composés inconnus, la qualité prédictive chute. Ces mauvaises performances ne se retrouvent pas en cross-validation pour la stratégie "Compounds Out", ce qui signifie, probablement, que ces modèles obtenus par cette stratégie sont sur-appris ou indiquent que les modèles sont employés pour extrapolation des données. "Mixtures out" est la stratégie qui donne les mêmes performances en cross-validation et en validation externe (prédictions pour L1). On ne peut exclure

des possibles erreurs dans les données expérimentales. Enfin, il faut relativiser la discussion au regard de faible nombre de mélanges dans le jeu de données externe.

Tableau 4-1 Résultats N-CV pour la modélisation de la courbe VLE Y=f(X) pour 224 mélanges

*Modélisation indirecte;
Le modèle de stacking est basé sur les prédictions des modèles ASNN, SVM et PLS*.

	Points out			Mixtures Out			Compounds Out								
	SVM	ASNN	PLS	SVM	ASNN	PLS	Stacking	SVM	ASNN	PLS	Stacking	SVM	ASNN	PLS	Stacking
R²	0.95	0.99	0.75	0.91	0.94	0.74	0.93	0.95	0.88	0.90	0.72	0.91	0.92	0.92	0.92
RMSE	0.07	0.04	0.15	0.09	0.07	0.15	0.08	0.07	0.10	0.10	0.16	0.09	0.08	0.08	0.08
MAE	0.05	0.02	0.11	0.06	0.04	0.11	0.05	0.04	0.07	0.07	0.11	0.06	0.06	0.06	0.06

Tableau 4-2 Prédiction de jeu de test pour la modélisation de la courbe VLE Y=f(X) pour 17 mélanges.

Modélisation indirecte; L1 - les 2 composés du mélange sont connus (5 mélanges); L2 - un des composés du mélange est inconnu (6 mélanges); L3 – les 2 composés des mélanges sont inconnus (6 mélanges). Le modèle de stacking est basé sur les prédictions des modèles ASNN, SVM et PLS.

	L1			L2			L3			L1+L2+L3										
	SVM	ASNN	PLS	SVM	ASNN	PLS	SVM	ASNN	PLS	SVM	ASNN	PLS	SVM	ASNN	PLS	Stacking	SVM	ASNN	PLS	Stacking
R²	0.88	0.96	0.70	0.95	0.97	0.39	0.62	0.45	0.57	0.60	0.67	0.71	0.55	0.85	0.84	0.62	0.75	0.56	0.77	0.78
RMSE	0.11	0.06	0.17	0.07	0.05	0.28	0.21	0.26	0.23	0.22	0.18	0.17	0.20	0.12	0.12	0.20	0.17	0.22	0.16	0.15
MAE	0.08	0.04	0.12	0.05	0.04	0.20	0.15	0.20	0.15	0.15	0.14	0.13	0.17	0.09	0.10	0.14	0.11	0.16	0.10	0.10

4.3.3 Domaine d'applicabilité

Afin de comprendre la raison de l'impossibilité de prédire certains mélanges, plusieurs DA ont été étudiées. Le DA Fragment Control permet d'améliorer très légèrement les résultats. Cela signifie que le simple Fragment Control n'est pas suffisant pour les mélanges binaires. Pour cette raison, le DA plus strict, nommé Fragment Control IVAB a été utilisé (cf. 1.5.1). Ceci est très restrictif et par conséquent seulement deux mélanges, en plus de ceux de L1, vont être prédits. Néanmoins, les résultats obtenus sont très satisfaisants (Tableau 4-3).

Tableau 4-3 Prédictions de jeu de test avec DA=FrgCtrl IVAB pour la modélisation de la courbe VLE Y=f(X) pour sept mélanges.

* Modélisation indirecte; Le modèle de stacking est basé sur les prédictions des modèles ASNN, SVM et PLS*

	SVM	ASNN	PLS	PLS*	Stacking
R ²	0.84	0.96	0.68	0.94	0.92
RMSE	0.12	0.06	0.18	0.07	0.08
MAE	0.09	0.04	0.13	0.06	0.06

4.3.4 Benchmark

Le Tableau 4-4 montre les résultats obtenus pour la prédiction de 220 mélanges appartenant au jeu de données d'apprentissage, pour lesquels le modèle COSMO-RS a pu être utilisé. Les prédictions COSMO-RS peuvent rentrer dans le cadre de la stratégie "Compounds out", pour cette raison les résultats ont été comparés avec ceux donnés par nos modèles pour cette stratégie.

Tableau 4-4 Benchmark pour le jeu d'apprentissage (220 mélanges).

Le modèle de stacking est basé sur les prédictions des modèles ASNN, SVM et PLS*.

	SVM	ASNN	PLS	PLS*	Stacking	COSMO
R ²	0.89	0.90	0.73	0.91	0.95	0.70
RMSE	0.10	0.09	0.16	0.09	0.07	0.17
MAE	0.07	0.07	0.11	0.06	0.04	0.12

Les résultats obtenus en CV sont nettement meilleurs par rapport à ceux obtenus avec COSMO-RS. La Figure 4-8 montre trois exemples de prédiction de la courbe d'équilibre vapeur-liquide Y=f(X).

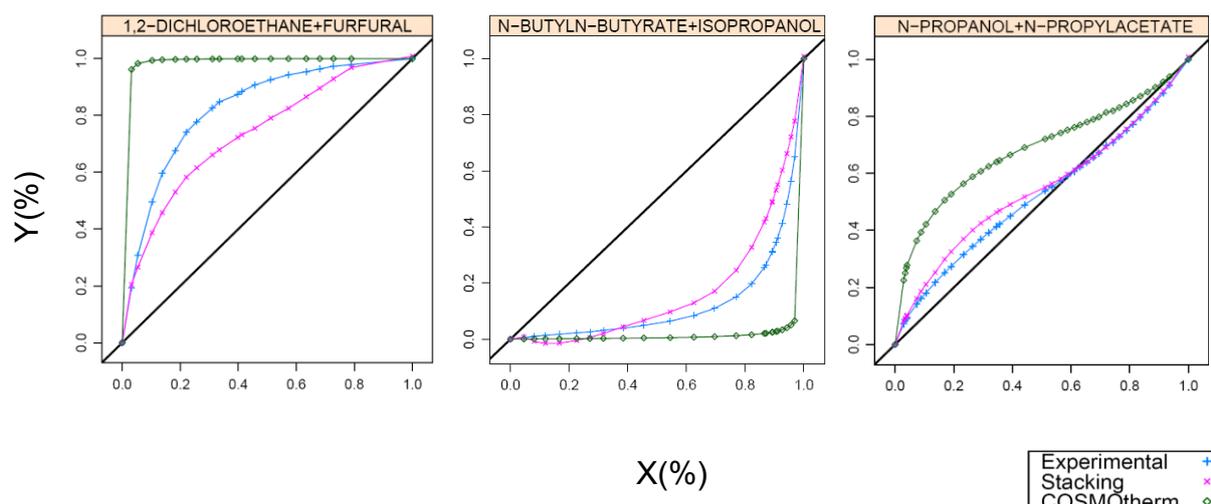


Figure 4-8 Exemple de prédictions de la courbe VLE $Y=f(X)$. Comparaison des prédictions de nos modèles avec ceux de COSMO-RS.

Dans le Tableau 4-5 les prédictions données par nos modèles ont été comparées avec des prédictions données par les modèles COSMO-RS, pour le jeu de données de validation. Il est intéressant de noter que COSMO-RS a les mêmes difficultés pour prédire les courbes d'équilibre $y=f(x)$ que nos modèles sur le jeu de test. Dans ce cas aussi les modèles ASNN et PLS* se révèlent supérieures.

	SVM	ASNN	PLS	PLS*	Stacking	COSMO
R²	0.58	0.77	0.57	0.75	0.77	0.64
RMSE	0.22	0.16	0.22	0.17	0.16	0.20
MAE	0.15	0.10	0.16	0.10	0.10	0.13

Tableau 4-5 Benchmark pour le jeu de test (13 mélanges)

* Modélisation indirecte ; Le modèle de stacking est basé sur les prédictions des modèles ASNN, SVM et PLS*.

Une deuxième comparaison a été faite sur l'ensemble de mélanges inclus dans le DA des modèles développés (Tableau 4-6). Cette fois-ci l'écart entre les performances de nos modèles (6% en stacking) et celle de COSMO-RS (17%) est beaucoup plus important. La Figure 4-9 sont tracées les courbes VLE pour ces 7 mélanges.

Tableau 4-6 Prédictions de la courbe VLE Y=f(X) pour le jeu de test en utilisant le DA=FrgCtrl IVAB (7 mélanges)

* Modélisation indirecte ; Le Stacking prend en compte les prédictions des modèles ASNN, SVM et PLS*.

	SVM	ASNN	PLS	PLS*	Stacking	COSMO
R ²	0.84	0.96	0.68	0.94	0.97	0.42
RMSE	0.12	0.06	0.18	0.07	0.06	0.24
MAE	0.09	0.04	0.13	0.06	0.04	0.17

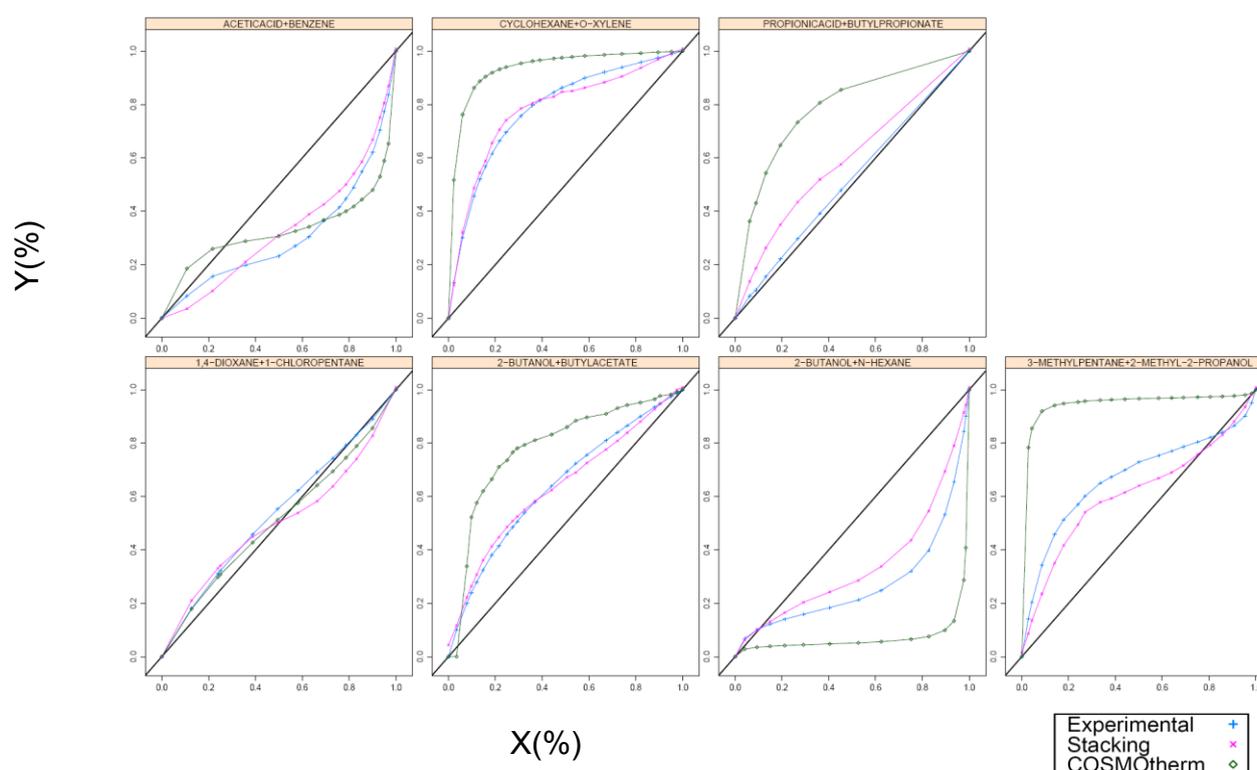


Figure 4-9 Prédictions de la courbe VLE Y=f(X) pour le jeu de test avec DA=FrgCtrl IVAB (7 mélanges).

4.4 Conclusion

Ce travail s'est concentré sur la prédiction de la courbe d'équilibre vapeur-liquide Y=f(X), en partant d'un jeu de données de 224 mélanges. Deux types de modélisation ont été étudiés: directe en utilisant les modèles ASNN, SVM et PLS et indirecte en utilisant les modèles PLS. Les modèles ont des performances comparables, à l'exception de PLS en modélisation directe. Ceci montre que les modèles linéaires sont utiles que pour une modélisation indirecte.

La modélisation indirecte d'avère être une bonne approche: les résultats obtenus par la modélisation indirecte sont très souvent meilleures que ceux obtenus par la modélisation directe.

Nous avons démontré aussi que les modèles obtenus peuvent être meilleurs que COSMO-RS, outil commercial souvent utilisé pour la prédiction des courbes VLE. Nos modèles peuvent être applicables à des mélanges contenant des composés purs connus où très semblables structurellement à ceux contenus dans le jeu d'apprentissage. Dans ce cas les modèles ont de bonnes performances i.e. pour les modèles de stacking la valeur de RMSE est de 6%.

Dans le cas de la modélisation "Compounds Out" l'erreur est de 10% ce qui montre que les modèles de stacking développés peuvent être utilisés même pour des nouveaux mélanges contenant des composés inconnus ce qui n'était pas le cas pour les modèles de la courbe d'ébullition ($T_b=f(x)$). Ces performances peuvent être expliquées par une meilleure qualité des données, les mélanges ayant été choisis plus rigoureusement. De plus la stratégie de modélisation prend en compte chaque mélange deux fois ce qui produit deux prédictions par mélange qui sont ensuite moyennés ce qui a un effet de compensation d'erreur. De plus, les courbes ont été ajustées afin de tenir compte respecter la contrainte que chaque courbe doit passer par les points (0,0) et (1,1), se qui améliore encore les prédictions.

4.5 Références

1. Kang, J.W., et al., *Development and current status of the Korea Thermophysical Properties Databank (KDB)*. International Journal of Thermophysics, 2001. **22**(2): p. 487-494.
2. RDevelopmentCoreTeam, *R: A language and environment for statistical computing* 2004, R Foundation for Statistical Computing: Vienna, Austria.
3. Mevik, B.H. and R. Wehrens, *The pls package: Principal component and partial least squares regression in R*. Journal of Statistical Software, 2007. **18**(2).
4. Klamt, A. and F. Eckert, *COSMOtherm, a powerful tool for the calculation of solvation effects and phase equilibria*. Abstracts of Papers of the American Chemical Society, 2000. **220**: p. U234-U234.
5. Seewald, A.K., *How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness*, in *Proceedings of the Nineteenth International Conference on Machine Learning*. 2002, Morgan Kaufmann Publishers Inc. p. 554-561.

5 Classification azéotrope/zéotrope

Le but de ce travail a été de développer des modèles capables de prédire la formation d'un azéotrope dans un mélange liquide binaire sans avoir besoin des données expérimentales pour ceci.

5.1 Revue

Jusqu'à présent peu de travaux ont été fait sur la prédiction d'un azéotrope.

En général une méthode mathématique puissante est combinée avec des équations d'état ou des expressions de l'énergie libre d'excès (NRTL, UNIFAC, Wilson...) afin de résoudre les équations décrivant la condition d'azéotropie. Stadtherr [1] présente une stratégie basée sur des méthodes d'analyse de Newton par intervalle avec la bisection généralisée (IN/GB)[2]. Cette méthode mathématique permet d'identifier toutes les racines de l'équation décrivant la condition d'azéotropie dans un mélange par des modèles de composition locale (Wilson, NRTL and UNIQUAC). Ces racines représentent les points azéotropiques homogènes de mélange.

Le travail de Salomone et Espinosa[3] représente une amélioration de celui présenté par Stadtherr [1] avec l'ajout de la théorie d'indice topologique de Zharov-Serafimov[4].

Dong et son équipe [5] propose quatre approches basés sur le modèle UNIFAC, afin de prédire les azéotropes binaires homogènes sans aucune donnée expérimentale. Le critère thermodynamique des quatre approches est la première dérivée de la pression par rapport à la composition qui est égale à zéro au point azéotrope. Si la phase vapeur du mélange binaire est considérée comme un gaz parfait, seul le modèle UNIFAC est utilisé, sinon, le modèle UNIFAC, l'équation d'état Peng-Robinson, et une des règles de mélange (HV, HVOS, MHV1) sont combinés pour calculer l'équilibre de phase. Le signe positif ou négatif de la dérivée seconde de la pression par rapport à la composition, indique le type d'azéotrope (positif ou négatif). Une critique sur ce travail est que l'affirmation de ne pas avoir utilisé des données expérimentales est fausse, car l'équation d'état utilise des valeurs expérimentales (paramètres critiques T_c , P_c , V_c). De plus, les coefficients d'Antoine utilisés pour le calcul de la pression de vapeur saturée proviennent des valeurs expérimentales. Même si ces valeurs expérimentales n'ont pas été mesurées par Dong et son équipe elles sont incluses dans les calculs, ce que signifie que pour un

nouveau mélange pour lequel ces valeurs ne sont pas disponibles le calcul ne peut pas être réalisé. De plus la théorie a été validée sur seulement 8 systèmes binaires réfrigérants ce qui est faible pour un jeu de validation.

Le travail de Lee et Kim [6] présente un critère simple de formation d'un réfrigérant binaire azéotropique basé sur la théorie des solutions régulières[7]. Selon Lee et Kim ce critère ne nécessite aucun calcul complexe et peut être facilement appliqué lorsque la pression de vapeur et des données de densité saturée de composants purs sont disponibles. Ce critère a été appliqué à seulement trois mélanges de réfrigérants ce qui ne démontre pas vraiment son efficacité et son utilité.

AZEOPERT [8-10] est un logiciel développé pour la prédiction de la présence et du type d'azéotrope dans un mélange binaire. Le fonctionnement d'AZEOPERT se base sur plusieurs niveaux en commençant par une recherche dans la base de données contenant environ 20000 données azéotropiques incluant les caractéristiques : température, pression et composition. Si cette requête ne donne pas de résultats des règles heuristiques sont appliqués. Ces règles sont formulées à partir des études de cas sur la base de données azéotropiques. Les composés sont classés dans des séries d'homologues dont le comportement azéotropique est considéré être le même. Si cette classification n'est pas possible des méthodes numériques (NRTL, UNIQUAC, UNIFAC) sont utilisées.

Le type azéotropique (négatif/positif) peut être estimé en tenant compte des températures d'ébullition des corps purs. La température de l'azéotrope est toujours inférieure (azéotrope positif) ou supérieure (azéotrope négatif) à la température d'ébullition de ses composants. Si le logiciel ne peut pas vérifier cette condition, en règle générale l'azéotrope est considéré positif et ensuite des règles heuristiques sont vérifiées pour savoir si l'azéotrope est négatif.

Comme on peut le constater cette approche utilise une très grande base de données contenant de nombreuses valeurs expérimentales sur des corps purs mais aussi sur les coordonnées azéotropiques. AZEOPERT semble pouvoir donner des résultats intéressants, même si aucun résultat de validation n'a pas été présenté. De plus, aucune trace de ce logiciel n'est trouvée après 1999, ce que nous laisse douter de son utilisation.

Une méthode QSPR de prédiction de l'existence d'un azéotrope est présentée dans [11]. Le modèle utilise le réseau de neurones afin de corréliser les données

azéotropiques et uniquement des propriétés de différents composants purs du mélange (huit au total). Un jeu de données de 490 mélanges a été divisé aléatoirement en trois jeux de données: d'apprentissage, de test et de validation. L'ordre des composés est important, c'est pour cette raison que le premier composé du mélange sera toujours celui qui est le plus volatil. Étant donné que la sortie n'est pas une valeur binaire, deux seuils ont été utilisés pour les valeurs prédites: si la valeur est inférieure à 0.2 le mélange est considéré comme zéotrope et si elle est supérieure à 0.8 le mélange est considéré comme étant un azéotrope. Entre ces deux valeurs aucune prédiction n'est fournie. Les résultats présentés pour le jeu de validation de 42 mélanges sont très bons avec deux erreurs et quatre mélanges non prédits. Un important problème avec ce travail est l'utilisation d'un seul jeu de validation au lieu d'une validation croisée. Tous les mélanges contenant l'eau, le chloroforme ainsi que les azéotropes hétérogènes et les azéotropes homogènes négatifs ont été placés dans le jeu d'apprentissage, ce qui améliore largement la qualité de la prédiction. Enfin, le fait de s'assurer que les données des jeux de test et de validation sont dans les limites des variables d'entrée du jeu d'entraînement (bonding box), favorise l'élimination des mélanges difficiles.

5.2 Méthodologie

5.2.1 Données

Le jeu de modélisation contient 400 mélanges binaires mesurés à la pression atmosphérique provenant de la KDB[12] et de la compilation de Horsley [13]. Le ratio zéotrope/azéotrope dans le jeu d'apprentissage est de 1, afin d'être cohérent avec la réalité. Les mélanges sélectionnés contiennent les mêmes types de composés purs utilisés pour les deux modélisations antérieures: des hydrocarbures, esters, alcools, cétones, composés chlorés (cf. Annexe 11.3).

Un premier jeu de test (**VS1**) de 96 mélanges (22 azéotropes et 74 de zéotropes) provenant des mêmes sources, a été utilisé. Ce jeu de données contient que des composés purs se trouvent déjà dans le jeu de données de modélisation (cf. Annexe 11.4).

Un deuxième jeu de test (**VS2**) contenant 499 mélanges a été fourni par la Société Processium [14]. Ce jeu de données est très divers, contenant des composés très différents structurellement par rapport au jeu d'apprentissage

(composés nitro, amides, heterocycles, composés iodés et soufrés, pyridine, etc.)(cf. 11.5).

5.2.2 Matrice de descripteurs

Différents types de descripteurs ISIDA sont utilisés pour la modélisation: séquences d'atomes (IA), séquences d'atomes et de liaisons (IAB), atomes unis (IIIA, IIIAB, IIIB, IVA, IVAB, IVB) avec les paramètres n_{min} variant entre 2 et 4 et le paramètre n_{max} compris entre n_{min} et 15.

Les atomes d'hydrogène sont pris en compte explicitement, parce que leur contribution au comportement azéotropique d'un mélange n'est pas négligeable.

En plus de descripteurs ISIDA les descripteurs générés par MOE[15] ont été utilisés pour cette étude, afin de comparer leur efficacité. Parmi les descripteurs 2D, 61 ont été choisis (cf. Annexe 11.1).

La matrice de descripteurs pour un mélange a été obtenue de la manière suivante: Après le calcul des descripteurs pour chaque composé individuel, les deux matrices résultantes sont combinées afin d'obtenir une matrice symétrique (Figure 5-1).

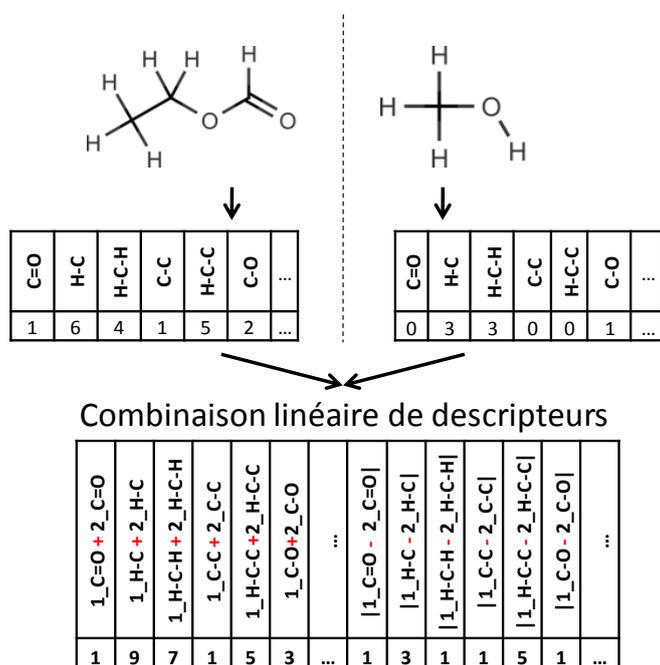


Figure 5-1 Calcul de la matrice de descripteurs pour des mélanges pour la classification

5.2.3 Machines d'apprentissage

Pour cette étude plusieurs machines d'apprentissage de classification ont été utilisées: Support Vector Machine (SVM), le Perceptron Votant et Random Forest (RF).

Pour la SVM le noyau Tanimoto $k_t(\vec{x}_i, \vec{x}_j)$ entre deux instances (vecteurs de descripteurs) \vec{x}_i et \vec{x}_j a été implémenté dans la logiciel *libsvm* version 2.8 et utilisé. Le produit scalaire est noté $\langle \cdot \rangle$.

$$k_t(\vec{x}_i, \vec{x}_j) = \frac{\langle \vec{x}_i, \vec{x}_j \rangle}{\langle \vec{x}_i, \vec{x}_i \rangle + \langle \vec{x}_j, \vec{x}_j \rangle - \langle \vec{x}_i, \vec{x}_j \rangle} \quad (5-1)$$

L'expression (5-1) est un noyau si les composantes des vecteurs sont positifs ou nuls comme cela a été montré auparavant [16].

Le Perceptron Votant (VP) a été présenté dans la partie théorique (cf. 1.2.3). Le principal avantage de VP est que l'algorithme ne dépend pas explicitement d'un paramètre. Le temps d'apprentissage (le nombre d'époque) est choisi avant mais l'algorithme peut être arrêté lorsque l'erreur d'apprentissage se stabilise. Dans notre cas le nombre d'époques a été de 100, car cette valeur suffit pour arriver à stabilisation de l'erreur.

Pour la classification avec Random Forest logiciel la version 3-6-4 de Weka a été utilisé et lancé en batch. 100 forêts ont été calculées et un consensus a été obtenu.

5.2.4 Validation croisée

Pendant cette étude, la robustesse de modèles a été évaluée en utilisant la validation croisée (5 paquets).

Pour la validation croisée utilisant les descripteurs MOE, pour chaque paquet le jeu de test à été normalisé en utilisant les paramètres (moyenne et variance) obtenus sur le jeu d'entraînement du même paquet.

5.2.5 Benchmark

Une étude de benchmark a été effectuée en comparant les résultats de nos modèles avec ceux donnés par les modèles UNIFAC-Dortmund, UNIFAC et COSMO-RS. Les calculs UNIFAC ont été réalisés avec AspenPLUS® v.7.1., tandis que les calculs COSMO ont été faits à l'aide de COSMOtherm v.C21_0108. L'UNIFAC-Dortmund[17] représente une mise-à-jour de l'UNIFAC original, en introduisant des paramètres dépendants de la température. Ainsi, 404 paramètres d'interaction ont été ajoutés ou modifiés dans la matrice originale UNIFAC contenant 635 paramètres dans la dernière publication[18].

5.3 Résultats

Des modèles individuels pour chaque type de fragment ont été développés codant la propriété 0 s'il s'agit d'un zéotrope ou 1 pour un azéotrope. Ensuite les meilleurs modèles, avec une valeur de BA supérieure à 0.75 ont été choisis pour le modèle consensus. Les 30 modèles sélectionnés votent à la majorité pour obtenir la prédiction finale. En cas d'égalité, le mélange ne peut pas être prédit.

5.3.1 Validation croisée

Les résultats de la cross-validation sont présentés dans le Tableau 5-1. Les deux méthodes de calculs SVM et VP ont des performances prédictives très similaires. De même, les matrices de confusion sont similaires (Tableau 5-2). Dans les deux cas les modèles prédisent plus facilement un zéotrope comme étant un azéotrope, qu'à l'inverse, ce qui signifie que les modèles sont plus précis pour retrouver les mélanges zéotropiques.

	SVM	VP
BA	0.82	0.82
ROC AUC	0.84	0.84
Recall (Z)	0.78	0.77
Recall (A)	0.85	0.86

Tableau 5-1 Résultats en 5-CV pour la classification azéotrope/zéotrope

		SVM		VP	
		Actuel			
Prédit		A	Z	A	Z
	A	170	44	172	46
	Z	30	156	28	154

Tableau 5-2 Matrices de confusion pour la classification azéotrope/zéotrope obtenus par validation croisée sur le jeu de 400 mélanges.

Dans la Figure 5-2 sont représentés les courbes ROC pour la validation croisée. Les courbes obtenues par les deux méthodes d'apprentissage sont similaires, ainsi que leur aire sous la courbe ROC (AUC). Cette valeur de 0.84 correspond à un critère de précision de la prédiction.

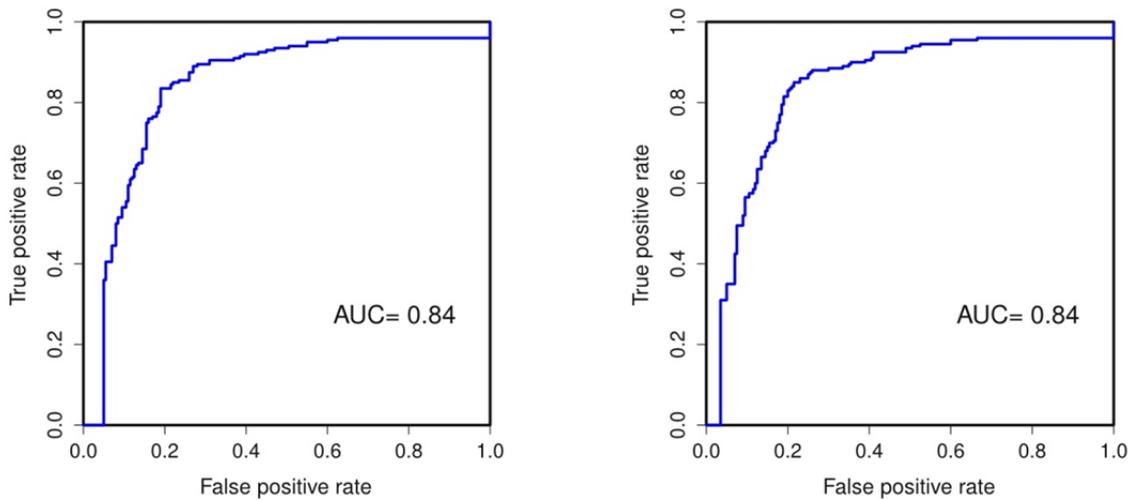


Figure 5-2 Représentation des courbes ROC pour SVM (gauche) et VP (droit) en 5-CV.

170 données expérimentales utilisées ont été recueillies à partir des courbes VLE. Une analyse de ces courbes indique que pour 13% des mélanges, environ, il est difficile d'établir le comportement azéotropique, parce que le point azéotropique, s'il existe, se trouve très proche de $X=1$ (Figure 2-11). Dans ces cas les informations trouvées dans la littérature sont souvent contradictoires. Cette incertitude expérimentale de 13% est proche de l'erreur de prédiction de nos modèles, ce qui est cohérent, car la prédictivité est toujours limitée par l'erreur expérimentale.

Les résultats présentés ont été obtenus en utilisant les descripteurs fragmentaux ISIDA (SMF). Comme nous l'avons annoncé (5.2.2) des modèles basés sur de descripteurs MOE ont été développés également. De plus, l'approche SVM a été comparée avec l'approche RF. Le Tableau 5-3 englobe les valeurs de BA obtenus avec les descripteurs SMF, MOE et leur combinaison avec l'approche RF et SVM. Nous constatons que les descripteurs SMF conduisent à des résultats semblables indépendamment de l'approche utilisée. Les descripteurs MOE seuls sont moins performants, tandis que en combinaison avec les descripteurs SMF donnent de meilleurs résultats avec l'approche RF. La différence n'est pas importante par rapport aux descripteurs SMF seuls. Cette analyse montre que les descripteurs SMF sont suffisants pour ce type de modélisation, et plus généralement pour la modélisation des propriétés des mélanges. Cela justifie le choix d'utiliser ces descripteurs pour tous les développements de cette thèse.

Tableau 5-3 BA en fonction de méthodes d'apprentissage et des descripteurs.

	RF	SVM
SMF(consensus)	0.83	0.82
MOE	0.80	0.79
SMF+MOE (consensus)	0.84	0.79

5.3.2 Y-Randomization

Cette procédure a été inspirée par les travaux de Rücker[19]. Les propriétés des tous les mélanges ont été réattribués aléatoirement aux mélanges. Ensuite des modèles ont été refaits en validation croisée. Les performances de ces modèles doivent se trouver alentour de $BA=0.5$ ce qui correspond au pur hasard. La Figure 5-3 représente les valeurs de BA pour les modèles résultant d'Y-randomization et les modèles développés sur le jeu original. Nous constatons que même le meilleur modèle obtenu après Y-randomization ($BA=0.6$) est moins bon que le pire modèle obtenu sur le jeu de données original ($BA=0.62$), c'est-à-dire il n'y a pas de superposition de ces deux résultats. Cela montre que les modèles développés sur le jeu de données original ne sont pas fortuits et qu'il y a bien eu un apprentissage.

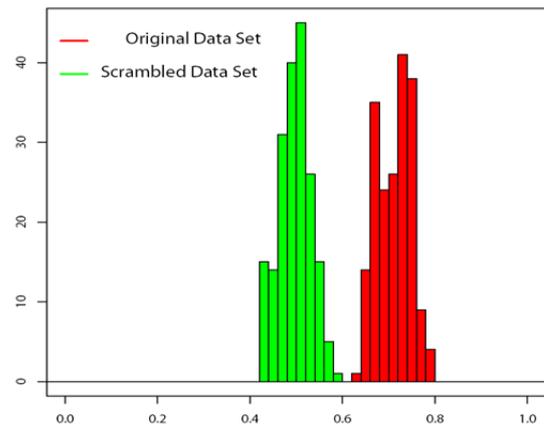


Figure 5-3 Comparaison entre les modèles obtenus par Y-randomization et le jeu de données original.

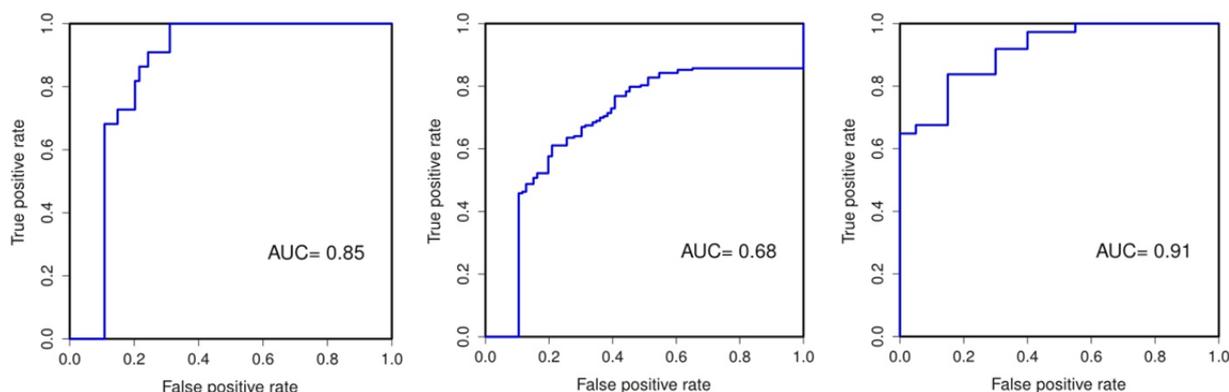
5.3.3 Validation externe

La validation externe a été faite sur deux jeux de données VS1 et VS2. Les résultats provenant de consensus sont regroupés dans Tableau 5-4. Des valeurs très bonnes ($BA>0.82$) sont obtenus pour VS1 et pour VS2 ne contenant que des composés connus (« Mixtures Out »). Par contre en considérant les mélanges de VS2 contenant au moins un composé connu, les performances obtenues sont décevantes. De plus, l'utilisation d'un DA n'améliore pas les résultats pour le VS2 entier ou contenant au moins un composé connu.

Tableau 5-4 Validation VS1 et VS2. * au moins un composé connu ; ** Mixtures Out; sDA- sans DA ;DA – Fragment Control.

	VS1	VS2		VS2*		VS2**
		sDA	DA	sDA	DA	
Mélanges	96	499	349	401	289	57
BA	0.83	0.59	0.64	0.62	0.67	0.84
Recall(-1)	0.74	0.57	0.64	0.55	0.70	0.85
Recall(1)	0.91	0.62	0.63	0.69	0.64	0.84

La Figure 5-4 représente les courbes ROC pour la validation de VS1 (*gauche*) et VS2 contenant un ou deux composés purs connus respectivement (*milieu et droite*). Ceci est une autre représentation des résultats montrés dans le Tableau 5-4. De très bons résultats sont obtenus pour la prédiction des mélanges contenant que des composés connus.

Figure 5-4 Courbes ROC pour la VS1 (*gauche*) et VS2 contenant respectivement un ou deux composés purs connus (*milieu et droite*).

5.3.4 *Domaine d'applicabilité*

Comme nous l'avons vu dans le Tableau 5-4 les seuls mélanges qui sont bien prédits sont ceux qui contiennent des composés connus, composés déjà existant dans le jeu d'apprentissage. Dans cette étude nous essayons de trouver un DA capable de préserver les performances.

Trois DA ont été essayés: Fragment Control pour chaque modèle (FrgCtrl), Fragment Control IVAB (FrgCtrl IVAB) et 1-SVM.

Le Tableau 5-5 résume les résultats obtenus en validation sur le deuxième jeu de données de test, avec différents DA. Seul le DA (FrgCtrl IVAB) améliore

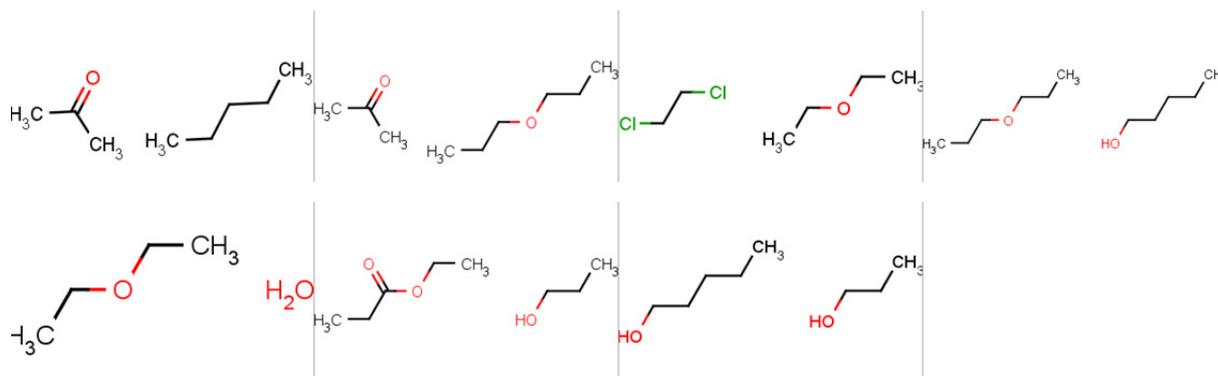
significativement les résultats des prédictions. Par contre, seulement 64 mélanges sont prédits, dont 57 ne contiennent que des composés connus.

Tableau 5-5 Résultats obtenus pour la validation de VS2 avec différents AD. sAD –sans AD.

	sAD	FrgCtrl	1-SVM	FrgCtrl IVAB
Mélanges prédits	499	349	284	64
BA	0.59	0.64	0.63	0.84
Recall(Z)	0.57	0.64	0.56	0.83
Recall(A)	0.62	0.63	0.70	0.85

Sept mélanges contenant un composé nouveau (n-pentane, 1-propoxypropane, éthyloxyéthane, 1-pentanol et ethyl propanoate) ont été acceptés dans le DA et, de plus, ils ont été tous bien prédits. Les structures affichées dans la Figure 5-5 sont très similaires aux composés qui se trouvent dans le jeu d'apprentissage, ce qui explique leur sélection dans le DA.

Figure 5-5 Mélanges dans le DA IVAB(2-4), contenant un composé nouveau



5.3.5 Benchmark

L'étude de benchmark nous a permis de comparer les résultats obtenus avec nos modèles avec ceux d'UNIFAC, UNIFAC Dortmund et COSMO-RS (Tableau 5-6). Étant donné que la méthode UNIFAC Dortmund est une amélioration de l'UNIFAC nous nous attendons à obtenir des meilleurs résultats dans le premier cas, ce qui est confirmé par les résultats ci-dessous. La méthode COSMO-RS donne des bons résultats avec une valeur de BA entre 0.77 et 0.84. Nos modèles sont plus robustes, fait démontré par une variation très faible de BA, entre 0.82 et 0.84. De plus, ces résultats sont parmi les meilleures du tableau.

Tableau 5-6 Étude de benchmark pour la classification.*UNIFAC Dortmund ;**56 mélanges ont été prédits.

	Jeu d'modélisation				VS1				VS2			
Mélanges	400				96				57			
Méthode	SVM	Unifac	Unifac*	COSMO	SVM	Unifac	Unifac*	COSMO	SVM	Unifac	Unifac*	COSMO**
BA	0.82	0.70	0.74	0.77	0.83	0.66	0.80	0.76	0.84	0.84	0.80	0.86
Recall(Z)	0.84	0.90	0.92	0.84	0.74	0.91	0.96	0.78	0.85	1	0.95	0.85
Recall(A)	0.78	0.5	0.56	0.70	0.91	0.41	0.64	0.73	0.84	0.68	0.65	0.86

5.3.6 Classification à partir des courbes VLE

A partir de courbes d'équilibres VLE un mélange peut être classé comme étant un azéotrope ou un zéotrope. Il est intéressant de réinterpréter les résultats obtenus par validation croisée pour la régression de la température d'ébullition et de la composition en phase gazeuse (chapitre 3 et 4), afin de classer les mélanges.

Si un mélange est un azéotrope, la courbe d'ébullition présente un minimum (azéotrope positif) ou un maximum (azéotrope négatif), ce qui n'est pas le cas pour un zéotrope.

De plus, la courbe d'équilibre $y=f(x)$ croise la diagonale ($y=x$) dans le cas d'un azéotrope, ce qui n'est pas valable pour un zéotrope.

Par conséquent, les deux approches de modélisation pour la VLE peuvent apporter la réponse à la question : « étant considéré un mélange binaire, peut-il former un azéotrope ? »

A partir de courbes d'ébullition ($T_b = f(x)$) les mélanges de jeu d'apprentissage et de test peuvent être classés en identifiant un minimum/maximum sur la courbe de chaque mélange. S'il existe, le mélange est considéré comme azéotrope, autrement il est classé comme zéotrope.

Les résultats ont été comparés avec ceux donnés par le modèle COSMO-RS. La comparaison a été faite pour 166 mélanges du jeu de modélisation et 89 mélanges pour le jeu de test, étant donné que seuls ces mélanges ont pu être prédits avec COSMO-RS. Les seuls modèles qui conduisent à des résultats satisfaisants sont les modèles de stacking en validation croisée pour la stratégie « Mixtures Out ». L'obtention des résultats décevants peut être expliquée par le fait que les courbes prédites ou expérimentales ne sont pas lisses dans le cas de beaucoup de mélanges, ce qui explique que dans ce cas, ils sont facilement confondus avec

azéotropes. Les résultats de Tableau 5-7 et le Tableau 5-8 montrent que nos modèles prédisent plus facilement un mélange comme étant azéotrope, tandis que le modèle COSMO-RS prédit plus facilement un mélange comme étant zéotrope (d'où la valeur de Recall(A) et Recall(Z) respectivement élevés).

Tableau 5-7 Classification à partir des courbes d'ébullition pour le jeu de modélisation (n-CV) et jeu de validation externe.

	Jeu de modélisation (n-CV) 166 mélanges			Jeu de test 89 mélanges	
	Mixtures Out (Stacking)	Compounds Out (Stacking)	COSMO	Stacking	COSMO
BA	0.79	0.61	0.65	0.60	0.57
Recall (Z)	0.71	0.31	0.74	0.26	0.93
Recall (A)	0.86	0.91	0.5	0.94	0.22

Tableau 5-8 Matrices de confusion pour la classification à partir des courbes d'ébullition pour le jeu de test de 89 mélanges.

		Stacking		COSMO	
		A	Z	A	Z
Prédit	Actuel				
	A	47	29	11	3
	Z	3	10	39	36

Les résultats de la classification à partir des courbes d'équilibre $y=f(x)$ sont donnés dans les tableaux ci-dessous. Si la courbe obtenue pour un mélange, par régression, croise la diagonale ($y=x$), le mélange est classé comme azéotrope. Si la courbe obtenue pour un mélange, par régression, ne croise pas la diagonale ($y=x$), le mélange est classé comme zéotrope.

Une comparaison avec le modèle COSMO-RS a été aussi présentée. Les prédictions de nos modèles de stacking sont satisfaisantes tant en validation croisée (Tableau 5-9, Tableau 5-10), que pour la validation externe (Tableau 5-11).

Tableau 5-9 Résultats de classification à partir des courbes d'équilibre $y=f(x)$ prédites par modèles de stacking par n-CV pour les 224 mélanges.

	Mixtures Out	Compounds Out	COSMO
BA	0.80	0.78	0.87
Recall (Z)	0.75	0.68	0.91
Recall (A)	0.84	0.89	0.82

Tableau 5-10 Matrices de confusion de classification à partir des courbes d'équilibre $y=f(x)$ prédites par modèles de stacking par n-CV pour les 224 mélanges

		Mixtures Out		Compounds Out		COSMO	
		Actuel					
Prédit		A	Z	A	Z	A	Z
	A	92	28	98	37	90	10
	Z	18	86	12	77	20	104

Le Tableau 5-11 résume les matrices de confusion obtenues pour le jeu de validation ne contenant que les mélanges du DA (FrgCtrl IVAB). On constate que les modèles COSMO-RS ne sont pas capables de prédire correctement quel que soit le type de mélange, tandis que nos modèles les prédisent correctement.

Tableau 5-11 Matrice de confusion de classification à partir des courbes d'équilibre $y=f(x)$ pour les 7 mélanges de test set contenant que des mélanges dans le DA (FrgCtrl IVAB).

		Stacking		COSMO	
		Actuel			
Prédit		A	Z	A	Z
	A	3	1	1	2
	Z	0	3	2	2

5.4 Conclusion

Dans cette partie des modèles de classification ont été développés à partir d'une base de données de 400 mélanges (200 azéotropes/200 zéotropes) et ont été validés par validation croisée (5 paquets). Deux validations supplémentaires ont été faites sur deux jeux de données externes, un de 96 mélanges et un autre de 499 mélanges. Nous avons pu démontrer que les modèles prédisent avec une précision de 84%, le comportement d'un mélange ne contenant que des composés connus ou des composés très proches structurellement aux composés de jeu d'entraînement.

De plus, une étude de benchmark montre que nos modèles mènent à des résultats similaires, voire meilleurs, que ceux fournis par COSMO-RS ou UNIFAC. Cette comparaison n'est pas tout à fait juste, car aucune information n'est obtenue sur la présence de mélanges prédits dans le jeu d'entraînement d'UNIFAC et COSMO, ce qui peut surestimer les performances de ces derniers. Par contre, dans le cas de nos modèles, ces jeux de données de test ont été choisis de façon à ne pas être pollués par des informations présentes dans le jeu d'apprentissage.

Nous avons démontré que les types de descripteurs pour les mélanges utilisés sont adaptés pour caractériser des mélanges et à développer des modèles prédictif de leurs comportement azéotropique. L'application des modèles reste restreinte car elle ne permet que de remplir la matrice formée par les 65 composés purs de jeu de données initial, i.e. 3825 nouveaux mélanges, et des mélanges contenant des composés très proches structurellement de ceux déjà contenus dans le jeu d'apprentissage, et qui se trouvent dans le DA des modèles. Les molécules fournis par Processium permettront d'élargir considérablement ce champ d'application.

La classification à partir de courbes VLE donne de bons résultats dans le cas de la courbe d'équilibre $y=f(x)$, ce qui s'explique surtout par la qualité des courbes, qui sont beaucoup plus lisses et plus précises par rapport aux courbes d'ébullition pour lesquelles les résultats sont faibles. Le fait que les modèles COSMO-RS montrent la même tendance en classifiant bien les mélanges que à partir de la courbe d'équilibre $y=f(x)$, montre cette stratégie est plus adaptée pour la classification. Les modèles obtenus par cette approche sont robustes et montrent de bonnes performances même pour la stratégie "Compounds Out" (cf. Tableau 5-9), ce qui est très prometteur, parce que cela permette d'élargir le champ d'application.

Les modèles de classification ont une précision légèrement meilleure par rapport aux modèles de régression ($y=f(x)$), par contre ces dernières amènent une information en plus sur la composition en phase liquide (x) et gazeuse (y) du point azéotropique.

5.5 Références

1. Stadtherr, M.A., R.W. Maier, and J.F. Brennecke, *Reliable computation of homogeneous azeotropes*. Aiche Journal, 1998. **44**(8): p. 1745-1755.
2. Kearfott, R., *Interval Newton/generalized bisection when there are singularities near roots*. Annals of Operations Research, 1990. **25**(1): p. 181-196.
3. Salomone, E. and J. Espinosa, *Prediction of homogeneous azeotropes with interval analysis techniques exploiting topological considerations*. Industrial & Engineering Chemistry Research, 2001. **40**(6): p. 1580-1588.
4. Zharov, W.T. and L.A. Serafinov, *Physico-chemical Fundamentals of Distillation and Recification*. 1975, Leningrad: Khimiya.
5. Gong, M.Q., et al., *Prediction of Homogeneous Azeotropes by the UNIFAC Method for Binary Refrigerant Mixtures*. Journal of Chemical and Engineering Data, 2010. **55**(1): p. 52-57.
6. Lee, J. and H. Kim, *Development of a criterion for azeotrope prediction of binary refrigerant mixtures*. Korean Journal of Chemical Engineering, 2002. **19**(5): p. 863-865.
7. Hildebrand, J.H., *The Term 'Regular Solution'*. Nature, 1951. **168**(4281): p. 868-868.
8. Schembecker, G. and K.H. Simmrock, *Azeopert - a Heuristic-Numeric System for the Prediction of Azeotrope Formation*. Computers & Chemical Engineering, 1995. **19**: p. S253-S258.
9. Kim, Y.J. and K.H. Simmrock, *Azeopert: An expert system for the prediction of azeotrope formation .1. Binary azeotropes*. Computers & Chemical Engineering, 1997. **21**(1): p. 93-111.
10. Kim, Y.J. and S.K. Kang, *Prediction of the types of binary azeotropes in an expert system, AZEOPERT*. Journal of Industrial and Engineering Chemistry, 1999. **5**(2): p. 105-115.
11. Nascimento, C.A.O., R.M.B. Alves, and F.H. Quina, *New approach for the prediction of azeotropy in binary systems*. Computers & Chemical Engineering, 2003. **27**(12): p. 1755-1759.
12. Kang, J.W., et al., *Development and current status of the Korea Thermophysical Properties Databank (KDB)*. International Journal of Thermophysics, 2001. **22**(2): p. 487-494.
13. Horsley L, H., *Table of Azeotropes and Nonazeotropes*, in *AZEOTROPIC DATA*. 1973, American Chemical Society. p. 1-314.
14. Processium: Available from: <http://www.processium.com/>.
15. *Molecular Operating Environment (MOE)*. 2009, Chemical Computing Group Inc: Montreal, Quebec, Canada.
16. Ralaivola, L., et al., *Graph kernels for chemical informatics*. Neural Netw, 2005. **18**(8): p. 1093-110.
17. Lohmann, J., R. Joh, and J. Gmehling, *From UNIFAC to modified UNIFAC (Dortmund)*. Industrial & Engineering Chemistry Research, 2001. **40**(3): p. 957-964.
18. Jakob, A., et al., *Further development of modified UNIFAC (Dortmund): Revision and extension 5*. Industrial & Engineering Chemistry Research, 2006. **45**(23): p. 7924-7933.
19. Rucker, C., G. Rucker, and M. Meringer, *y-Randomization and its variants in QSPR/QSAR*. Journal of Chemical Information and Modeling, 2007. **47**(6): p. 2345-57.

6 Prédiction du point azéotrope (T_{az} X_{wt} %)

Le point azéotrope, température d'ébullition (T_{az}) et le pourcentage massique (X_{wt}) de 176 mélanges binaires azéotropiques ont été modélisés par régression multilinéaire (ISIDA MLR). Les modèles ont été validés par validations croisées (5 paquets) et par une validation externe supplémentaire sur un jeu de données de 24 nouveaux azéotropes.

Les erreurs de prédiction (3-4K pour T_{az} et 10-14% X_{wt}) sont comparables avec le bruit dans les données expérimentales. De plus, une relation empirique simple reliant T_{az} et les températures d'ébullition des deux constituants purs de l'azéotrope a été suggérée.

L'approche QSPR remédie aux inconvénients des méthodes UNIFAC concernant la disponibilité des paramètres d'interaction et le domaine d'applicabilité. Cela montre que la modélisation QSPR pourrait devenir une alternative intéressante pour les méthodes connues, basées sur la thermodynamique.

La méthodologie et les résultats sont décrits dans l'article publié le mois Novembre 2011, dans *Industrial & Engineering Chemistry Research*.

Quantitative Structure–Property Relationship (QSPR) Modeling of Normal Boiling Point Temperature and Composition of Binary Azeotropes

Vitaly P. Solov'ev

Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Leninskiy prospect, 31a, 119991, Moscow, Russia

Ioana Oprisiu, Gilles Marcou, and Alexandre Varnek*

Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg, 4, rue B.Pascal, Strasbourg, 67000, France

S Supporting Information

ABSTRACT: Quantitative structure–property relationship (QSPR) modeling of normal boiling point temperature (T_{az}) and the composition (weight fraction, X_{1w}) of 176 binary azeotropic mixtures was performed using ensemble multiple linear regression analysis and fragment descriptors implemented in ISIDA software. The models have been validated in external 5-fold cross-validations procedure and on an additional test set of 24 azeotropes. The prediction errors (3–4 K for T_{az} and 10–14 wt % for X_{1w}) are comparable with the noise in experimental data. A simple empirical relationship linking T_{az} with boiling points of two molecular components of azeotrope has been suggested.

1. INTRODUCTION

Azeotropic data are most important for the design of distillation processes.¹ Their theoretical assessment could significantly reduce the costs of selection of proper agents for industrial processes. Several different theoretical approaches could potentially be used to study azeotropes: (i) equations of state including van der Waals equation² and its more complex derivatives (Redlich–Kwong,³ Peng–Robinson,⁴ etc.), (ii) thermodynamic equations involving group contribution techniques (the Wilson,⁵ NRTL,⁶ UNIQUAC⁷ and UNIFAC⁸), (iii) first principles molecular modeling methods including quantum chemistry⁹ and force field molecular dynamics or Monte Carlo¹⁰ approaches explicitly accounting for intermolecular interactions, and (iv) Quantitative Structure–Property Relationships (QSPR).

Among these, the UNIFAC is the most popular method to assess the azeotropic temperature and composition.^{11–15} UNIFAC is based on the thermodynamic equation for the activity coefficient (γ) of liquid 1 in the environment of liquid 2. To calculate γ , UNIFAC considers interactions of selected “structural groups” i ($i \in 1$) and j ($j \in 2$) accounted for the “energy parameters” A_{ij} and A_{ji} . The latter are fitted on available experimental data. An extensive table of UNIFAC group–interaction parameters was first published by Fredenslund et al.¹⁶ in 1977. It has been several times revised because of the growing volume of experimental data. The latest UNIFAC update was based on the Dortmund Data Bank containing more than 39 000 VLE data. Currently, UNIFAC considers more than 60 structural groups representing both simple molecular fragments (e.g., methyl, carbonyl, ether) and the entire molecules (e.g., DMSO, acrylonitrile, methylpyrrolidone).

Despite its solid thermodynamic basis and excellent results of quantitative estimations of vapor–liquid equilibria, UNIFAC has

two serious drawbacks. The first one concerns the energy parameters A_{ij} and A_{ji} which cannot be assessed from the parameters of individual groups i and j but must be fitted directly on experimental VLE data for the binary mixtures. Thus, the total number of the energy parameters rises as a squared number of the group. According to our estimations, for the current list of some 60 groups, less than half of the energy parameters were calculated. The second drawback is related to inflexible strategy of groups' selection which in most cases makes it impossible to represent a complex system as an ensemble of groups. For instance, in a recently reported UNIFAC model for ionic liquids (IL),¹⁷ 12 new groups have been introduced and each of them represents an individual IL. Taking into account a large number of existing ILs, further development of UNIFAC models for these systems looks unrealistic. A similar conclusion could be drawn for a variety of liquid binary systems involving heterocyclic molecules because each of them should be also represented as an individual group. The above problems significantly restrict UNIFAC applications and motivate one to look for an alternative solution to model azeotropes.

Quantitative structure–property relationship (QSPR) approach could be considered as such an alternative. A QSPR model relates a given physical property with chemical structure encoding by molecular descriptors. Although, QSPR technique is traditionally used to model individual compounds, some efforts have been recently made to model their mixtures. Thus, Ajmani et al.^{18–20} reported QSPR models for infinite-dilution activity coefficients,

Received: August 19, 2011

Accepted: November 9, 2011

Revised: November 5, 2011

Published: November 09, 2011

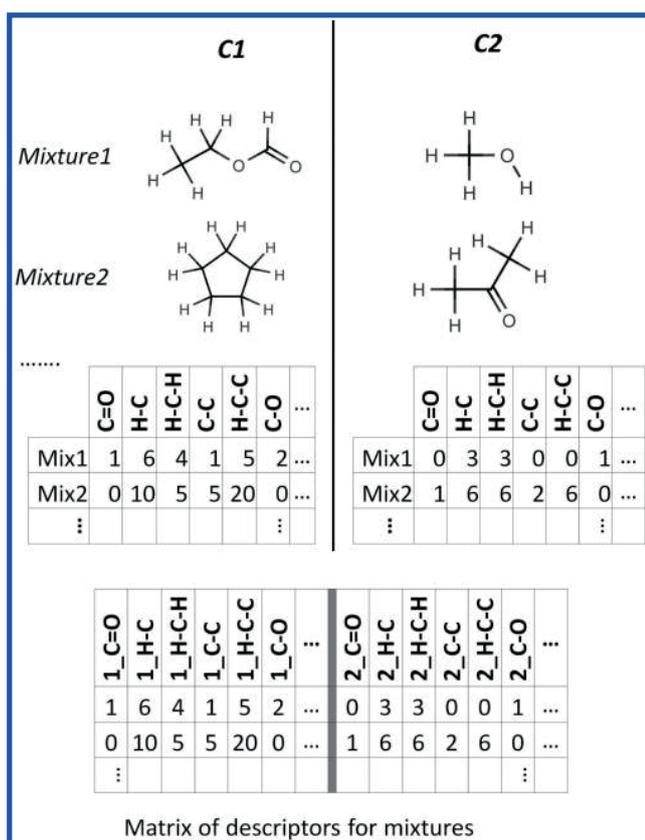


Figure 1. ISIDA fragment descriptors to model binary azeotropes. For a binary mixture, descriptor vector is formed by concatenation of descriptors of the individual molecular components C1 and C2. Between two components, the compound with lower boiling point temperature is always taken as C1.

excess molar volume, and density of liquid binary mixtures using special mixture descriptors constructed from those describing individual compounds. Recently, Katritzky et al.²¹ performed QSPR modeling of normal boiling point temperature of azeotropes (T_{az}) using the CODESSA PRO program. Two different strategies have been used to prepare mixture descriptors: either simple arithmetic average of those calculated for individual molecular components 1 and 2, or weighting 1 and 2 by their molar ratios in the azeotrope. The latter is of limited practical interest because experimental measures of the mixture's composition at the azeotropic point is always coupled with obtaining T_{az} . Predictive performance obtained in²¹ linear models is rather weak: the standard deviation of about 23 K has been obtained at the fitting stage and has not even been reported for the external test set.

In this paper, we describe QSPR modeling of both normal boiling point temperature T_{az} and the composition (weight fraction, X_{1w}) of binary azeotropes using ensemble Multiple Linear Regression (e-MLR) approach and fragment descriptors implemented in the ISIDA (In Silico design and Data Analysis) software. The errors of predictions are comparable with the noise in experimental data used in the models' development.

2. METHODS

2.1. Descriptors. Substructural molecular fragments (SMF) of the ISIDA program^{22,23} as subgraphs of molecular graph were

used as descriptors. They represent either sequences (the shortest topological paths with explicit presentation of all atoms and bonds), or atom pairs (the paths where only terminal atoms and bonds as well topological distance between them are given).²⁴ For searching the shortest paths, the Floyd algorithm²⁵ is used. For each type of sequences, the minimal ($n_{min} \geq 2$) and maximal ($n_{max} \leq 15$) number of constituent atoms is defined. For the given combination n_{min} and n_{max} , all intermediate shortest paths with n atoms ($n_{min} < n < n_{max}$) are also generated.

Descriptor vector for a binary mixture was generated by concatenation of descriptors of its individual molecular components (Figure 1). Thus, similar molecular fragments belonging to two different mixture components are considered as different. Between two components, the compound with lower boiling point temperature was always taken as the first one. This strategy to construct descriptors may be restrictive in the case when experimental boiling points are not available. By this reason, the model predicting boiling points of individual compounds has been built and implemented in the platform of virtual screening available at <http://infochim.u-strasbg.fr/webserv/VSEngine.html>.

2.2. Data Sets. 176 azeotropes represented by binary mixtures of 42 pure compounds were taken from the book²⁶ (see Supporting Information) in which the authors made an attempt to select the most reliable data. The 2D sketcher *EdChemS*^{23,27,28} and the Structure Data File (SDF) manager *EdiSDF*^{23,27,29} were used to prepare the 2D structures of the azeotrope molecular components expressed with explicit hydrogen atoms. Both structures of the azeotrope components are kept as one SDF record. Distributions of normal boiling point temperature (T_{az}) and the composition (weight fraction of first component, X_{1w}) are given in Figure 2. The values of boiling point temperature T_{az} and weight fraction X_{1w} vary in the range of 293–411 K and 16–99 wt %, correspondingly (Figure 2).

To validate the models, we used an external test set containing 24 binary mixtures (see Supporting Information) collected from reference 1. These mixtures are composed of 32 pure organic compounds, 10 of which were not included in the modeling data set. The values of boiling point temperature T_{az} and weight fraction X_{1w} vary in the range of 313–374 K and 30–98 wt %, correspondingly.

2.3. Obtaining and Validating QSPR Models. The e-MLR module of the ISIDA software^{29–31} has been applied to build ensemble of QSPR models, each of them described by eq 1

$$Y = a_0 + \sum a_i N_i \quad (1)$$

where Y is modeling property ($Y = T_{az}$ or X_{1w}), N_i is the count of the i -th SMF, a_i is its contribution, and a_0 is a free term. Concatenated fragments always occurring in the same combination in each compound of the training set are interpreted as one extended fragment. Rare fragments (i.e., found in less than m molecules, here $m < 3$) were excluded. Original forward and backward stepwise techniques^{31–33} have been used to select the most pertinent variables X from the initial pools of SMF descriptors. The parameters a_0 and a_i in eq 1 have been fitted by the Singular Value Decomposition method.³⁴ Varying the minimal and maximal length of the fragments, 120 initial pools of descriptors were generated followed by the building of QSPR model from each of them. The most robust models were selected according to leave-one-out cross-validation correlation coefficient $Q^2 > Q^2_{lim}$, where Q^2_{lim} is a user-defined threshold. Here, $Q^2_{lim} \geq 0.8$ (T_{az}) and $Q^2_{lim} \geq 0.5$ (X_{1w}) have been used.

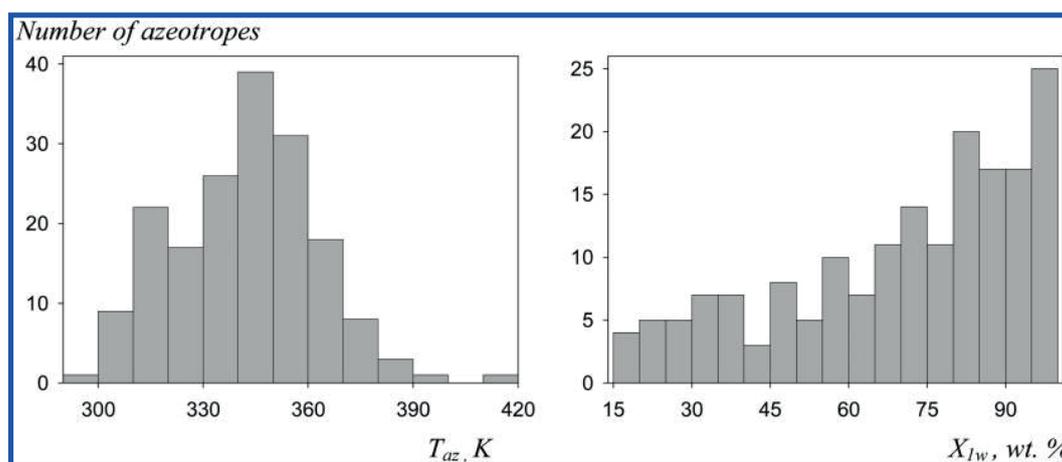


Figure 2. Distribution of experimental values of normal boiling point temperature (left) and composition (right) in the modeling data set of 176 binary azeotropic mixtures.

Selected individual models are used to build a Consensus Model (CM). For each test set compound, the program computes the property as an arithmetic mean of values obtained by ensemble of individual models excluding those leading to outlying values according to Tompson's rule and a ranked series method.³⁵ As shown in refs 27, 30, 32, and 33 CM smoothes inaccuracies of individual models and ensures more reliable predictions.

When applying an individual model, the program checks its applicability domain AD^{29,36} which measures a similarity between a test compound and the training set compounds. If the test compound is identified as being outside AD, the corresponding individual model is not used in CM calculations. Here, two AD approaches have been simultaneously used: bounding box, considering AD as a multidimension descriptor space confined by minimal and maximal values of counts of SMF descriptors, and fragment control rejecting test compounds fragments non-occurring in the initial SMF descriptors pool.

To validate Consensus Models, the external 5-fold cross-validation (5-CV) was applied.^{33,37} In this procedure, the entire modeling data set is divided in 5 nonoverlapping pairs of training and test sets. Each training set covers $4/5$ of the data set, while the related test set covers the remaining $1/5$. Because each molecule belongs to one of 5 test sets, all molecules from the initial data set are predicted. In such a way, one can avoid an ambiguity linked to selection of one only test set. Predictive performance of the models has been assessed by determination coefficient (R^2) and mean-average error (MAE) for the entire modeling set taken as a combination of all five test sets:

$$R^2 = 1 - \frac{\sum (Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum (Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2} \quad (2)$$

$$\text{MAE} = \frac{\sum |Y_{\text{exp}} - Y_{\text{pred}}|}{n} \quad (3)$$

were Y_{pred} and Y_{exp} are, respectively, predicted and experimental property values, and $\langle Y \rangle_{\text{exp}}$ is an average over experimental values.

At the final step, the SMF initial pools providing the best CM models in 5-CV have been used to build the models on the entire modeling data set followed by their validation on the external test set.

3. RESULTS AND DISCUSSION

In 5-CV, 82–171 (T_{az}) and 176–199 (X_{1w}) QSPR models were selected to build CMs for each of 5 training/test set combinations. The individual models contained, on average, 36 (T_{az}) and 30 (X_{1w}) SMF variables. A reasonable performance of prediction of T_{az} has been achieved: the squared determination coefficient is high ($R^2 = 0.880$) and MAE is 4.2 K. The latter is rather small compared to the range of experimental T_{az} in the modeling set (293–411 K, Figure 3). The CMs predictions of the X_{1w} resulted in $R^2 = 0.697$ and MAE = 9.6 wt % (Figure 3).

Another way to analyze the model's performance is Regression Error Curve which displays a fraction of the data set predicted within a given threshold of absolute prediction error $|T_{\text{az exp}} - T_{\text{az pred}}|$ for boiling point temperature and $|X_{1w \text{ exp}} - X_{1w \text{ pred}}|$ for the composition. Figure 4 shows that for 60% binary azeotropic systems this error is about 4.2 K for the temperature and 10 wt % for the composition.

An empirical relationship linking T_{az} with normal boiling point temperatures of first (T_{C1}) and second (T_{C2}) components could be derived from the analysis of 176 experimental data in the modeling set

$$T_{\text{az}} = 0.984(\pm 0.002)(T_{C1} + T_{C2})/2 - 0.414(\pm 0.023)|T_{C2} - T_{C1}| \quad (4)$$

Remarkable statistical parameters ($R^2 = 0.94$, MAE = 3.3 K, Figure 3) demonstrate its good predictive performance.

The developed e-MLR consensus model was validated on the additional test set. Among 24 azeotropes, 12 have been found outside of the model's applicability domain and, therefore, have not been predicted (see Supporting Information). The following statistical parameters of predictions have been obtained for the remaining 12 compounds: $R^2 = 0.84$ and MAE = 3.7 K for T_{az} and $R^2 = 0.32$ and MAE = 14.4 wt % for X_{1w} (Figure 3). For the boiling temperature, these parameters are similar to those obtained in cross-validation procedure on the modeling set thus demonstrating the robustness of the developed models. For the composition, the predictions on the additional test set are less good than those in 5-CV; this shows that the modeling of X_{1w} is more difficult. Notice that both e-MLR model and eq 4 detected the pyridine/acetic acid as outlier thus showing a potential problem

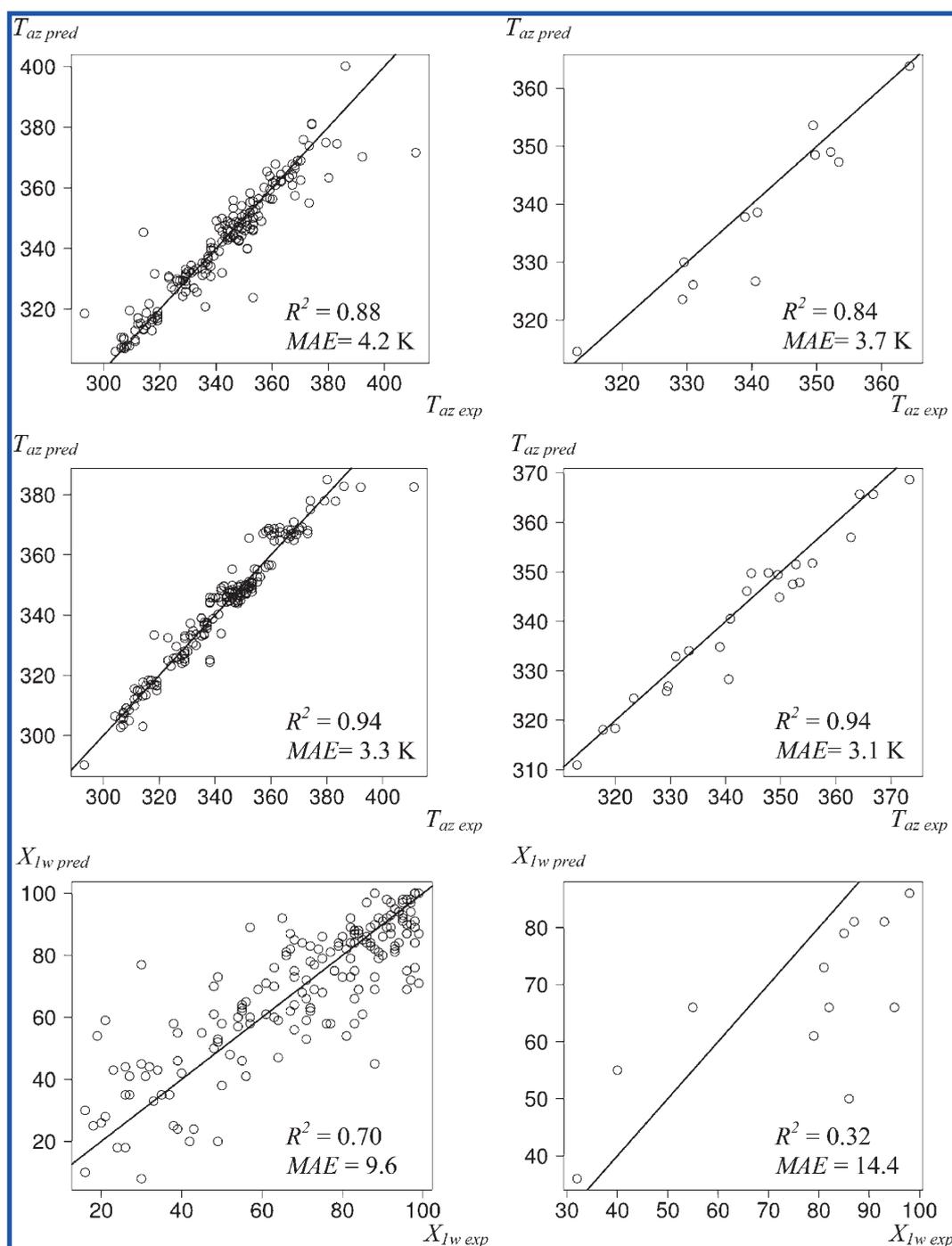


Figure 3. Calculations performed on the modeling set (left) and additional test set (right). Predicted ($T_{az\ pred}$) vs experimental ($T_{az\ exp}$) values of normal boiling point temperature of binary azeotropes computed by the consensus MLR models (top) and eq 1 (middle). Predicted $X_{1w\ pred}$ vs experimental $X_{1w\ exp}$ values of the composition of azeotropes for the consensus MLR models (bottom). Predicted values correspond to all test sets of external 5-fold cross-validations.

with the accuracy of the experimental data for this azeotropic mixture.

Being applied to the entire additional test set, empirical relationship 4 demonstrated its good predictive performance: $R^2 = 0.94$ and $MAE = 3.1\text{ K}$ (Figure 3).

The question arises whether our predictions are acceptable. Analysis of the experimental data from references 38–40 shows

that for one same system, the boiling temperature and composition may vary from 0.3 to 7.6 K and from 0.4 to 10.6 wt % as a function of data source (Table 1). These support an observation by Gmehling and Böls¹ concerning difficulties to assess a reliability of the thermodynamics data reported in the literature. One can see that the values representing the noise in experimental data (Table 1) are close to the prediction errors of the models developed

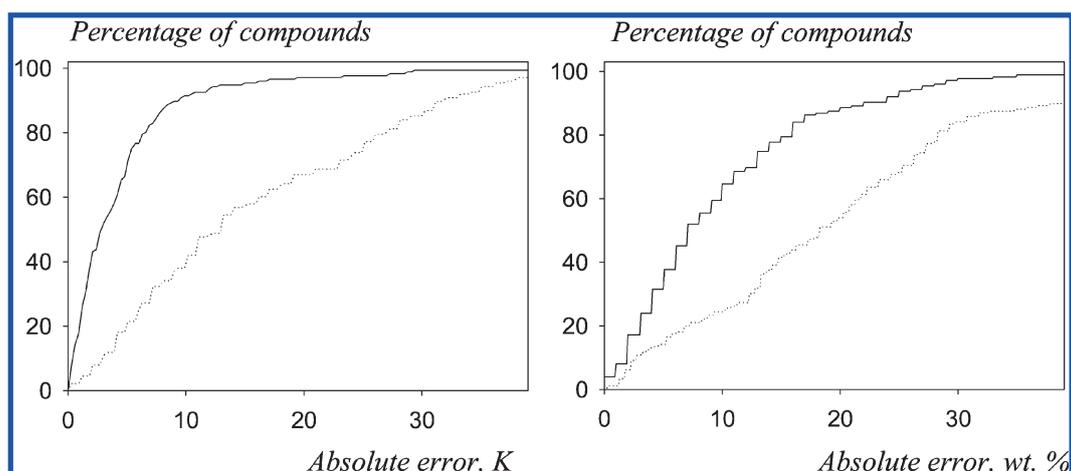


Figure 4. Regression Error Curves: percentage of compounds vs absolute prediction error; continuous line corresponds to the predictions of $T_{\text{bp az}}$ (left) and X_{1w} (right) by MLR CMs. The dotted line corresponds to “no model” calculations in which the average over experimental data set has been used as a predicted value for every compound.

Table 1. Discrepancy Levels in Experimental Normal Boiling Point Temperature ($T_{\text{bp az}}$) and the Composition (X_{1w}) of Two-Component Liquid Azeotropes According to Different Sources^{38–40}

no.	component 1	component 2	min X_{1w} , %	max X_{1w} , %	ΔX_{1w} , %	min $T_{\text{bp az}}$, K	max $T_{\text{bp az}}$, K	ΔT_{bp}
1	methyl ethyl ketone	benzene	37.5	47	9.5	351.15	351.55	0.4
2	2-butanol	H ₂ O	68	73.2	5.2	360.15	361.65	1.5
3	benzene	cyclohexane	49.7	55	5.3	350.55	351.15	0.6
4	acetone	methyl acetate	48	50	2	328.15	328.95	0.8
5	nitromethane	1,4-dioxane	56.5	57	0.5	373.15	373.7	0.55
6	H ₂ O	formic acid	22.5	26	3.5	373.15	380.8	7.65
7	ethyl acetate	ethanol	69	74.2	5.2	344.95	345.33	0.38
8	1-propanol	H ₂ O	70.9	72	1.1	360.91	361.25	0.34
9	cyclohexane	1,2-dichloroethane	50	50.4	0.4	347.55	348.15	0.6
10	CCl ₄	methyl ethyl ketone	71	81.6	10.6	346.85	347.15	0.3

in this work. Notice that boiling points data are significantly more reliable than mixtures composition ones, which explains reasonably good predictions of T_{az} and poor predictions of X_{1w} . Thus, one can conclude that data quality is the major problem in modeling of azeotropic composition.

It should be noted that any new molecule, more or less similar to that in the training set, can be easily constructed from sub-structural molecular fragments used here as descriptors. Thus, compared to UNIFAC, QSPR approach provides a much more flexible solution to predict properties of new azeotrope systems.

4. CONCLUSIONS

QSAR models for normal boiling point temperature and composition of binary azeotropes have been built on a relatively small data set of 176 experimental systems using fragment descriptors and ensemble modeling approach. These models display a reasonable predictive performance for T_{az} models and poor performance for X_{1w} models. The latter can be explained by noise in experimental data. QSAR approach overcomes the drawbacks of UNIFAC with respect to the availability of method's parameters and applicability domain. This shows that QSPR modeling could become a valuable alternative for popular thermodynamic-based methods.

■ ASSOCIATED CONTENT

S Supporting Information. SM1: Experimental and predicted values of normal boiling point temperature ($T_{\text{bp az}}$) and the composition (X_{1w}) of 176 two-component liquid azeotropes. SM2: Experimental and predicted values of normal boiling point temperature ($T_{\text{bp az}}$) and composition (X_{1w}) for extra test set of 24 two-component liquid azeotropes. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: varnek@chimie.u-strasbg.fr.

■ ACKNOWLEDGMENT

V.S. thanks the ARCUS “Alsace-Russia/Ukraine” project, GDRI SupraChem, and the Russian Foundation for Basic Research (project 09-03-93106) for the support.

■ REFERENCES

(1) Gmehling, J.; Böls, R. Azeotropic Data for Binary and Ternary Systems at Moderate Pressures. *J. Chem. Eng. Data* **1996**, *41* (2), 202–209.

- (2) van der Waals, J. D. Over de Constnuteit van den gas-en Vloeistofoestand. Doctoral Dissertation. 1873, Leiden, Holland.
- (3) Redlich, O.; Kwong, J. N. On the thermodynamics of solutions; an equation of state; fugacities of gaseous solutions. *Chem. Rev.* **1949**, *44* (1), 233–244.
- (4) Peng, D.-Y.; Robinson, D. B. A New Two-Constant Equation of State. *Ind. Eng. Chem., Fundam.* **1976**, *15* (1), 59–64.
- (5) Wilson, G. M. Vapor-Liquid Equilibrium. XI. A New Expression for the Excess Free Energy of Mixing. *J. Am. Chem. Soc.* **1963**, *86* (2), 127–130.
- (6) Renon, H.; Prausnitz, J. M. Local compositions in thermodynamic excess functions for liquid mixtures. *AIChE J.* **1968**, *14* (1), 135–144.
- (7) Anderson, T. F.; Prausnitz, J. M. Application of Uniquac Equation to Calculation of Multicomponent Phase-Equilibria 0.1. Vapor-Liquid-Equilibria. *Ind. Eng. Chem. Proc. Des. Dev.* **1978**, *17* (4), 552–561.
- (8) Gmehling, J.; Rasmussen, P.; Fredenslund, A. A Survey of the Calculation of Phase-Equilibria with the Aid of the Unifac-Method. *Chem.-Ing.-Tech.* **1980**, *52* (9), 724.
- (9) Prausnitz, J. M.; Tavares, F. W. Thermodynamics of fluid-phase equilibria for standard chemical engineering operations. *AIChE J.* **2004**, *50* (4), 739–761.
- (10) Punnathanam, S.; Monson, P. A. Crystal nucleation in binary hard sphere mixtures: A Monte Carlo simulation study. *J. Chem. Phys.* **2006**, *125* (2), 024508-1–024508-11.
- (11) Salomone, E.; Espinosa, J. Prediction of Homogeneous Azeotropes with Interval Analysis Techniques Exploiting Topological Considerations. *Ind. Eng. Chem. Res.* **2001**, *40* (6), 1580–1588.
- (12) Dong, X.; Gong, M.; Zhang, Y.; Liu, J.; Wu, J. Prediction of Homogeneous Azeotropes by the UNIFAC Method for Binary Refrigerant Mixtures. *J. Chem. Eng. Data* **2010**, *55* (1), 52–57.
- (13) Maier, R. W.; Brennecke, J. F.; Stadtherr, M. A. Reliable computation of reactive azeotropes. *Comput. Chem. Eng.* **2000**, *24* (8), 1851–1858.
- (14) Maier, R. W.; Brennecke, J. F.; Stadtherr, M. A. Reliable Computation of Homogeneous Azeotropes. *AIChE J.* **1998**, *44* (8), 1745–1755.
- (15) Harding, S. T.; Maranas, C. D.; McDonald, C. M.; Floudas, C. A. Locating All Homogeneous Azeotropes in Multicomponent Mixtures. *Ind. Eng. Chem. Res.* **1997**, *36* (1), 160–178.
- (16) Fredenslund, A.; Gmehling, J.; Michelsen, M. L.; Rasmussen, P.; Prausnitz, J. M. Computerized Design of Multicomponent Distillation-Columns Using Unifac Group Contribution Method for Calculation of Activity-Coefficients. *Ind. Eng. Chem. Proc. Des. Dev.* **1977**, *16* (4), 450–462.
- (17) Lei, Z. G.; Zhang, J. G.; Li, Q. S.; Chen, B. H. UNIFAC Model for Ionic Liquids. *Ind. Eng. Chem. Res.* **2009**, *48* (5), 2697–2704.
- (18) Ajmani, S.; Rogers, S. C.; Barley, M. H.; Burgess, A. N.; Livingstone, D. J. Characterization of Mixtures Part 1: Prediction of Infinite-Dilution Activity Coefficients Using Neural Network-Based QSPR Models. *QSAR Comb. Sci.* **2008**, *27* (11–12), 1346–1361.
- (19) Ajmani, S.; Rogers, S. C.; Barley, M. H.; Burgess, A. N.; Livingstone, D. J. Characterization of Mixtures. Part 2: QSPR Models for Prediction of Excess Molar Volume and Liquid Density Using Neural Networks. *Mol. Inform.* **2010**, *29* (8–9), 645–653.
- (20) Ajmani, S.; Rogers, S. C.; Barley, M. H.; Livingstone, D. J. Application of QSPR to mixtures. *J. Chem. Inf. Model.* **2006**, *46* (5), 2043–2055.
- (21) Katritzky, A. R.; Stoyanova-Slavova, I. B.; Tamm, K.; Tamm, T.; Karelson, M. Application of the QSPR Approach to the Boiling Points of Azeotropes. *J. Phys. Chem. A* **2011**, *115* (15), 3475–3479.
- (22) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 847–858.
- (23) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comp.-Aided Mol. Des.* **2005**, *19* (9–10), 693–703.
- (24) Baskin, I.; Varnek, A. Fragment Descriptors in SAR/QSPR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. In *Chemoinformatic Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Eds.; RCS Publishing, 2008; p 1–43.
- (25) Swamy, M. N. S.; Thulasiraman, K. *Graphs, Networks, and Algorithms*; John Wiley & Sons: New York, 1981.
- (26) Gordon, A. J.; Ford, R. A. *The Chemist's Companion. A Handbook of Practical Data, Techniques, and References*; John Wiley and Sons: New York, 1972; p 537.
- (27) Solov'ev, V. P.; Kireeva, N.; Tsivadze, A. Y.; Varnek, A. Structure-Property Modelling of Complex Formation of Strontium with Organic Ligands in Water. *J. Struct. Chem.* **2006**, *47* (2), 298–311.
- (28) Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. Quantitative Structure-Property Relationship Modeling of beta-Cyclodextrin Complexation Free Energies. *J. Chem. Inf. Comp. Sci.* **2004**, *44* (2), 529–541.
- (29) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 191–198.
- (30) Varnek, A.; Solov'ev, V. P. "In Silico" Design of Potential Anti-HIV Actives Using Fragment Descriptors. *Comb. Chem. High Throughput Screening* **2005**, *8* (5), 403–416.
- (31) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generation-How Much Effort May the Mining for Successful QSAR Models Take? *J. Chem. Inf. Model.* **2007**, *47* (3), 927–939.
- (32) Solov'ev, V. P.; Varnek, A. Structure-Property Modeling of Metal Binders Using Molecular Fragments. *Russ. Chem. Bull.* **2004**, *53* (7), 1434–1445.
- (33) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I.; Solov'ev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **2007**, *47* (3), 1111–1122.
- (34) Golub, G. H.; Reinsch, C. Singular Value Decomposition and Least Squares Solutions. *Numer. Math.* **1970**, *14*, 403–420.
- (35) Muller, P. H.; Neumann, P.; Storm, R. *Tafeln der mathematischen Statistik*; VEB Fachbuchverlag: Leipzig, 1979, p 280.
- (36) Varnek, A.; Fourches, D.; Kireeva, N.; Klimchuk, O.; Marcou, G.; Tsivadze, A.; Solov'ev, V. Computer-Aided Design of New Metal Binders. *Radiochim. Acta* **2008**, *96* (8), 505–511.
- (37) Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P. Benchmarking of Linear and Non-Linear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Organic Ligands. *J. Chem. Inf. Model.* **2006**, *46* (2), 808–819.
- (38) Dean, J. A. *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: New York, 2005.
- (39) Horsley, L. H. *Tables of Azeotropes and Nonazeotropes, in Azeotropic Data-III*; American Chemical Society: Washington, DC, 1973; pp 1–613.
- (40) Ponton, J. W. *Azeotrope Databank*. 2001; available from <http://eweb.chemeng.ed.ac.uk/jack/newWork/Chemeng/azeotrope/>.

7 Température d'ébullition des corps purs

La température d'ébullition des corps purs est une donnée nécessaire pour l'obtention des modèles QSPR permettant d'estimer la température d'ébullition et la fraction molaire d'un azéotrope décrit précédemment.

Par ailleurs ces estimations sur les corps purs pourraient être utilisées pour corriger les modèles de prédiction de la courbe de bulle, reliant la température d'ébullition du mélange à sa composition. Ces courbes atteignent les températures d'ébullition des corps purs quand $x=0$ et quand $x=1$.

Dans ce projet, un jeu de données grand et divers contenant 2098 molécules organiques a été modélisé. Plusieurs machines d'apprentissage linéaires et non-linéaires ont été utilisées (ASNN, SQS, ISIDA MLR). Les performances des modèles obtenus sont comparables aux ceux déjà publiés (RMSE entre 9.3 K et 17.9K).

Par rapport aux travaux antérieurs, la valeur ajoutée de notre travail est:

- l'utilisation de modèles consensus, c'est-à-dire le calcul d'une moyenne des prédictions de plusieurs centaines de modèles;
- l'usage d'une estimation des performances prédictives par validations croisées;
- la prise en compte du DA, ce qui permet l'utilisation des modèles pour le criblage virtuel d'une grande base de données;
- la publication des modèles via une interface WEB.

La méthodologie et les résultats sont présentés dans l'article récemment soumis dans le *Thermodynamica Acta*.

Publicly available models to predict normal boiling points of organic compounds.

*Ioana Oprisiu^a, Gilles Marcou^a, Dragos Horvath^a, Damien Bernard Brunel^b, Fabien Rivollet^b, Alexandre Varnek^{*a}*

^aLaboratoire d'Infochimie UMR 7177 CNRS, Université de Strasbourg, 4, rue B. Pascal, Strasbourg 67000, France;

^bProcessium, C.E.I.3 - 62 Bd Niels Bohr - BP 2132, F-69603 VILLEURBANNE, France.

*Corresponding Author: Tel.: +33(0)368851560; Fax: +33(0)368851589; E-mail: varnek@unistra.fr.

ABSTRACT

Quantitative structure-property models to predict the Normal Boiling Point (T_b) of organic compounds were developed using non-linear ASNN (Associative Neural Networks) as well as Multiple Linear Regression – ISIDA-MLR and SQS (Stochastic QSAR Sampler). Models were built on a diverse set of 2098 organic compounds with T_b varying in the range of 185 - 491 K. In ISIDA-MLR and ASNN calculations, fragment descriptors were used, whereas fragment, FPT (Fuzzy Pharmacophore Triplets), and ChemAxon descriptors were employed in SQS models. Prediction quality of the models has been assessed in 5-fold cross validation. Obtained models were implemented in the on-line ISIDA Predictor at <http://infochim.u-strasbg.fr/webserv/VSEngine.html>.

KEYWORDS

QSPR/QSAR, Normal Boiling Point, Publicly Available Predictor

1. INTRODUCTION

Knowledge of normal boiling points (T_b) of organic compounds is widely needed in chemical industry, mainly for assessment of vapor pressure[1], heat of vaporization[2], liquid viscosity[3],

critical temperature[4] and other properties. Since experimental T_b values are available only for a few thousands of compounds out of the 60 million registered in the CAS database[5], theoretical models may represent a fast and convenient way to estimate this property for compounds synthesized so far, and the only option for molecules to be synthesized.

A SciFinder[6] query with the keywords “*QSPR Boiling Point OR Prediction Boiling Point OR Group Contribution Boiling Point*” returned 1075 publications. Most of models reported in the literature were developed on small data sets (less than 200) representing specific compound classes such as hydrocarbons[7-9], haloalkanes[10-12], chlorosilanes[13], saturated alcohols[14], sulfides and thiols[15], etc. Modeling large and diverse data sets, the only method to obtain predictive equations having an applicability domain which covers a significant chemical space, is however much more challenging, thus rarely addressed[16].

In chemical industry, the classical approach to property estimation is the group contribution method (GC) [17-19], where each molecule is considered as an ensemble of “fundamental” groups linearly contributing to the property of interest. GC approaches are thus classical multiple regression models. However, unlike conventional QSPR techniques, in which a subset of descriptors involved in the model is selected from a large initial pool of descriptors, in GC methods the restricted list of descriptors (functional groups) is suggested by the authors empirically. By contrast to GC approaches, in Group-Interaction Contribution-based approaches (GIC)[20] the interactions between groups are taken into account.

Typical GC methods are those of Joback and Reid[17], Constantinou and Gani[18] and Stein and Brown[19], which served as a benchmark for other recent work. Some of the most recent publications presenting the use of improved GC method for the prediction of normal boiling point, along with some QSPR models[16, 20-26] are summarized in Table 1.

Two main QSPR approaches were applied: Multiple Linear regression (MLR) and Neural Networks (NN).

Insert Table 1

Performance of the models listed in Table 1 is difficult to compare because several reasons. In some cases the accuracy of predictions was measured by Root-Mean Squared Error (RMSE,

eq. 4), in other cases by Mean Absolute Error (MAE, eq. 5). Notice that RMSE is more sensitive for outliers than MAE, thus $RMSE \geq MAE$. Furthermore, these results pertain to the specific test and training sets, and may be strongly influenced by the peculiar choice of training vs. test compounds. RMSE values derived on hand of cross-validation approaches, where each compound is alternatively used for model fitting and kept out for prediction, are more robust indicators of the quality of the models, but still dramatically depend on the actual training data set. Rigorous benchmarking cannot be done unless different models are confronted with respect to a common compound collection – and even then, it cannot be foretold whether the relative ranking in predictive performance would be maintained if benchmarking were to be repeated with respect to a novel test set.

It should be noted that there exist some on-line applications for T_b prediction:

- “Artist Property Estimation”[27] is a product of Dortmund Data Bank (DDB)[28]. The online-estimation service on the website started recently and the free version uses the simple Joback [17] group contribution method.

- “PirikaLight”[29] is the on-line demo version of a physical property prediction tool. It predicts only molecules containing maximum 8 carbon atoms, two oxygen atoms and one nitrogen atom. It uses Neural Networks to estimate T_b .

- ACD/Labs [30] tools incorporated in SciFinder predict T_b only for the compounds stored in the CAS databases, and it is not freely available.

In this work, we report the predictive models obtained on large divers dataset using the ISIDA fragment and pharmacophoric descriptors and different linear and non-linear machine learning methods. Compared to previous publications, our models have several clear advantages.

- Models were trained on a large set of ~2100 molecules, which compares favorably to previous work, by far exceeding all but two of the above cited studies.

- This data set is diverse, regrouping various compound classes. Unfortunately, it is impossible to directly compare the diversity of the set to the ones used by predecessors.

- In spite of size and diversity, herein obtained prediction errors are well within the state-of-the-art values.

- An ensemble modeling approach has been used. Thus, for prediction calculations on the external data, we use *consensus* model integrating hundreds of individual models.
- Prediction performances were assessed using robust external 5-Fold cross-validation[31] and extensive testing on external sets.
- Applicability domain of models is taken into account. This allows one to use the models in virtual screening of large databases.
- Models are publically available via WEB based application. Only a browser is required to the user to apply the models. Unlike the abovementioned on-line applications, our tools support batch prediction jobs of arbitrarily large compound collections.

2. METHODOLOGY

2.1 Data preparation

2.1.1 Preliminary Benchmarking Studies.

The set of Hall and Story[24] has been used for preliminary exploration of the behavior of our QSPR tools in the context of boiling point modeling, in order to quickly compare the predictive potential of the herein considered descriptors with respect to other molecular descriptors used in the field.

2.1.2 Training set

The initial set of T_b for 2962 compounds collected from the literature was provided by *Processium* [32]. Only organic compounds containing C, H, F, Cl, Br, I, N, O and S were selected. Compounds having less than two carbon atoms and less than three atoms in total were also discarded, since some of the employed molecular descriptors (in particular the pharmacophore triplets) were specifically designed for complex organic “drug-like” compounds. After these operations 2853 compounds remained.

A second filter was applied on this set eliminating molecules containing rare fragments of type IAB2-6 (see *Descriptors* section). The compounds containing rare fragments were put aside to constitute an independent and particularly difficult test set (TS1).

Last but not least, compounds having a large variation in experimental T_b reported in different publications[33, 34] were also discarded. Failure to do so lead to models in which such

molecules appeared as outliers (results not shown), which underlines the robustness of the herein used training set (overfitting of experimental errors did not seem to occur). The boiling point data are expected to be highly accurate (in the range of 1–2 degrees), nevertheless, the literature data contain important discrepancies as for the following molecules: *carbosulfan* (399.15K in ref.[33] and 492.45K in ref.[35]); 1-methylantracene (472.65K [33] vs. 636.2K [34]); 2-bromobutyryl bromide : T_b between 415.15 K and 447.15 K; Benzilic acid (460.15K [36] vs. decomposition at 453.15K [33]); 5-Methylbenzofuran (471.15K [36], but at 453.15 K, according to ref.[33] it sublimes).

Such discrepancies are most likely due to the fact that T_b is measured at different pressures, followed by extrapolation to 1 atm using the Clausius-Clapeyron equation or some of its simplifications. Stein and Brown [37] found that extrapolations from two measured values to the same pressure of 1 atm depend on the selected equations. Such extrapolations had a MAE in the range of 12 to 18 K, which should be of a same magnitude order as the average error of an ideal T_b prediction model[38].

Insert Figure 1

The training set used for modeling was formed of 2098 organic compounds, containing different chemical groups, with temperature varying in a range of 185 K and 491 K (see Figure 1). The distribution of the temperature has a maximum at 430K, while compounds with boiling points above 480 K are rare.

2.1.3 Test sets

Two external test sets, TS1 and TS2, were used:

TS1, the “difficult” set contains the 594 compounds discarded during the second filtering stage, due to presence of rare fragments.

TS2 is based on data from reference [39], out of which 304 compounds, not part of the training set and containing more than one carbon atom were retained.

Preliminary predictions identified many “outliers” that were proven to be erroneous data, consequently 516 compounds in TS1 and 290 compounds in TS2, were retained for final predictions.

Notice that TS2 is more similar to the training set than TS1. Thus, maximum Tanimoto coefficient between each molecule from the test set and the training set was computed, using "Chemistry/Overlap Analysis" ChemAxon tool[40]. Its average value is 0.57 for TS1 and 0.74 for TS2.

2.1.4 Descriptors

Prior to descriptor calculations, all molecule sets were standardized with the ChemAxon standardizer [40]: transformation of implicit H atoms to explicit hydrogens and aromatization were performed.

In ISIDA-MLR and ASNN substructural molecular fragments (SMF) developed in the *Laboratory of ChemoInformatics*, Strasbourg, were used. Two different types of fragments are considered (Figure 2): “sequences” (**I**) and “augmented atoms” (**II**).

The sequences correspond to sequential set of atoms linked by chemical bonds where either atom types (C, N, O, ...) or bond types (single, double, ...), or both of them are considered explicitly. In the following, we specify the number of atoms of a given sequence. For instance, **IAB2-6** refers to all sequences containing from 2 to 6 atoms connected by bonds of specified type. For **IA2-6**, the definition is similar, but bond types are omitted. Only shortest paths from one atom to the other are used, as shown in Figure 2.

An “augmented atom” represents a selected atom within its nearest environment including either neighboring atoms and bonds (**AB**), or atoms only (**A**), or bonds only (**B**). The neighborhood is described by concatenating all sequences starting at a given atom and of a given length. For instance **IIAB2-3** refers to fragments concatenating sequences of length of 2 to 3 around the selected atom.

In QSPR models each fragment is monitored by an individual descriptor: its count in the given molecule is the descriptor value.

Insert Figure 2

2D Fuzzy Pharmacophore Triplets [41] (FTP) were used with SQS. FTPs are fuzzy counts of topological pharmacophore triplets. Any triplet of three atoms, classified into pharmacophore types (hydrophobic, aromatic, hydrogen bond donor & acceptor, anion or cation) represents a triangle of pharmacophore types, with edge lengths equal to the inter-atomic topological distances (number of interposed bonds, on the shortest path). The triangles found in a molecule are fuzzily counted, by mapping them onto a basis set of reference triangles. Each molecular triangle will increment the fuzzy population level of similar basis triangles, by an amount proportional to the similarity it bears with respect to the latter.

Other, third-party descriptors employed here include Chemaxon [40] terms: BCUT, LogD, LogP, TPSA (Topological Polar Surface Area) and PF (pairwise Pharmacophore Fingerprints).

2.2 Modeling tools

2.2.1 ASNN

An associative neural network (ASNN) is a combination of an ensemble of 100 feed-forward neural networks and the k-NN (Nearest Neighbor) technique. The method uses correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbor technique. This method corrects a bias of a global model for a considered data case by analyzing the biases of its nearest neighbors determined in the space of calculated models.

For calculations we used the ASNN 1.0 program provided by Dr. Igor V. Tetko[42].

2.2.2 ISIDA [43]

ISIDA/MLR builds Multiple Linear Regression models based on SMF. The MLR algorithm is coupled with a greedy variable selection algorithm [44]. Once a given compound is split into constitutive fragments, the modeled property Y is calculated from the selected fragment population levels N_i using multi linear fitting:

$$Y = c_0 + \sum_{i=1}^n c_i N_i \quad (1)$$

where n is the total descriptor number, c_i are fragment contributions and N_i is the number of fragments of type i .

2.2.3 SQS [45]

The Stochastic QSAR Sampler (SQS) constructs linear and non-linear QSAR models using a Genetic Algorithm (GA) to provide an effective, combined descriptor and nonlinear function selection procedure for the fitting of QSAR models according to the following equation:

$$Y_m = c_0 + \sum_{i=1}^N c_i T_i(D_m^i) \quad (2)$$

where N is the total descriptors number in the model. c_i represent the contribution of the descriptor D^i of molecule m , transformed by a function T which can be null ($T_i(D_m^i) = 0$), linear ($T_i(D_m^i) = D_m^i$) or non-linear.

The final functional form of the QSAR model is thus a linear combination of nonlinearly transformed descriptors. However SQS can be restricted to generate linear models only (T null or linear). This allows, on one hand, to restrict the volume of the problem space and focus on descriptor selection, and on the other hand to estimate the actual benefit of non-linearity by comparing the relative performance of linear versus non-linear models.

2.2.4 5 fold Cross-Validation[31]

For all methods (except ISIDA, which implicitly keeps every 5th compound in the test set) the initial set was randomly split into 5 subsets, each of which was iteratively ignored at the training stage, in order to serve as internal validation set while the four others formed, together, the learning set. For each of these 5 splitting schemes, models were built followed by prediction calculations on the corresponding validation set. Finally, all values calculated for five test sets are merged into one file (combined test set) to analyze overall linear correlations between experimental and predicted property. One can use Determination Coefficient (R^2), Root Mean-Squared Error ($RMSE$) or Mean Average Error (MAE), to estimate the quality of the linear correlation between predicted (Y_{pred}) and experimental (Y_{exp}) data for n compounds. Formulas for the statistical parameters are formulated below.

Determination coefficient:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{pred,i} - y_{exp,i})^2}{\sum_{i=1}^n (y_{exp,i} - \bar{y}_{exp})^2} \quad (3)$$

Root-mean square error:

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

Mean average error:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{pred,i} - y_{exp,i}| \quad (5)$$

2.2.5 Ensemble modeling

ISIDA calculates a Consensus Model (CM) combining the information issued from several models. At the first step, hundreds of models are built using different initial pools of descriptors corresponding to different fragmentation types. Then predictive performance (R^2_{LOO}) is estimated using Leave One Out (LOO) procedure and the best models ($R^2_{LOO} > 0.7$) are combined into a consensus model. In the “leave one out” method, each compound is predicted in turn, based on a model learned from all other compounds. Predicted values are compared to experimental value, to compute leave one out cross-validation determination coefficient. For each compound from the test set, the program computes the property as an arithmetic mean of values obtained with these best models; those leading to outlying values were excluded according to Grubbs’s statistics [46]. Generally, some 30 individual MLR models were used in consensus calculations.

SQS generates thousands of MLR equations for each cross-validation round, each being validated (R^2_v) in terms of the currently left-out fifth of total compound set. Using the determination coefficient of the best validating model $R^2_{v,max}$ as an internal standard, only equations with $R^2_v \geq 0.8R^2_{v,max}$ are used to build the consensus model at current cross-validation round. Together with this, the three single top R^2_v scoring models are kept for future predictions. Thus, for 5-fold cross-validation, in two (linear and non-linear) model search modes, $5 \times 2 \times (3+1) = 40$ models will be used for final prediction. The property is estimated as an average of predictions for each model. Alternatively, only the average of outputs of models including the

molecule in its applicability domain is returned. Depending on the number of the latter models, the most robust of the two alternative estimates is returned as the ‘final’ prediction, together with a trustworthiness score (see AD section below).

Several hundreds of ASNN models were built using different fragmentation types. Only 52 models, with $R^2 > 0.8$, were retained for the consensus predictions on TS1 and TS2.

2.2.6 Applicability Domain (AD)

QSPR models are obtained on a training set, which, no matter how large, may never represent a significant sample of the entire chemical space. An applicability domain defines a region of the chemical space for which the training set constitutes a good sample. In this region, statistical models can deliver reliable predictions. Outside of the AD, the models can no longer be trustfully used.

Two AD types were used for ISIDA MLR and ASNN models:

Fragment control doesn't allow predicting molecules containing new fragments, unknown to the training set.

Bounding Box allows prediction if the value of the descriptor of the new molecule, included in the model, lies between a maximum and a minimum value. These bounds are defined as the first and last 0.5 percentile respectively, from the empirical distribution of each descriptor. This constitutes an estimate of a 99% confidence interval for the descriptor.

For SQS models the main applicability criteria are:

- For each individual model, the Dice dissimilarity of the predicted compound with respect to its closest well-predicted training molecule, MINDIS-OK, determines whether the model is “competent” to predict that molecule or not.
- Globally, the variance of the prediction of individual models included in the consensus – is specifically calculated over all the models, and over the “competent” models, in the sense above. For details on both these criteria, see [47].

These two aspects were empirically combined into a categorical trustworthiness assessment returning OPTIMAL, GOOD, MEDIUM, POOR or NONE.

For TS1, compounds labeled “OPTIMAL” and “GOOD” were considered within the AD, whereas in the case of TS2 “MEDIUM” molecules were also added, in order to maintain the AD coverage of predicted compounds above 50% of the entire set.

3. RESULTS AND DISCUSSION

3.1 Preliminary Benchmarking

Trying to reproduce the models performance reported by Hall and Story [24] we built ISIDA MLR and ASNN models, using ISIDA fragment counts, on two different training sets of the same size (268 molecules) then validated on two different test sets of 30 molecules. The former training and test sets were identical to the ones used in the original publication, whereas the latter followed a random reshuffling scheme of the initial 268+30 molecules. With the former, test set compound prediction resulted in RMSE = 10.0 K and 5.22 K for MLR and ASNN, respectively, whereas the second led to significantly different results (RMSE = 17.0 K and 12.9 K for MLR and ASNN, respectively). The ASNN model following the original training/test splitting scheme effortlessly matched the previously reported results (also based on neural networks, but using E-state descriptors). Hence, ISIDA fragments seem as competent with respect to T_b modeling as other state-of-the-art descriptors. However, training/test reshuffling showed that the initial split – albeit random, as reported in [23] – happened to be a very lucky, modeling-friendly draw. Different train/test configurations are less prone to allow such a smooth extrapolation of the learned model to test set compounds. This clearly demonstrates the necessity of cross-validation procedure, in order to escape the potential bias due to a unique training/test set split, as it has been done in the paper by Artemenko et al. [39]

3.2 Model Building & Validation

Table 2 summarizes the performance of our models in 5-fold cross validation for the given set of 2098 organic compounds and external test sets. Herein, ISIDA MLR and ASNN models were built on hand of ISIDA fragment counts, whereas SQS operated on fuzzy pharmacophore triplet counts 2D-FPT and ChemAxon terms. The results below concern ensemble models, *i.e.*, for each of the five alternative training/test splitting schemes, ensemble models were built on the basis of training compounds, and used to predict test molecules as explained in section 2.2.5.

RMSE and MAE are taken by comparing experimental values to the calculated value corresponding to the splitting scheme in which the compound had been in the test set.

Insert Table 2

In terms of cross-validation results, the performance of SQS is slightly less good than ISIDA MLR models. These models are based on different descriptor sets, where ISIDA MLR and ASNN are a consensus over models based on various fragmentation schemes, capturing different structural aspects. The SQS descriptor pool focuses on the physico-chemical/pharmacophoric nature of the compounds, and is as such more homogeneous. However, SQS models are the only to pass the external validation test over the entire pool of TS1 molecules. The behavior of the large consensus over thousands of individual equations is very robust and less prone to over-optimistic estimation of quality during cross-validation. The lesser cross-validation performance of SQS has, at least, the merit of being more realistic – optimistic with respect to the actual external validation results, indeed, but not overoptimistic like the ones of ISIDA MLR/ASNN.

Insert Figure 3

Analysis of linear correlations between predicted and experimental points (Figure 3) reveals some "outliers": incorrect predicted compounds whatever the method used. Thus, eight major "outliers" were identified. Molecules are quite different structurally, nevertheless one can observe that the majority contains several halogens and at least one atom of oxygen (Heptanoyl chloride, 2,2,2-Trichloro-1-ethoxyethanol, Tetrafluorosuccinic anhydride, 1-Methyl-3-piperidinemethanol, β -Propiolactone). We distinguish also a group of small fluorinated compounds (Tetrafluoro-ethene, 1,1-difluoroethylene, Hexafluoroethane) with T_b located in the in the lower region of the curve, where the points are sparse. Most of those compounds are also unsatisfactory predicted by the Gani model, but not by ACD/Labs. This let us wonder if it is a question of method or maybe experimental values are erroneous, which ACD/Labs is not able to identify, because these values are in its modeling set.

Data are taken mostly from handbooks[33] or catalogs [48] that do not provide data sources. The only way to check data for one compound is to collect it from different sources and to verify that the variance of the values is very small. In this case data is considered to be correct.

3.3 External Validation

Prediction errors for test subsets within the applicability domain of each of three models (see Table 2) are expectedly higher than in 5-CV and two external test sets TS1 and TS2. Exclusion of the molecules outside of AD decreases the test coverage to 54-70 % (TS1) and 85-98 % (TS2). Relatively small coverage observed for TS1 can be explained by the fact that this test set is much more different from the training set than TS2 (see *Data Preparation* section).

Regression Error Curves (REC)[49] have been used to compare the predictive performance of different machine-learning methods used in this paper with that of commercially available tools. These curves plot the error tolerance on the *x*-axis versus the percentage of points predicted within the tolerance on the *y*-axis. It allows quick visual estimation of the relative merit of many regression curves by examining their relative positions: the fastest raising curve corresponds to the most performing method. One can see that among methods applied in this work, ASNN models predict better than ISIDA MLR models, which outperforms SQS (Figure 4). The Gani group-contribution-based method [18] implemented in the *Propred* program v4.1 [50] is much less predictive than the models developed in this work. On the other hand, the ACD/Labs models do best. However, one can't exclude that some of molecules in TS1 and TS2 could be part of ACD/Labs training set which might explain small prediction errors of these models.

Insert Figure 4

3.4 Software implementation

The developed models (ASNN, MLR and SQS) were implemented in software which is publically available at <http://infochim.u-strasbg.fr/webserv/VSEngine.html>, via a WEB interface. A test molecule could be either manually prepared using online structure sketcher or uploaded from the file. The user can also submit for calculation a dataset containing many molecules. For each molecule, the program calculates an arithmetic mean of T_b values predicted by individual models which have been accepted by applicability domain and related variance.

ASN and MLR models can be accessed through a simple interface, showing the output (predicted value and AD status) of each individual model, and a consensus prediction for each molecule.

For SQS predictions, T_b are given together with their variance, the trustworthiness level (OPTIMAL, GOOD, MEDIUM, POOR or NONE) and information supporting the returned trustworthiness level (Figure 5). The trustworthiness is defined by the number of individual models accepted by AD and the variance value. Predictions are considered reliable if the number of models is larger than 5 and the variance is rather small. The program is user friendly and supports batch calculations of large compound collections.

Insert Figure 5

4. CONCLUSION

Predictive ensemble models for boiling point of organic molecules has been developed using ISIDA fragment descriptors and both linear (multiple linear regression) and non-linear (neural networks) machine-learning methods. Each model has been validated in 5-fold external cross-validation for the training set and on two external test sets. Several different approaches have been used to define the models' applicability domain. Application of AD improves predictive performance of the models but decreases the test set coverage. Root Mean Square Error of predictions is about 20K, which corresponds to the performance of previously published models [16, 26] built on large and diverse datasets. This is encouraging, albeit not strict benchmarking, for results refer to different data sets.

Direct comparison with commercial boiling point predictors does not allow drawing irrefutable conclusions either, in as far as it is unknown whether the “external” test compounds were or not included in the third-party model training set. This question is peculiarly relevant for the comparison to the excellently performing ACDLabs tool, which returns experimental data if the compound to predict is part of its database. Potential statistical bias notwithstanding, practically speaking ACDLabs is the clear winner of the herein performed benchmarking test TS2. The hypothesized partial overlap of TS2 and its example database would – in this case – be a merit, not an issue of the ACDLabs tool. The problem is that the quality of extrapolation of

ACDLabs models to unknown structures cannot be objectively determined. The other tested software, based on Gani group contributions, returned clearly worse predictions than the herein reported models.

The fact that many outliers seem to be caused by erroneous experimental data or structures, demonstrates that our models are able to recognize these inconsistencies, and therefore they can be suggested as errors identification tools.

Developed models have been implemented in WEB-based software freely available for users via any INTERNET browser. For any test molecule, the program reports both predicted T_b value and the trustworthiness estimation. To our knowledge, this is the only publicly available tool for T_b .

ACKNOWLEDGEMENTS

The authors thank Dr. Vitaly Solov'ev for his help and discussions, Dr. Igor V. Tetko for providing us with the ASNN software and Dr. Baskin for providing us with the external data for test set2.

FIGURES

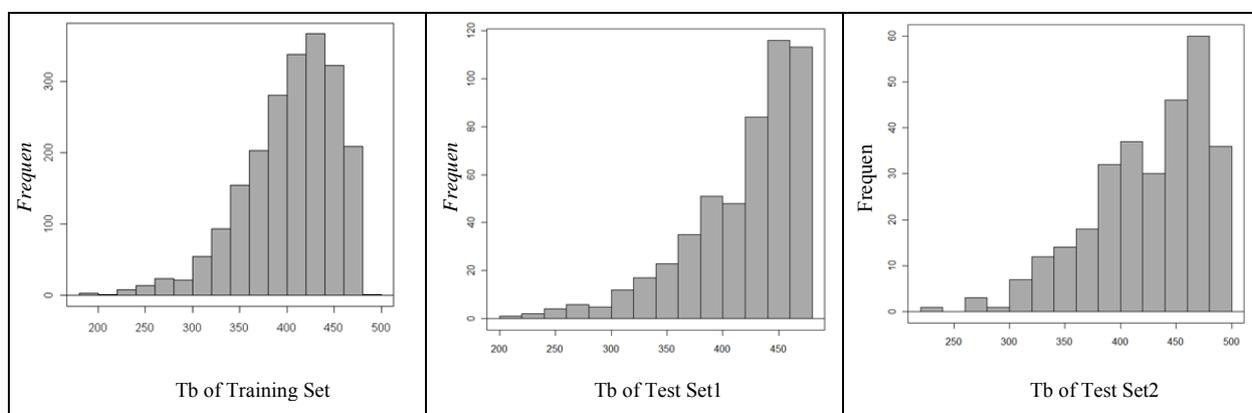
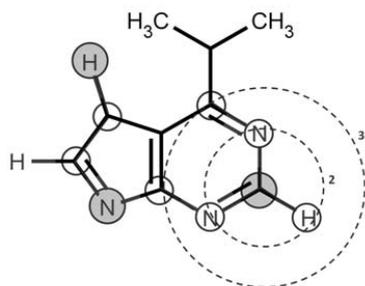


Figure 1 Boiling point (T_b) distribution for data sets used in this study



	SEQUENCES (I)	AUGMENTED ATOMS (II)
Atoms and Bonds	N=C-C-H; N=C-C; C-C-H; N=C; C-C; C-H;	C(=N)(-N)(-H)(=N-C)(-N=C)
Atoms	NCCH; NCC; CCH; NC; CC; CH;	C(N)(N)(H)(NC)(NC)
Bonds		C(=)(-)(-)(=)(=)

Figure 2 ISIDA Fragmentation. Two classes of substructural fragments: atom/bond sequences and augmented atoms. From top to bottom: the sequences (I) correspond to the I (AB, 2-4) and I (A, 2-4) types involving the shortest paths between each pair of atoms. Augmented atoms (II) correspond to the II (AB, 2-3), II (A, 2-3) and II (B, 2-3) types.

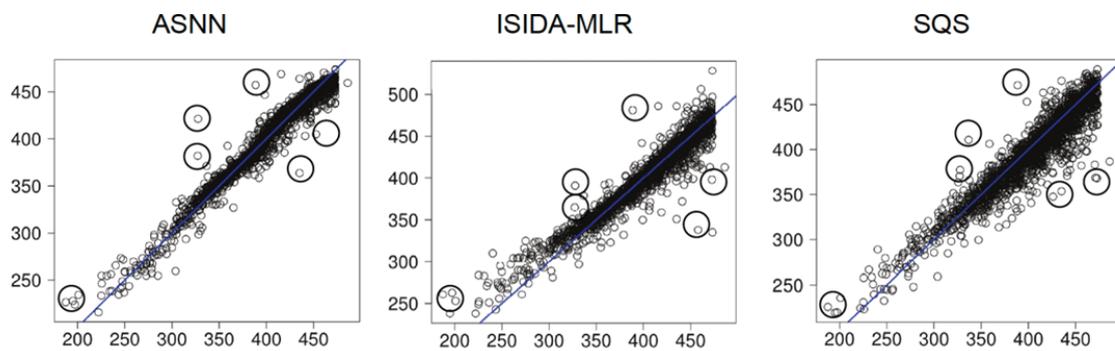


Figure 3 Correlation of predicted in 5-CV and experimental T_b . for different machine-learning methods.

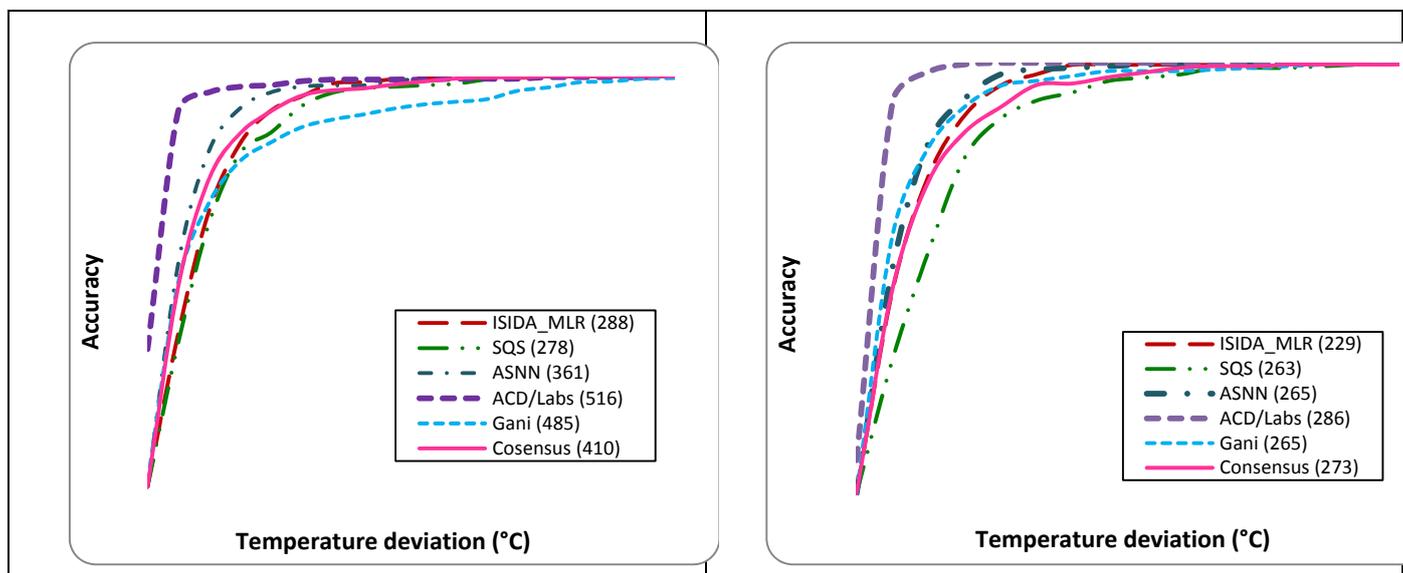


Figure 4. REC curves of different models predictions for TS1 (left) and TS2(right). For each model the number of predicted compounds is indicated. "Consensus" results from the average of ISIDA-MLR, SQS and ASNN predictions.

Column Header Legend

A - #Mol: Current number of the molecule in the submitted set
 B - STRUCTURE: standardized STRUCTURE serving as basis of descriptor calculation
 C - NMOD: number of local models including current compound in their applicability domains: if there are none, the total number of local models is given
 D - Tbp0: Consensus Average of predicted property over all the AVAILABLE models, ignoring applicability domain considerations
 E - VAR0: Consensus Variances of predicted property over all the AVAILABLE models, ignoring applicability domain considerations
 F - TbpApp: Consensus Average of predicted property over all the APPLICABLE models - if missing, or if NMOD is low, report to the (less trustworthy) Tbp0
 G - VARApp: Consensus Variances of predicted property over all the APPLICABLE models - if missing, or if NMOD is low, report to the (less trustworthy) VAR0
 H - **Tbp: Returned prediction - the most trustworthy of Tbp0 and Tbp1**
 I - VAR: Variances associated to Returned prediction
 J - TRUST: Generic estimation of the degree of trust associated to this prediction
 K - REASON: explanation of the trust estimator

#Mol	STRUCTURE	NMOD	Tbp0	VAR0	TbpApp	VARApp	Tbp	VAR	TRUST	REASON
1		10	482.11	9.651	481.30	6.404	481.30	6.404	OPTIMAL	-
2		4	480.79	21.024	481.31	5.951	480.79	21.024	MEDIUM	- There are too few (less than 5) local models containing molecule within applicability domain - global consensus is preferred - Furthermore, the other local models disagree with the prediction of the minority containing compound inside their applicability domain
3		3	466.86	22.432	466.67	17.518	466.86	22.432	POOR	- There are too few (less than 5) local models containing molecule within applicability domain - global consensus is preferred - Furthermore, the other local models disagree with the prediction of the minority containing compound inside their applicability domain - Individual models failed to reach unanimity - prediction variance exceeds 4.0% of the property range width
4		6	424.98	19.500	432.74	23.571	432.74	23.571	GOOD	- Individual models failed to reach unanimity - prediction variance exceeds 4.0% of the property range width

Figure 5. WEB interface for T_b predictions. Example of output for 4 different organic compounds.

TABLES

Table 1 Summary of models for T_b prediction

Reference	Method ^c	Number of compounds		Prediction error (K)	
		Training set	Test set	Training	Test set
Nannoolal and Rarey	GC	2812	199	6.6 ^a	6.4 ^a
Ericksen, Wilding et al.	GC	1141	384	7.8 ^a	13 ^a
Marero-Morejon and Pardillo-Fontdevila	GIC	407	99	6.4 ^a	5.2 ^a
Egolf and Jurs	NN	268	30	11.9 ^b	13.2 ^b
Hall and Story	NN	268	30	5.3 ^b	5.9 ^b
Katritzky	MLR	584	28	15.5 ^b	9.7 ^b
Chalk et al.	NN	6000	629	16.5 ^b	19.0 ^b
Artemenko and Baskin	MLR	510	51 ^d	14.6 ^b	21.2 ^b
	NN		51 ^d	9.5 ^b	18.1 ^b

^aMAE calculated by formula 5); ^b RMSE calculated by formula 4).^c method used: neural networks (NN); multiple linear regression (MLR); Group Contribution (GC); group-interaction contribution' (GIC); ^c size of test sets in 10 fold cross validation.

Table 2 Performance of our models in cross validation and external test set.

Model	5-CV			TS1 without AD			TS1 with AD				TS2 without AD			TS2 with AD			
	R ²	RMSE	MAE	R ²	RMSE	MAE	N	R ²	RMSE	MAE	R ²	RMSE	MAE	N	R ²	RMSE	MAE
ISIDA MLR	0.93	13.7	8.6	0.1	47.7	33.9	288	0.80	22.9	17.0	0.59	32.6	21.3	229	0.85	20.4	14.7
ASNN	0.96	9.3	6.2	0.05	48.9	29.4	361	0.86	19.2	13.0	0.64	30.6	17.3	265	0.88	17.8	12.8
SQS	0.87	17.3	12.3	0.37	39.8	28.6	278	0.74	28.2	19.9	0.67	28.96	21.5	261	0.75	25.9	19.2

REFERENCES

- [1] Moller, B., J. Rarey, and D. Ramjugernath, Estimation of the vapour pressure of non-electrolyte organic compounds via group contributions and group interactions, *Journal of Molecular Liquids* 143 (2008) 52-63.
- [2] Ceriani, R., R. Gani, and A.J.A. Meirelles, Prediction of heat capacities and heats of vaporization of organic liquids by group contribution methods, *Fluid Phase Equilibria* 283 (2009) 49-55.
- [3] Nannoolal, Y., J. Rarey, and D. Ramjugernath, Estimation of pure component properties. Part 4: Estimation of the saturated liquid viscosity of non-electrolyte organic compounds via group contributions and group interactions, *Fluid Phase Equilibria* 281 (2009) 97-119.
- [4] Nannoolal, Y., J. Rarey, and D. Ramjugernath, Estimation of pure component properties. Part 2. Estimation of critical property data by group contribution, *Fluid Phase Equilibria* 252 (2007) 1-27.
- [5] CAS REGISTRY, 2010, See: <http://www.cas.org>.
- [6] Wagner, A.B., SciFinder Scholar 2006: An empirical analysis of research topic query processing, *Journal of Chemical Information and Modeling* 46 (2006) 767-774.
- [7] Katritzky, A.R., et al., Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties, *Journal of Chemical Information and Computer Sciences* 40 (2000) 1-18.
- [8] Toropov, A.A., et al., Using the maximal topological distance matrix for QSPR modeling of the boiling points of cyclic hydrocarbons, *J. Struct. Chem.* 40 (1999) 169-172.
- [9] Tsygankova, I.G., Combination of fragmental and topological descriptors for QSPR estimations of boiling temperature, *Qsar & Combinatorial Science* 23 (2004) 629-636.
- [10] Balaban, A.T., et al., Correlations between Chemical-Structure and Normal Boiling Points of Halogenated Alkanes C-1-C-4, *Journal of Chemical Information and Computer Sciences* 32 (1992) 233-237.
- [11] Rucker, C., M. Meringer, and A. Kerber, QSPR using MOLGEN-QSPR: The challenge of fluoroalkane boiling points, *Journal of Chemical Information and Modeling* 45 (2005) 74-80.
- [12] Toropov, A.A., et al., Testing the atomic orbital graph as a basis for QSPR modeling of the boiling points of haloalkanes, *J. Struct. Chem.* 40 (1999) 950-958.
- [13] Bunz, A.P., B. Braun, and R. Janowsky, Application of quantitative structure-performance relationship and neural network models for the prediction of physical properties from molecular structure, *Industrial & Engineering Chemistry Research* 37 (1998) 3043-3051.
- [14] Kompany-Zareh, M., A QSPR study of boiling point of saturated alcohols using genetic algorithm, *Acta Chimica Slovenica* 50 (2003) 259-273.
- [15] Roy, K. and A. Saha, QSPR with TAU indices: Boiling points of sulfides and thiols, *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* 43 (2004) 1369-1376.
- [16] Chalk, A.J., B. Beck, and T. Clark, A quantum mechanical/neural net model for boiling points with error estimation, *Journal of Chemical Information and Computer Sciences* 41 (2001) 457-462.
- [17] Joback, K.G. and R.C. Reid, Estimation of Pure-Component Properties from Group-Contributions, *Chemical Engineering Communications* 57 (1987) 233-243.

- [18] Constantinou, L. and R. Gani, New Group-Contribution Method for Estimating Properties of Pure Compounds, *Aiche Journal* 40 (1994) 1697-1710.
- [19] Stein, S.E. and R.L. Brown, Estimation of Normal Boiling Points from Group Contributions, *Journal of Chemical Information and Computer Sciences* 34 (1994) 581-587.
- [20] Marrero-Morejon, J. and E. Pardillo-Fontdevila, Estimation of pure compound properties using group-interaction contributions, *AIChE Journal* 45 (1999) 615-621.
- [21] Nannoolal, Y., et al., Estimation of pure component properties Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions, *Fluid Phase Equilibria* 226 (2004) 45-63.
- [22] Ericksen, D., et al., Use of the DIPPR database for development of QSPR correlations: Normal boiling point, *Journal of Chemical and Engineering Data* 47 (2002) 1293-1302.
- [23] Egolf, L.M., M.D. Wessel, and P.C. Jurs, Prediction of Boiling Points and Critical-Temperatures of Industrially Important Organic-Compounds from Molecular-Structure, *Journal of Chemical Information and Computer Sciences* 34 (1994) 947-956.
- [24] Hall, L.H. and C.T. Story, Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks, *Journal of Chemical Information and Computer Sciences* 36 (1996) 1004-1014.
- [25] Katritzky, A.R., V.S. Lobanov, and M. Karelson, Normal boiling points for organic compounds: Correlation and prediction by a quantitative structure-property relationship, *Journal of Chemical Information and Computer Sciences* 38 (1998) 28-41.
- [26] Artemenko, N.V., et al., Prediction of physical properties of organic compounds using artificial neural networks within the substructure approach, *Doklady Chemistry* 381 (2001) 317-320.
- [27] DDB Online Property Estimation by the Joback Method, 2010, See: http://www.ddbst.com/en/online/Online_Est_Artist.php.
- [28] Onken, U., J. Rareynies, and J. Gmehling, The Dortmund Data-Bank - a Computerized System for Retrieval, Correlation, and Prediction of Thermodynamic Properties of Mixtures, *International Journal of Thermophysics* 10 (1989) 739-747.
- [29] Boiling Point Estimation by JAVA applet, 1991, See: <http://www.pirika.com/hiroka/PirikaLight/PirikaLight.html> .
- [30] ACD/Labs, 2009, See: <http://www.acdlabs.com/home/>.
- [31] Golub, G.H., M. Heath, and G. Wahba, Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, *Technometrics* 21 (1979) 215-223.
- [32] Processium, See: <http://www.processium.com/>
- [33] Trimble, V., *Crc Handbook of Chemistry and Physics - Weast,Rc, Scientist 1* (1987) 19-19.
- [34] NIST, WebBook de Chimie NIST, 2011, See: <http://webbook.nist.gov/chemistry/>.
- [35] Pesticide Properties DataBase, See: <http://sitem.herts.ac.uk/aeru/footprint/en/Reports/121.htm>.
- [36] PHYSPROP, in: S.R. Corporation, 1994.
- [37] Stein, S.E. and R.L. Brown, Estimation of normal boiling points from group contributions, *Journal of Chemical Information & Computer Sciences* 34 (1994) 581-587.
- [38] Tetko, I.V., The Prediction of Physicochemical Properties, in: *Computational Toxicology*, John Wiley & Sons, Inc.,2006, pp. 240-275.

- [39] Artemenko, N.V., et al., Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks within the Substructure Approach, *Doklady Chemistry* 381 (2001) 317-320.
- [40] ChemAxon, Instant JChem in, 2009.
- [41] Bonachera, F. and D. Horvath, Fuzzy tricentric pharmacophore fingerprints. 2. application of topological fuzzy pharmacophore triplets in quantitative structure-activity relationships, *Journal of Chemical Information and Modeling* 48 (2008) 409-425.
- [42] Tetko, I.V., Associative neural network., *Neural Processing Letters* 16 (2002) 187-199.
- [43] Varnek, A., et al., ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors, *Current Computer-Aided Drug Design* 4 (2008) 191-198.
- [44] Guyon, I. and A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.* 3 (2003) 1157-1182.
- [45] Horvath, D., et al., Stochastic versus stepwise strategies for quantitative structure - Activity relationship generation - How much effort may the mining for successful QSAR models take?, *Journal of Chemical Information and Modeling* 47 (2007) 927-939.
- [46] Grubbs, F.E., Procedures for detecting outlying observations in samples, *Technometrics* 11 (1969) 1-21.
- [47] Horvath, D., G. Marcou, and A. Varnek, Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models, *Journal of Chemical Information and Modeling* 49 (2009) 1762-1776.
- [48] Srouf, E.F. and T.L. Horvath, Identification of a novel murine CD45 negative bone marrow-derived cell population with in vivo marrow repopulating potential following hematopoietic specification in vitro, *Blood* 106 (2005) 490a-490a.
- [49] Bi, J. and K.P. Bennett, Regression Error Characteristic Curves, in: *Proceedings of the 20th International Conference on Machine Learning, Washington DC, 2003*, pp. 45-50.
- [50] Marrero, J. and R. Gani, Group-contribution based estimation of pure component properties, *Fluid Phase Equilibria* 183-184 (2001) 183-208.

8 Solubilité aqueuse

8.1 Introduction

Un projet qui n'est pas directement relié aux autres implique le développement de modèles prédictifs de la solubilité aqueuse. La modélisation de cette propriété est utile pour la société Processium, qui s'intéresse à des nombreuses propriétés physico-chimiques. La solubilité d'un liquide dans un autre est un problème important pour les mélanges, néanmoins cette question n'a pas été traitée en détail et juste une première étape, la modélisation de la solubilité dans l'eau, a été résolue.

La solubilité aqueuse (S , mol/L) représente la quantité maximale de soluté qui se dissout dans un litre d'eau. On peut distinguer au moins trois types de solubilités [1] :

- Intrinsèque (S_i): solubilité d'un composé neutre.
- Apparente (S_{pH}) : solubilité pour les composés ionisables à un certain pH. Cette solubilité est mesurée en solution tampon. $S_{pH} = S_i(1 + 10^{(pH - pK_a)\delta_i})$, où $\delta_i = \{-1, 1\}$ pour les groupements acides ou basiques.
- Dans l'eau pure : solubilité d'un composé dans l'eau pure sans solution tampon. C'est la solubilité apparente au pH de la solution.

Parmi ces trois solubilités, la plus utilisée est la solubilité intrinsèque, plutôt le logarithme de cette valeur, c'est-à-dire le $\log S$. La valeur de $\log S$ dépend de la méthode de mesure utilisée dans chaque laboratoire. La déviation standard est estimée à 0.49 [1] unités des \log pour les solubilités mesurées à la même température. Pour cette raison, les valeurs de la déviation standard ne pourront pas être inférieures à 0.5-0.6 \log pour nos modèles, ce qui correspond à une valeur de R^2 de 0.95 maximum.

8.2 Revue

Le développement de modèles prédictifs pour la solubilité est basé principalement sur trois méthodes :

- Méthodes basées sur des propriétés **physico-chimiques** expérimentales, comme le $\log P$ ou la température de fusion. L'équation présente l'exemple de la GSE (General Solubility Equation) introduite par Yalkowsky et Ran [2].

$$\log S = 0.5 - 0.01 \cdot (MP - 25) - \log P, \text{ où } MP = 25 \text{ pour les liquides} \quad (8-1)$$

- Basées sur des **approches théoriques** qui utilisent des énergies d'hydratation calculées à partir des modèles de solvation implicites ou explicites (Simulations Monte Carlo, calculs quantiques – COSMO-R). Ces méthodes ont besoin des structures 3D optimisés pour les molécules.
- Basées sur des **approches statistiques**. Ces méthodes consistent à développer des modèles QSPR prédictifs pour le logS. Des nombreux modèles ont été développés en utilisant différents types de descripteurs et différentes machines d'apprentissage. Le Tableau 8-1 présente quelques exemples pour lesquelles une base de données importante a été utilisée. Par contre, la stratégie de validation employée est très simple, parce que, dans tous les cas, un seul jeu de données est utilisé pour la validation externe.

Auteur	Nombre de molécules	Méthode	R ²	RMSE
Klopman	1168 (dont 120 pour le test)	MLR	0.91	0.79
Huuskonen	1297 (dont 413 pour le test)	MLR	0.88	0.71
		ANN	0.92	0.60
Tetko	1291 (dont 412 pour le test)	ANN	0.92	0.60
		MLR	0.85	0.81
Votano	4115 ⁱ (dont 10% pour le test)	MLR	0.72	1.01
		PLS	0.72	0.97
		ANN	0.77	0.74
	1849 ⁱⁱ (dont 10% pour le test)	MLR	0.76	0.83
		PLS	0.78	0.81
		ANN	0.84	0.61

Tableau 8-1 Étude des méthodes QSPR pour la prédiction de logS ;

ⁱnon-aromatiques; ⁱⁱaromatiques;

Les calculs de Klopman[3] se basent sur la théorie de la contribution des groupes(2.2.6) sur un jeu d'apprentissage de 1168 molécules. Cette méthode utilise une liste prédéfinie de coefficients de contribution de groupes ainsi que des coefficients obtenus par régression de ce nouveau jeu de données. La performance obtenue (RMSE=0.61) est améliorée (RMSE=0.5) en utilisant une transformation non-linéaire sur le logS. Une validation sur un jeu de test externe de 120 molécules a été effectuée conduisant à une valeur de RMSE de 0.79. Étant donné que le coefficient de détermination n'a pas été rapporté dans la publication, nous l'avons calculé (R²=0.91).

Huuskonen[4] et son équipe ont utilisé un jeu de données de 1297 molécules dont 413 sont choisies aléatoirement pour le set. Deux approches, MLR et ANN, sont

utilisées pour la modélisation. La validation externe conduit à des meilleurs résultats pour les modèles ANN (RMSE=0.60) par rapport aux modèles MLR (RMSE=0.71).

En utilisant les approches MLR et ANN améliorées, Tetko[5] obtient des résultats similaires par rapport au travail de Huuskonen[4] et démontre qu'un nombre réduit des paramètres suffit pour obtenir les mêmes performances.

Votano[6] utilise un grand jeu de données divisé en molécules contenant des cycles aromatiques ou pas. 10% de chaque groupe des molécules est utilisé comme jeu de validation.

Dans tous les cas présentés ci-dessus, la validation est faite juste sur un jeu de données choisi aléatoirement dans le jeu de données initial, ce qui peut biaiser les résultats. Tetko et Huuskonen utilisent, en plus, la validation Leave One Out (LOO), mais celle-ci est beaucoup plus optimiste qu'une validation croisée (5 ou 2 paquets).

Malgré l'existence de nombreux modèles prédictifs de la solubilité aqueuse une méthodologie QSPR plus stricte est nécessaire. Pour cette raison des nouveaux modèles pour la prédiction de la solubilité aqueuse ont été développés.

Le premier pas dans ce sens a été fait par Dr. Denis Fourches[7] pendant son travail de thèse au Laboratoire d'Infochimie. Le modèle développé est un modèle consensus contenant 3 modèles individuels et se basant sur des descripteurs fragmentaux ISIDA. Ce modèle a été implémenté dans le logiciel ISIDA/logS qui est utilisé par la Chimiothèque Nationale Essentielle (CNE).

Étant donné le nombre faible des modèles individuelles contenus dans le consensus et l'absence d'un domaine d'applicabilité, justifie le choix de retravailler sur cette propriété.

Pendant cette thèse le même set de données utilisé par Dr. Fourches a été utilisé afin de développer des nouveaux modèles pour la prédiction de la solubilité aqueuse.

8.3 Méthodologie

8.3.1 Données

Le set de données utilisé est une compilation de jeu de données étudié par Huuskonen[4], Yaffe[8], Jurs[9] et Ran[2]. Nous avons sélectionné 1635 composés organiques (liquides et solides) pour lesquelles la valeur de la solubilité aqueuse est connue. Les valeurs de logS varient entre -11.62 et 1.58.

8.3.2 Machines d'apprentissage et descripteurs

Les modèles QSPR ont été développés avec plusieurs machines d'apprentissage : ISIDA-MLR, SVM, ASNN et SQS. Les descripteurs fragmentaux ISIDA ont été utilisés pour la MLR, SVM et ASNN, tandis que pour le SQS une combinaison des descripteurs fragmentaux ISIDA, FPT[10] et descripteurs Chemaxon[11] a été utilisée.

8.3.3 Résultats

8.3.3.1 Validation croisée

Tableau 8-2 Résultats en 5-CV pour la modélisation de la solubilité (logS)

Méthode	R ² (5-CV)	R ² >Seuil	Nombre de modèles	RMSE	MAE
SVM	0.90	0.85	31	0.70	0.50
SQS	0.85	0.80	11343	0.83	0.63
ASNN	0.90	0.80	31	0.67	0.48
ISIDA-MLR	0.88	0.70	25	0.74	0.55

Les résultats présentés dans le Tableau 8-2 sont obtenus pour des modèles consensus englobant une trentaine de modèles ISIDA-MLR, SVM ou ASNN et plus de 10000 modèles SQS.

Les modèles non-linéaires ASNN et SVM mènent à des meilleurs résultats que les modèles linéaires ISIDA-MLR and SQS, avec une valeur de RMSE de 0.67 et 0.70 respectivement.

La Figure 8-1 représente la distribution des valeurs de logS prédites en fonction des valeurs expérimentales pour les modèles ASNN, SVM, SQS et MLR en validation croisée.

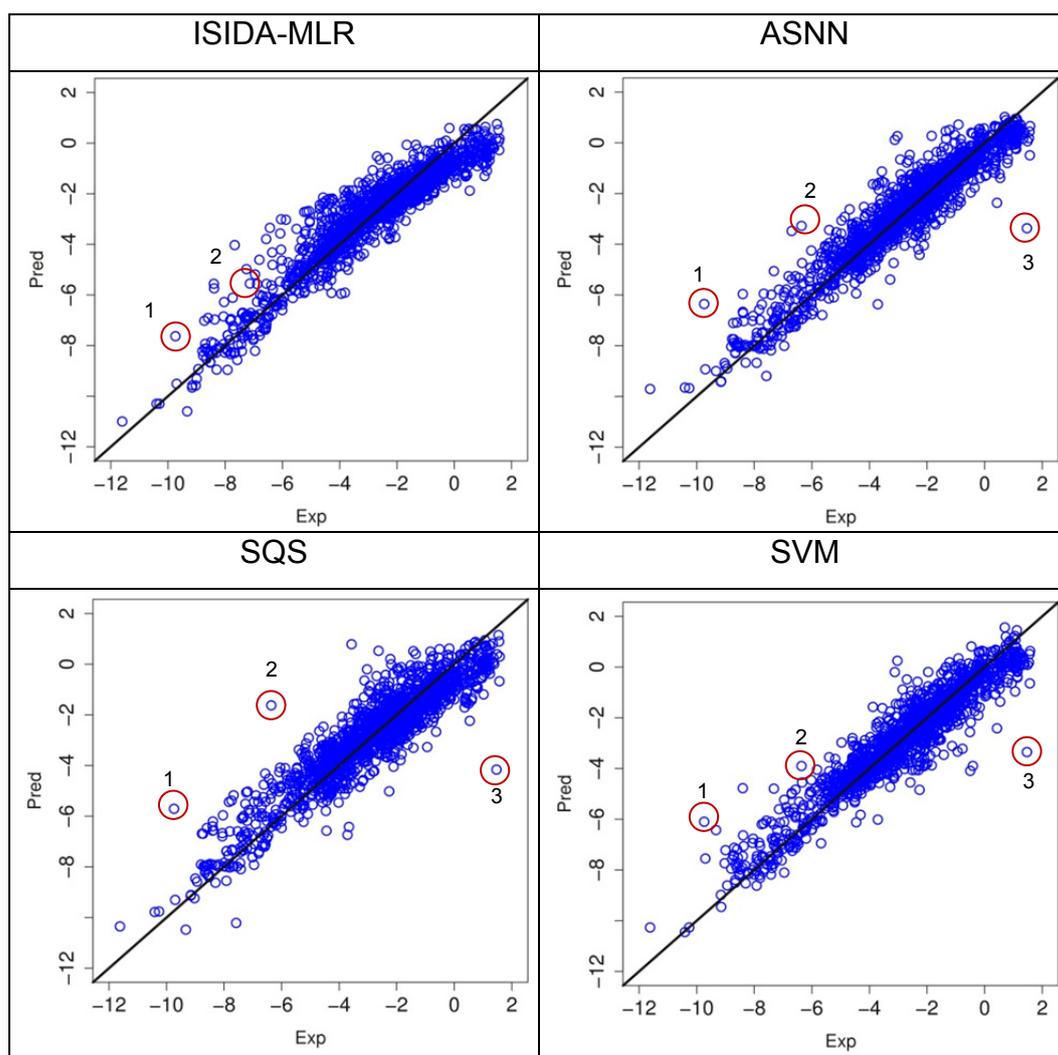
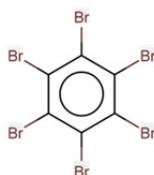


Figure 8-1 Distribution des valeurs de logS prédites en fonction des valeurs expérimentales pour les modèles ASNN, ASVM, SQS et MLR en 5-CV.

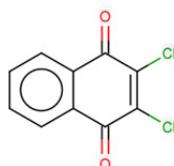
8.3.4 Analyse des points aberrants (outliers)

L'analyse des outliers consiste à déterminer des molécules qui sont constamment mal prédites quiconsoit l'approche utilisée. La Figure 8-1 montre deux molécules qui sont toujours prédites par toutes les modèles avec un grand écart par rapport à la valeur expérimentale. Une troisième molécule est mal prédite par les modèles ASNN, SVM et SQS.

outlier 1



outlier 2



outlier 3

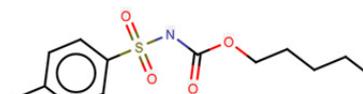


Figure 8-2 Structure des outliers

Les premières deux composés sont très peu solubles avec des valeurs de solubilité de -9.53 et -6.3 respectivement. Le fait que toutes nos modèles surestiment ces valeurs peut être dû à la petite densité des points dans cette zone des valeurs de logS, où les modèles n'ont pas appris assez. De plus dans cette zone où la solubilité est très faible (donc difficilement mesurable) l'erreur peut être beaucoup plus importante. Le troisième composé est mal prédit par ASNN, SVM et SQS, en sous-estimant sa valeur de solubilité. Pour le outlier 1 et 2 les valeurs expérimentales ont été confirmées par la référence [12], tandis pour le troisième composé aucune valeur expérimentale n'a pas été retrouvée, d'où l'impossibilité d'apprécier la véracité de cette valeur.

8.3.4.1 Validation externe

21 composés d'intérêt pharmaceutique ont été utilisés pour une validation externe. Le Tableau 8-3 résume les prédictions données par nos modèles ainsi que celles fournies par Klopman[3], Huuskonen[4] et Tetko[5], sur le même set.

Tableau 8-3 Résultats de la validation sur un jeu de données de 21 molécules. Comparaison avec d'autres travaux. * ANN models; ** MLR models.

Numéro	Composé	Exp	Tetko*	Huuskonen*	Klopman**	ISIDA	SVM	SQS	ASNN
1	2,2',4,5,5'-PCB	-7,89	-7,57	-7,21	-7,9	-7,58	-7,56	-7.54	-7.51
2	Benzocaine	-2,32	-1,63	-1,79	-1,71	-2,15	-1,92	-1.62	-1.95
3	Aspirin	-1,61	-1,81	-1,69	-1,52	-1,92	-2,02	-1.51	-1.54
4	Theophylline	-1,37	-0,69	-1,71	-1,07	-1,58	-1,41	-1.54	-1.47
5	Antipyrine	0,39	-0,89	-1,29	-2,76	-1,34	-0,97	-1.88	-0.99
6	Atrazine	-3,55	-3,7	-3,51	-3,05	-3,75	-4,11	-3.31	-4.06
7	Phenobarbital	-2,34	-2,89	-2,97	-2,08	-2,47	-2,44	-2.31	-2.50
8	Diuron	-3,76	-3,01	-2,86	-2,85	-3,41	-3,74	-3.2	-3.51
9	Nitrofurantoin	-3,38	-3,09	-3,42	-2,19	-1,22	-2,90	-3.83	-2.76
10	Phenytoin	-3,99	-3,52	-3,4	-3,47	-3,18	-2,90	-2.71	-3.47
11	Diazepam	-3,76	-4,37	-4,05	-6,54	-3,95	-3,93	-3.89	-4.33
12	Testosterone	-4,07	-4,13	-3,98	-5,17	-5,05	-4,29	-4.35	-4.30
13	Lindane	-4,6	-4,91	-4,71	-4,88	-4,13	-4,90	-5.01	-4.90
14	Parathion	-4,29	-4,31	-4,13	-3,94	-3,72	-3,50	-5.5	-4.42
15	Diazinon	-3,76	-3,43	-4,01	-5,29	-3,54	-3,08	-4.43	-3.68
16	Phenolphthalein	-2,9	-4,31	-3,99	-4,48	-4,10	-3,31	-5.14	-3.99
17	Malathion	-3,36	-3,73	-3,24	-2,94	-2,99	-3,42	-2.48	-2.62
18	Chlorpyrifos	-5,67	-5,31	-5,61	-5,77	-4,24	-4,80	-4.89	-5.60
19	prostaglandin	-2,47	-3,52	-3,29	-4,21	-3,55	-3,06	-4.02	-3.24
20	4,4-DDT	-8,08	-7,59	-7,67	-8	-7,46	-6,90	-6.96	-7.20
21	Chlordane	-6,86	-7,23	-7,29	-7,55	-6,68	-6,73	-7.25	-6.99
R²			0,9 (0.92)	0,91	0,63 (0.80)	0.88	0.91	0.75	0.92
RMSE			0,63 (0.56)	0,6	1,21 (0.84)	0.70	0.62	0.99	0.57

Les modèles SVM et ASNN donnent des résultats similaires, voire meilleurs par rapport aux modèles de Tetko et Huuskonen. Les modèles ISIDA-MLR et SQS sont moins performants mais toujours meilleures que les modèles de Klopman.

8.3.5 Domaine d'applicabilité

Le domaine d'applicabilité d'un modèle est important, pour estimer les chances qu'une molécule nouvelle soit bien prédite. Dans le cas des modèles ISIDA-MLR, ASNN, SVM le domaine d'applicabilité Fragment Control et MinMax on été utilisés simultanément. Même si toutes les molécules ont été prédites, dans le consessus ont été retenus entre 3 et 25 modèles, ce nombre dependant de la molécule. L'utilisation du DA permet de diminuer l'erreur de 0.2 logS.

Le DA pour les modèles ASNN et SVM n'apporte aucune ameliration.

Pour les modèles SQS la distance Dice à la plus proche molécule de jeu d'apprentissage est utilisée pour décider si une nouvelle molécule se trouve dans le domaine d'applicabilité du modèle. Cela correspond au domaine d'applicabilité MINDIS-OK introduite par Horvath et ses collaborateurs dans une récente publication[13].

8.4 "Solubility challenge"

8.4.1 Présentation

"Solubility challenge" a été un concours organisé par l'équipe de Goodman[14] qui propose 132 molécules d'intérêt pharmaceutique pour lesquelles il a mesuré la solubilité aqueuse intrinsèque en utilisant la même méthode de mesure, afin d'éviter des biais méthodologiques trop importantes.

Le jeu d'entrainement proposé contient 100 composés et Goodman demande de prédire les valeurs pour un jeu de 32 molécules.

Après un « nettoyage » (élimination de molécules trop solubles ou qui se décomposent ou pour lesquelles il n'y a pas des valeurs) 94 composés ont été sélectionnés pour l'entrainement et 28 pour le test.

Ce jeu de composés a servi d'un côté pour la création des modèles, mais aussi comme jeu de test pour les modèles développés précédemment.

8.4.2 Prédictions ISIDA-MLR

77 molécules ne se trouvant pas dans le jeu d'apprentissage ont été sélectionnées dans un jeu de validation additionnel. Les résultats obtenus pour ce

jeu sont moins bons que ceux obtenus précédemment. Le meilleur modèle est donné par ISIDA MLR avec une erreur de prédiction de 1.07 (RMSE) et un coefficient de détermination de 0.45. La difficulté de prédire ces composés peut être due, d'un côté, à un surapprentissage des modèles et de l'autre à la difficulté de décrire ces molécules par les descripteurs fragmentaux. Ce jeu de données était annoncé dans la publication comme étant difficile à modéliser, d'où le défi de ce concours.

8.4.2.1 Modélisation

Comme a été demandé pour le concours, des modèles ont été développés en se basant sur le jeu de données de 94 molécules et appliqués au jeu de données de 28 molécules. La Figure 8-3 montre les performances des modèles développés par 100 différentes équipes et le positionnement de nos modèles par rapport aux autres. On constate que les modèles ISIDA MLR ont de bonnes performances se situant parmi les premiers 10 meilleurs modèles. Les autres modèles sont moins bons mais toujours meilleurs que la majorité des modèles présentés dans le concours.

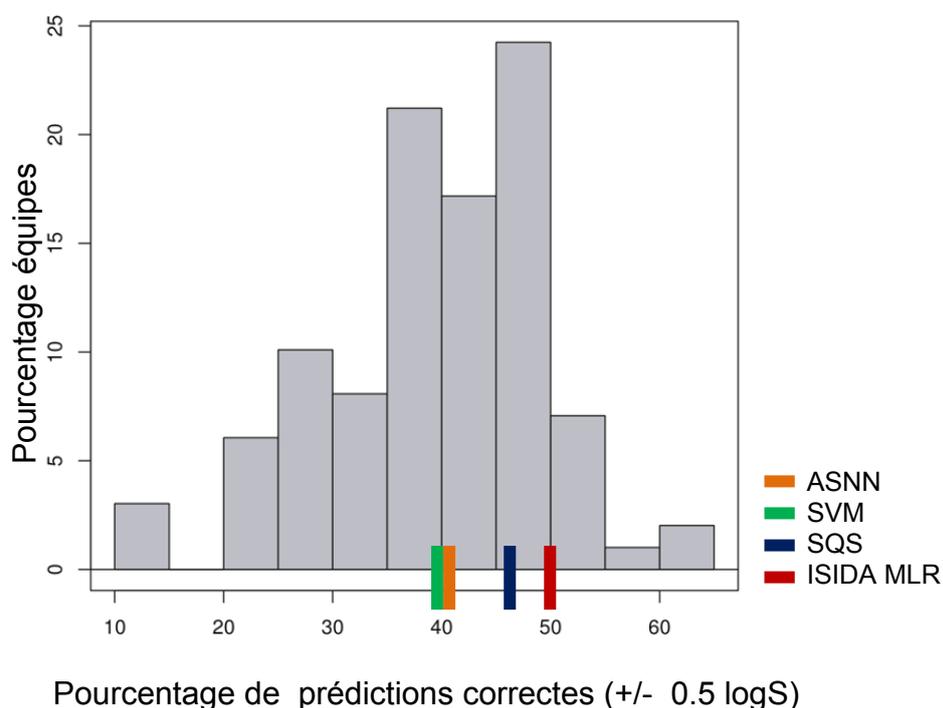


Figure 8-3 Performances de nos modèles par rapport aux ceux d'autres équipes

8.5 Prédiction des molécules de la Chimiothèque Nationale Essentielle (CNE)

Dans le cadre d'un projet concernant des molécules de la CNE un jeu de données à été utilisé pour prédire leur solubilité. Ces prédictions ont pu être comparées aux valeurs de solubilité expérimentales, toutes mesurées dans les mêmes conditions.

8.5.1 Données

Parmi les 313 molécules du jeu initial de molécules ayant une valeur expérimentale de la solubilité, 191 ont été retenus pour la prédiction. Ce filtrage a été effectué afin d'enlever toutes molécules pour lesquelles la mesure de la solubilité était incertaine (solution non saturée, composé non pur, plusieurs composés détectables, etc.).

8.5.2 Résultats

8.5.2.1 Prédiction

191 molécules ont été prédites en utilisant les modèles ISIDA MLR et SQS implémentés dans les logiciels Predictor et Virtual Screening, respectivement. Les logiciels MOE, ALOGPS et ISIDA_logS ont été également utilisé.

Dans la Figure 8-4 les performances de prédictions des quatre outils sont comparées, ainsi que leur couverture sur l'ensemble de données. Les modèles SQS et ISIDA MLR ont des meilleures performances, néanmoins l'utilisation d'un DA élimine un nombre important de composés. Les modèles ISIDA-MLR restent un bon compromis qualité prédiction/recouvrement : Pour plus de 79% de composés prédits l'erreur de prédiction est de 1.17 logS (RMSE). Cette valeur est similaire à celle obtenue pour le jeu de données antérieur (celui de la « Solubility challenge »).

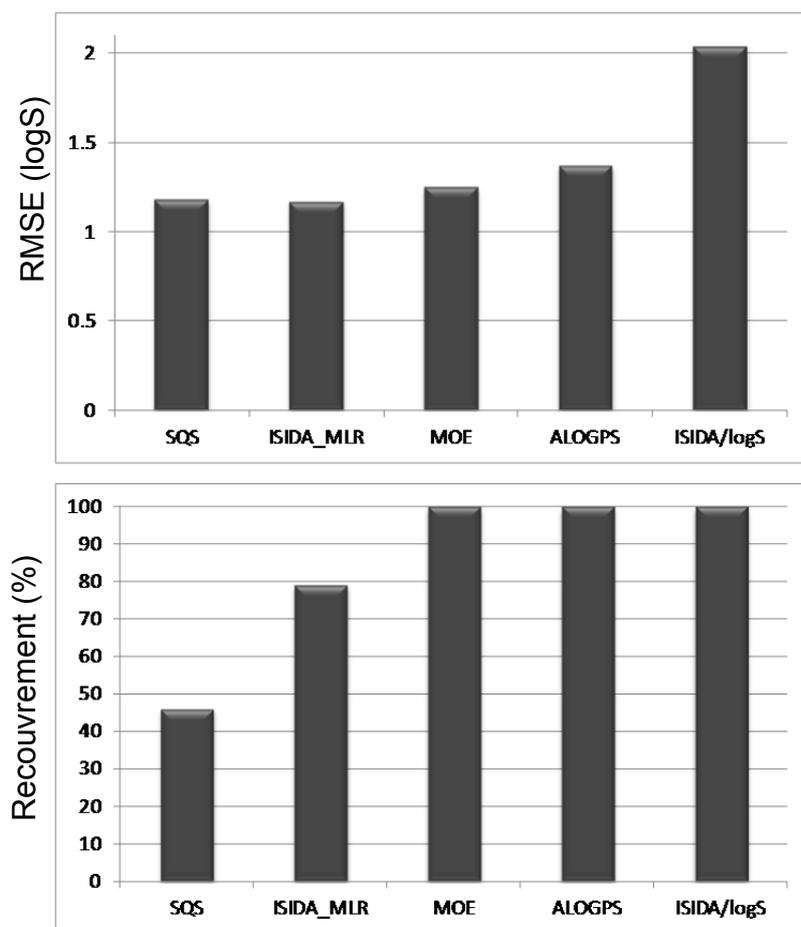


Figure 8-4 Prédications de la solubilité par le modèles SQS, ISIDA MLR, MOE et ALOGPS.

8.6 Conclusion

Ce projet a visé au développement de modèles prédictifs de la solubilité aqueuse. Il constitue une première étape vers l'étude de la miscibilité de deux composés pour former un mélange.

Cette propriété a été modélisée pour un jeu de données de 1635 molécules dont la valeur de la solubilité ($\log S$) est comprise entre -11,62 et 1,58. Les approches ISIDA MLR, SVM, ASNN et SQS ont été utilisées pour le développement des modèles. De plus, l'efficacité des modèles consensus a été démontrée.

Deux jeux de données supplémentaires, de Goodman[14] et de la CNE, ont été prédites avec un erreur moyenne de 1,1 $\log S$, ce qui est supérieure aux erreurs obtenues en validation croisée (de 0,67 à 0,83 $\log S$). Ceci est peut-être du à la complexité structurale et à la taille des composés de ces deux jeux, où 60% des molécules ont une masse moléculaire supérieure à 250 par rapport à seulement 25% dans le jeu d'entraînement. Un domaine d'applicabilité plus rigoureux pourra

améliorer les résultats, en acceptant que des molécules proches au jeu d'entraînement.

La comparaison des prédictions sur les molécules de la CNE avec le modèle ISIDA/logS, montre la supériorité de nos modèles. Par conséquent, ces modèles pourront être englobées dans la CNE, afin de remplacer l'ancien modèle.

8.7 Références

1. Tetko, I.V., K.V. Balakin, and N.P. Savchuk, *In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions*. Current Medicinal Chemistry, 2006. **13**(2): p. 223-241.
2. Ran, Y.Q. and S.H. Yalkowsky, *Prediction of drug solubility by the general solubility equation (GSE)*. Journal of Chemical Information and Computer Sciences, 2001. **41**(2): p. 354-357.
3. Klopman, G. and H. Zhu, *Estimation of the aqueous solubility of organic molecules by the group contribution approach*. Journal of Chemical Information and Computer Sciences, 2001. **41**(2): p. 439-445.
4. Huuskonen, J., *Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology*. Journal of Chemical Information and Computer Sciences, 2000. **40**(3): p. 773-777.
5. Tetko, I.V., et al., *Estimation of aqueous solubility of chemical compounds using E-state indices*. Journal of Chemical Information and Computer Sciences, 2001. **41**(6): p. 1488-1493.
6. Votano, J.R., et al., *Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation*. Chemistry & Biodiversity, 2004. **1**(11): p. 1829-1841.
7. Fourches, D., *Modèles multiples en QSAR/QSPR: Développement de nouvelles approches et leurs applications au design "in silico" de nouveaux extractants de métaux, aux propriétés ADMETox ainsi qu'à différentes activités biologiques de molécules organiques*. 2007.
8. Cohen, Y., et al., *A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) for predicting aqueous solubility of organic compounds*. Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1177-1207.
9. Jurs, P.C. and N.R. McElroy, *Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure*. Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1237-1247.
10. Bonachera, F., et al., *Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes*. Journal of Chemical Information and Modeling, 2006. **46**(6): p. 2457-2477.
11. ChemAxon, *Instant JChem* 2009.
12. *ChemIDplus*. [cited 2010 9.02.2010]; Available from: <http://chem.sis.nlm.nih.gov/chemidplus/>.
13. Dragos, H., M. Gilles, and V. Alexandre, *Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models*. Journal of Chemical Information and Modeling, 2009. **49**(7): p. 1762-1776.
14. Llinàs, A., R.C. Glen, and J.M. Goodman, *Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?* Journal of Chemical Information and Modeling, 2008. **48**(7): p. 1289-1303.

TROISIEME PARTIE:
**Développement d'un outil de
prédiction**

9 Développement

La dernière partie de ce travail a été consacrée au développement d'un logiciel pour valoriser les résultats précédents et ainsi favoriser leur application à des cas concrets de procédés.

Deux logiciels ont été développés:

Le **Multifragmentor** et le **MixturesPredictor**.

Ces deux logiciels ont été développés en FreePascal, un langage de programmation orienté objet utilisé souvent au laboratoire d'Infochimie.

Les applications sont multiplateformes pouvant être lancées sous différents systèmes d'exploitation (Windows, Mac OS X, Linux...).

9.1 "Multifragmentor"

9.1.1 Présentation

Le "Multifragmentor" permet de créer la matrice de descripteurs pour des mélanges. Cette matrice est utilisée ensuite par différentes machines d'apprentissage pour le développement des modèles QSPR.

La différence du "Fragmentor" déjà existant qui fonctionne pour une seule molécule, ce logiciel fragmente plusieurs molécules à la fois (deux ou plus). Le principe de fonctionnement est le suivant: chaque molécule est fragmenté séparément et ensuite les matrices de descripteurs sont soit concaténées soit symétrisées comme décrit dans les chapitres précédents (chapitre 3,4,5 et 6).

9.1.2 Fonctionnement

Le "Multifragmentor" se lance en ligne de commande et utilise plusieurs options :

--help: affiche la liste des options avec leur signification.

-i: le nom de fichier SDF en entrée. Pour chaque entrée du mélange il faut écrire d'abord "-i", ensuite le nom de fichier.

-o: le nom du fichier en sortie. Deux fichiers vont être créés: un fichier avec l'extension *.hdr* contenant la liste des fragments et un fichier avec l'extension *.svn* qui est le format d'entrée pour le logiciel LIBSVN. Ce fichier est facilement convertible en d'autres formats comme *.asnn* utilisé par ASNN ou PLS ou *.arff* qui est le format de Weka.

-t: le type de fragmentation souhaitée (cf. Tableau 9-1).

Le tableau ci-joint montre les types de fragments avec leur correspondance pour le logiciel.

Tableau 9-1 Correspondances t – type de fragment

Type fragment	AC	IA	IB	IAB	IIA	IIB	IIAB	IIHy	IIIA	IIIB	IIAB	IVA	IVB	IVAB
t	0	1	2	3	4	5	6	7	8	9	10	11	12	13

-l: la longueur minimale de chaque fragment. Elle est comprise entre 2 et 15.

-u: la longueur maximale du chaque fragment. Elle est comprise entre 2 et 15.

-w: le nom du champ dans le fichier SDF contenant les poids des molécules dans chaque mélange. Dans notre cas le poids est représenté par la fraction molaire d'un composé dans le mélange.

-c: le type de combinaison à utiliser pour les descripteurs des molécules de chaque mélange: 1 représente concaténation (cf. chapitre 4), 2 représente symétrisation (cf. chapitre 3).

-h: le nom du fichier header de référence pour la fragmentation. Dans le fichier header de sortie les premiers fragments dans la liste sont ceux appartenant au fichier header de référence. S'il y a des fragments nouveaux par rapport à la référence ils sont ajoutés à la fin du fichier header de sortie.

-strict: cette option impose que seulement les fragments qui se trouve dans le fichier header de référence doivent se trouver dans le fichier header de sortie.

Il faut souligner que le fichier header de référence est un fichier obtenu en ne fragmentant que des composés individuels, c'est-à-dire un seul fichier SDF.

9.2 "MixturesPredictor"

9.2.1 Présentation

"MixturesPredictor" est une application standalone. Elle englobe tous les modèles QSPR pour la prédiction des propriétés physico-chimiques des composés purs et des mélanges, développés lors de cette thèse, afin d'être utilisés pour la prédiction de nouveaux composés/mélanges de façon simple et rapide. "MixturesPredictor" est la mise à jour du logiciel Predictor déjà développé au laboratoire utilisé pour la prédiction de propriété des corps purs (constante de complexation des métaux et propriétés ADME).

Dans le "MixturesPredictor" des modèles pour la prédiction des propriétés physico-chimiques de composés purs et des mélanges ont été ajoutés:

- Corps purs:
 - Solubilité (logS) – modèles MLR
 - Point d'ébullition (Tbp) – modèles MLR,SVM ou ASNN
- Mélanges binaires:
 - Bubble point curve (Tbp vs. X%) – modèles ASNN et SVM
 - VLE curve (Y% vs. X%)– modèles ASNN et SVM
 - Classification (Azéotrope=1 / Zéotrope=0) – modèles SVM

9.2.2 Fonctionnement

Les molécules ou mélanges à prédire sont stockées dans un ou deux fichiers SDF respectivement et elles doivent être standardisées préalablement:

- logS: aromatisation, sans hydrogènes, acides forts déprotonés, bases fortes protonés, nitro(O–N+=O).
- Tbp: aromatisation, avec hydrogènes, nitro(O–N+=O).
- Propriétés mélanges binaires: aromatisation, avec hydrogènes, nitro(O–N+=O), tous les acides protonés.

L'utilisation du logiciel en interface graphique est intuitive et n'a pas besoin de plus d'explication. Ci-joint une capture d'écran de l'application.

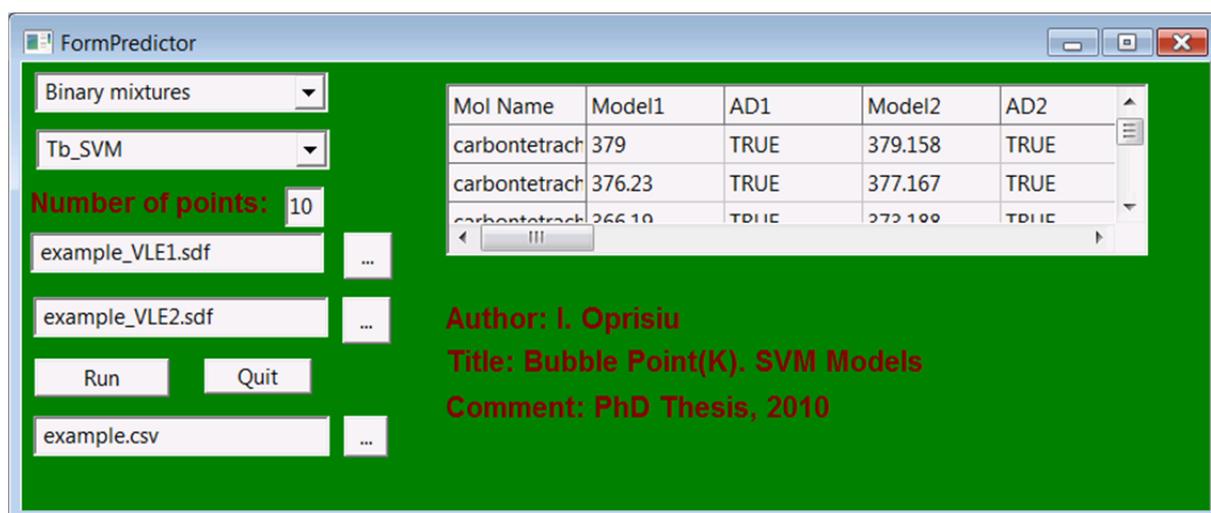


Figure 9-1 Capture d'écran de l'application graphique MixturesPredictor

Le logiciel en ligne de commande peut être utilisé en spécifiant différentes options :

Predictor_cmd.exe -m model(xml) -p nombre des points -i fichier sdf -c fichier csv des résultats

-m : indique le modèle à utiliser. Les modèles sont stockés dans un fichier *xml*. Sa structure est expliquée par la suite (cf. 9.2.3).

-p : est à utiliser pour les prédictions de Y% vs. X% ou de Tbp vs.X%. Pour ces cas il faut indiquer le nombre des points à prédire pour un mélange. Les points sont distribués uniformément entre 0 et 1.

-i : indique le fichier SDF à prédire. Pour les mélanges binaires il faut indiquer deux fichiers SDF de la façon suivante: *-i fichier1 SDF -i fichier2 SDF*

-c : indique le nom de fichier csv où les résultats vont être sauvegardés.

9.2.3 Fichiers XML

Le fichier XML (eXtensible Markup Language) est un élément important du "MixturesPredictor" apportant l'information nécessaire pour le fonctionnement du logiciel. Deux types de fichiers ont été utilisés: le fichier *cfg.xml* qui contient la liste des modèles disponibles et un fichier *.xml* attaché à chaque modèle de prédiction. Toutes les informations sur le modèle se retrouvent dans ce fichier: le type de machine d'apprentissage, le type de fragmentation, le nom de modèle, le domaine d'applicabilité utilisé, etc.

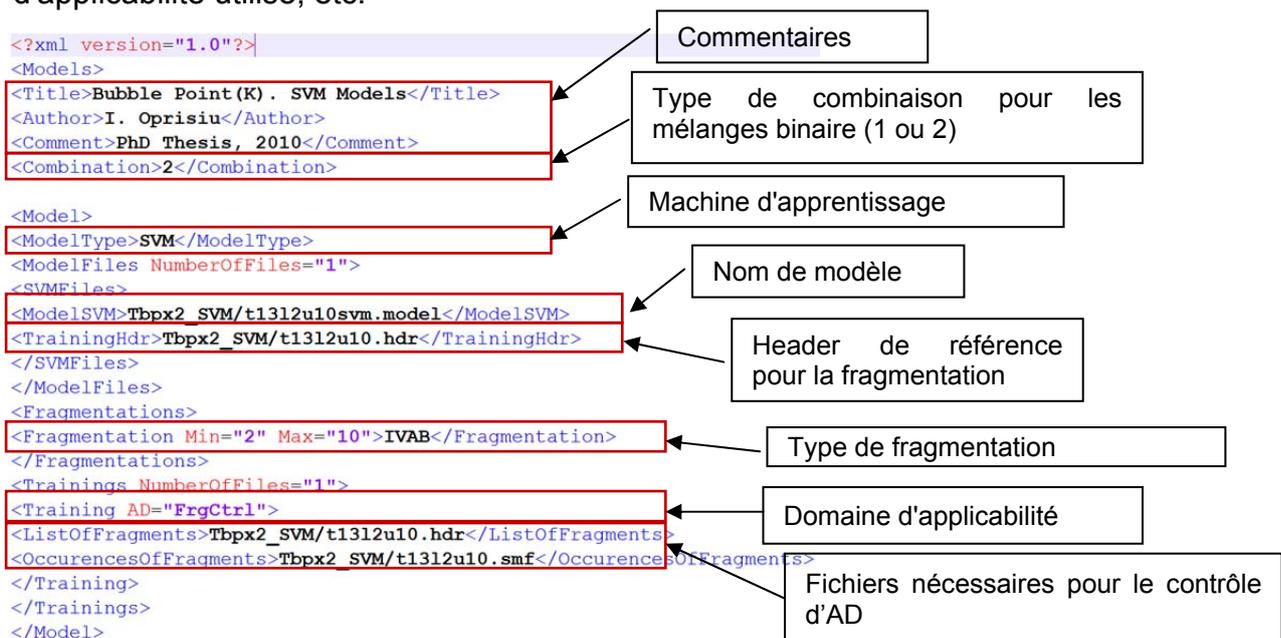


Figure 9-2 Fichier XML utilisé par le MixturesPredictor

9.2.4 Résultats

- Les résultats sont sauvegardés dans un fichier .csv. Il y a plusieurs colonnes (Models1, Model2...) qui représentent les prédictions faites pour chaque modèle de l'ensemble.

- Les colonnes correspondantes (Model1 AD, Model2 AD...) indique si la molécule/le mélange est dans le domaine d'applicabilité du modèle respectif (TRUE) ou ne l'est pas (FALSE).

- La colonne "Average" est la moyenne des toutes les prédictions de tous les modèles. La colonne "97.5% confidence" indique l'intervalle de confiance de la moyenne à 97.5%.

- La colonne "Average on AD" représente la moyenne des prédictions des modèles dont le DA contient la molécule/le mélange à prédire. La colonne "97.5% confidence on AD" indique l'intervalle de confiance de cette moyenne.

- La colonne "Applied models" indique le nombre de modèles inclus dans la moyenne tenant compte du DA. Ce nombre peut être inférieur au nombre total des modèles dont le DA est TRUE, car les valeurs extrêmes des prédictions sont éliminées.

- une dernière colonne est ajoutée pour la prédiction de VLE. Celle-ci représente les valeurs de X% pour laquelle la prédiction d'Y ou Tbp a été faite.

Compound Id	Compound(s) Name	Model1 AD	Model1 ...	Model9 AD	Model9	Average	97.5% confidence	Average on AD	97.5% confidence on AD	# Applied models	X(%)
1	carbontetrachloride isobutanol	TRUE	379.00 ...	TRUE	379.00	379.06	0.05	379.06	0.05	9	0.00
2	carbontetrachloride isobutanol	TRUE	376.23 ...	TRUE	369.03	373.80	2.15	373.80	2.15	9	0.11
3	carbontetrachloride isobutanol	TRUE	366.19 ...	TRUE	354.69	364.24	4.19	364.24	4.19	9	0.22
4	carbontetrachloride isobutanol	TRUE	354.94 ...	TRUE	341.87	354.99	6.05	354.99	6.05	9	0.33
5	carbontetrachloride isobutanol	TRUE	347.30 ...	TRUE	333.17	348.24	7.35	348.24	7.35	9	0.44
6	carbontetrachloride isobutanol	TRUE	343.47 ...	TRUE	330.64	343.99	6.71	343.99	6.71	9	0.56
7	carbontetrachloride isobutanol	TRUE	341.84 ...	TRUE	333.91	342.28	4.20	342.28	4.20	9	0.67
8	carbontetrachloride isobutanol	TRUE	341.51 ...	TRUE	339.93	342.59	1.52	342.59	1.52	9	0.78
9	carbontetrachloride isobutanol	TRUE	343.81 ...	TRUE	344.81	344.60	0.70	344.60	0.70	9	0.89
10	carbontetrachloride isobutanol	TRUE	349.72 ...	TRUE	350.05	349.41	0.42	349.41	0.42	9	1.00
11	tert-butanol n-butanol	TRUE	388.17 ...	FALSE	388.43	388.24	0.08	388.24	0.08	8	0.00
12	tert-butanol n-butanol	TRUE	373.07 ...	FALSE	377.20	375.22	1.35	375.22	1.35	8	0.11
13	tert-butanol n-butanol	TRUE	361.67 ...	FALSE	361.77	361.94	0.60	361.94	0.60	5	0.22
14	tert-butanol n-butanol	TRUE	356.53 ...	FALSE	345.80	357.48	1.51	357.48	1.51	6	0.33
15	tert-butanol n-butanol	TRUE	355.81 ...	FALSE	342.87	351.60	3.59	351.60	3.59	8	0.44
16	tert-butanol n-butanol	TRUE	352.20 ...	FALSE	344.95	349.66	2.45	349.66	2.45	8	0.56
17	tert-butanol n-butanol	TRUE	349.99 ...	FALSE	345.71	348.24	1.97	348.24	1.97	8	0.67
18	tert-butanol n-butanol	TRUE	349.89 ...	FALSE	347.39	348.85	1.39	348.85	1.39	8	0.78
19	tert-butanol n-butanol	TRUE	351.95 ...	FALSE	351.15	351.56	0.49	351.56	0.49	8	0.89
20	tert-butanol n-butanol	TRUE	356.45 ...	FALSE	356.02	356.22	0.17	356.22	0.17	8	1.00

Figure 9-3 Exemple de fichier des résultats .csv

En utilisant, ces résultats on peut tracer la courbe Tbp vs. X% pour le premier mélange, par exemple.

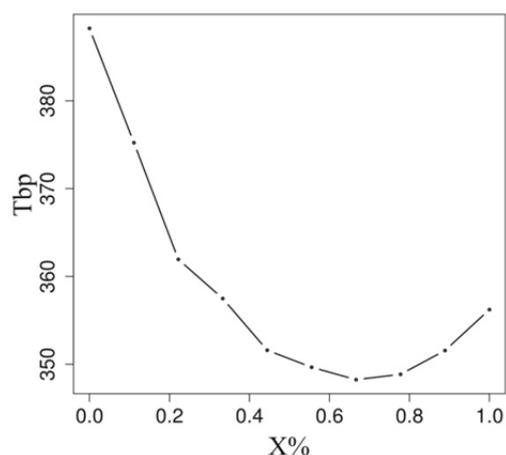


Figure 9-4 Exemple de courbe Tbp vs. X%

9.2.5 Interface WEB

La société PROCESSIUM a accepté de rendre disponibles publiquement les prédictions des modèles pour les corps purs et par conséquent les modèles de logS et Tbp ont été ajoutés à l'interface WEB de "Predictor" déjà développée pour d'autres propriétés des corps purs (constante de complexation des métaux et propriétés ADME).

Ci-joint une capture d'écran de cette interface WEB accessible à l'adresse

<http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi>

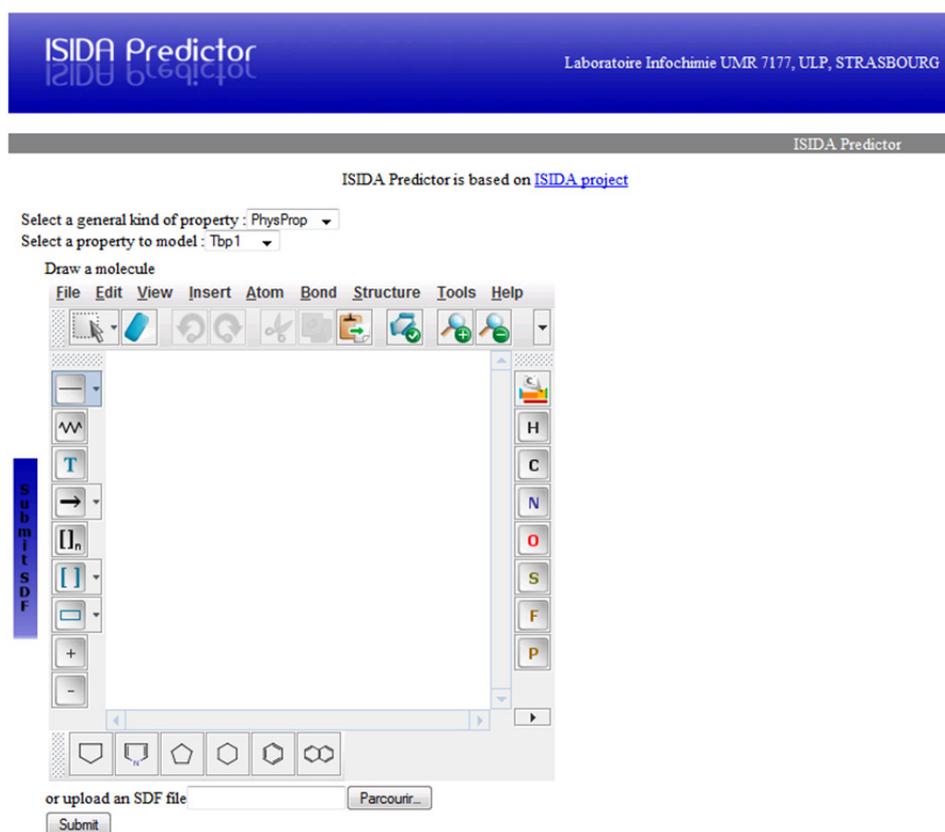


Figure 9-5 Capture d'écran de l'interface WEB de Predictor

L'application WEB ne nécessite aucune installation et elle peut être utilisée sur n'importe quel navigateur Internet. Un fichier SDF peut être soumis ou une molécule peut être dessinée à l'aide de l'applet Java de Chemaxon.

9.3 Conclusion

Les modèles développés font partie d'une plateforme de criblage virtuel au laboratoire d'Infochimie et qui est utilisée au sein de la société Processium. Contrairement à d'autres applications on-line, notre outil permet le traitement simultané de toutes les molécules/mélanges d'une base de données, au lieu d'un traitement cas par cas.

Toutes ces applications représentent la dernière étape de ce travail de la thèse, un travail complet, partant d'un concept et mené jusqu'au stade du développement d'applications logicielles pour des utilisateurs non spécialistes.

CONCLUSION GENERALE

Au cours de cette thèse nous avons proposé une nouvelle approche de modélisation de mélanges binaires non-additifs. Ce travail inclut le développement de descripteurs spéciaux pour des mélanges, un protocole d'obtention et de validation des modèles, ainsi qu'un domaine d'applicabilité robuste. Tous ces points ont été regroupés dans plusieurs publications dont la premières deux ont été publiées et une autre a été soumise. Deux autres publications sont en préparation.

Les descripteurs de mélanges, développés en combinant les descripteurs de ses composants purs, sont suffisants pour le développement de modèles QSPR pour les propriétés des mélanges, ce qui a été démontré dans la deuxième partie de ce travail.

Ce travail marque également un effort particulier sur les modèles de validation et sur la notion de domaine d'applicabilité. Le domaine d'applicabilité n'est pas évoqué dans les publications étudiées. L'importance du DA a été démontrée sur les différents modèles proposés et il semble indispensable en mélange car toute extrapolation non maîtrisée conduit rapidement à des erreurs de prédiction.

Afin de caractériser des mélanges d'équilibres liquide-vapeur, deux types d'approches ont été développés:

- La première approche concerne la détermination de l'existence d'un azéotrope. La précision obtenue est de 84% lorsque les composés en jeu sont présents dans le set initial d'ajustement ou sont des composés proches structurellement. Ceci justifie l'importance du domaine d'applicabilité mis en place.
- Dans une deuxième approche, la courbe d'équilibre $y=f(x)$ est recherchée et les modèles obtenus montrent une bonne capacité de classification des mélanges. Les modèles pour la prédiction de la température de bulle étant moins lisses n'ont pas donné des résultats satisfaisants pour la classification.

Pour obtenir le détail des azéotropes (température, composition) trois type de modèles ont été développés:

- Des modèles de régression du point azéotropique montrent des résultats avec des écarts cohérents avec les erreurs expérimentales. La température d'ébullition est prédite avec une erreur inférieure à 4K, tandis que la composition est prédite avec une erreur de 14.4% en validation externe.

- Les modèles pour prédire la courbe d'équilibre $y=f(x)$ ne peuvent prédire que la composition en phase gazeuse en fonction de la composition en phase liquide, et celle-ci est estimée avec une erreur de 4% en validation externe « Mixtures Out ».
- Les modèles pour prédire la température d'ébullition en fonction de la composition peuvent donner la température d'ébullition en fonction de la composition en phase liquide avec une erreur inférieure à 6K en validation externe « Mixtures Out ».

Ainsi, ce travail apporte plusieurs pistes aux ingénieurs pour répondre à leurs problématiques. (Schéma 2-1)

La comparaison avec d'autres modèles, COSMO ou UNIFAC, a montré que nos modèles présentent des résultats similaires ou meilleurs. Ceci montre que les modèles QSPR sont une alternative fiable aux modèles COSMO et UNIFAC.

Tout au long de ce travail, nous avons attaché beaucoup d'importance au développement d'une démarche rigoureuse et aux attentes des utilisateurs de tels modèles. Ainsi des logiciels ont été développés pour exploiter les modèles. Il faut retenir en particulier MixturesPredictor, qui fait partie d'une plateforme de criblage virtuel au laboratoire d'Infochimie et qui est utilisé au sein de la société Processium.

Ce travail nous ouvre de nombreuses perspectives, dont il faut mentionner:

- Le développement des DA pour des mélanges. Aujourd'hui les DA pour les corps purs ne sont pas adaptés pour les mélanges, ce que nous avons pu voir lors des développements des modèles QSPR. Ceci a été présenté plus en détail dans le chapitre 5.3.4, mentionnant le seul DA efficace, le « Fragment Control IVAB ».
- La difficulté de disposer des données pour des propriétés des mélanges (cf. 2.3) se répercute directement sur la qualité des modèles QSPR. Nous pouvons alors espérer qu'en élargissant nos bases de données (plus de données pour des mélanges contenant des nouveaux corps purs), les modèles pourront être reproduits et leurs performances augmenteront. De plus, leur domaine d'applicabilité sera élargi.
- Une autre perspective qui ne doit pas être écartée, est le développement des modèles QSPR tout en gardant en vue l'aspect thermodynamique, c'est-à-dire développer des modèles QSPR basés sur des équations de thermodynamique décrivant le comportement des mélanges (voir. 2.2.4).

- Etant donné que les descripteurs développés s'appliquent aussi aux mélanges contenant plus de deux composants, des modèles QSPR pourront être développés pour prédire des propriétés de ces mélanges.
- La gamme de problèmes auquel cette méthodologie peut répondre est beaucoup plus large. Nous pouvons envisager de développer des modèles QSPR pour prédire d'autres propriétés non-additives des mélanges dépendant de la composition, comme la densité, la viscosité, la toxicité, etc.

10 Communications

10.1 Publications

Solov'ev, V.P., I. Oprisiu, G. Marcou, A. Varnek., *Quantitative Structure-Property Relationship (QSPR) Modeling of Normal Boiling Point Temperature and Composition of Binary Azeotropes*. Industrial & Engineering Chemistry Research, 2011. 50(24): p. 14162-14167

I. Oprisiu, G. Marcou, D. Horvath, D. Bernard Brunel, F. Rivollet, A. Varnek, "*Publically available models to predict normal boiling point of organic compounds.*", (soumis au Thermodynamica Acta).

I. Oprisiu, E. Varlamova, E. Muratov, A. Artemenko, G. Marcou, P. Polischuk, V. Kuz'min, A. Varnek, "*QSPR approach to predict non-additive properties of multicomponents mixtures. Application to Bubble point temperatures of binary liquid mixtures.*", (publié dans Molecular Informatics).

DOI: 10.1002/minf.201200006

I. Oprisiu, G. Marcou, F. Rivollet, A. Varnek, "*Classification model azeotrope/zeotrope involving fragment descriptors.*", (en préparation)

I. Oprisiu, G. Marcou, F. Rivollet, A. Varnek, "*QSPR modeling of vapor-liquid equilibrium of binary liquid mixtures*", (en préparation)

10.2 Communications orales

I. Oprisiu, "QSPR approach to predict non-additive properties of multicomponents mixtures. Application to the azeotropic behavior of binary liquid mixtures", Journée des doctorants en chimie, Strasbourg, octobre 2010

I. Oprisiu "QSPR approach to predict non-additive properties of multicomponents mixtures. Application to the azeotropic behavior of binary liquid mixtures", Cabourg, octobre 2011

10.3 Communications par affiche

I. Oprisiu, G. Marcou, D. Horvath et A. Varnek, "*Predictive models for aqueous solubility based on the ISIDA descriptors*", Journées Nationales de la Chémoinformatique, Montpellier, juin 2009

I. Oprisiu, E. Muratov, E. Varlamova, V. Kuz'min, G. Marcou, D. Horvath et A. Varnek "*Predictive QSPR models for Bubble Point curve of binary liquid mixtures*", Summer School on Chemoinformatics, Obernai, juin 2010.

I. Oprisiu, G. Marcou, F. Rivollet, D. Horvath et A. Varnek, "*Predictive QSPR models for the azeotropic behavior of binary liquid mixtures*", 18th Euro-QSAR, Rhodes, septembre 2010.

E. Varlamova, I. Oprisiu, E. Muratov, A. Artemenko, G. Marcou, P. Polischuk, V. Kuz'min et A. Varnek, "QSPR Analysis of Boiling Temperatures of 2-Component Systems", MACC4, Lviv, juin 2011

11 Annexes

11.1 Descripteurs moléculaires calculés avec MOE

Code	Description
apol	Sum of the atomic polarizabilities (including implicit hydrogens) with polarizabilities taken from [CRC 1994].
bpol	Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens) with polarizabilities taken from [CRC 1994].
density	Molecular mass density: Weight divided by vdw_vol (amu/Å ³).
FCharge	Total charge of the molecule (sum of formal charges).
mr	Molecular refractivity (including implicit hydrogens). This property is calculated from an 11 descriptor linear model [MREF 1998] with $r^2 = 0.997$, RMSE = 0.168 on 1,947 small molecules.
SMR	Molecular refractivity (including implicit hydrogens). This property is an atomic contribution model [Crippen 1999] that assumes the correct protonation state (washed structures). The model was trained on ~7000 structures and results may vary from the mr descriptor.
Weight	Molecular weight (including implicit hydrogens) in atomic mass units with atomic weights taken from [CRC 1994].
logP(o/w)	Log of the octanol/water partition coefficient (including implicit hydrogens). This property is calculated from a linear atom type model [LOGP 1998] with $r^2 = 0.931$, RMSE=0.393 on 1,827 molecules.
logS	Log of the aqueous solubility (mol/L). This property is calculated from an atom contribution linear atom type model [Hou 2004] with $r^2 = 0.90$, ~1,200 molecules.
reactive	Indicator of the presence of reactive groups. A non-zero value indicates that the molecule contains a reactive group. The table of reactive groups is based on the Oprea set [Oprea 2000] and includes metals, phospho-, N/O/S-N/O/S

	single bonds, thiols, acyl halides, Michael Acceptors, azides, esters, etc.
TPSA	Polar surface area (\AA^2) calculated using group contributions to approximate the polar surface area from connection table information only. The parameterization is that of Ertl <i>et al.</i> [Ertl 2000].
vdw_vol	van der Waals volume (\AA^3) calculated using a connection table approximation.
vdw_area	Area of van der Waals surface (\AA^2) calculated using a connection table approximation.
a_aro	Number of aromatic atoms.
a_count	Number of atoms (including implicit hydrogens). This is calculated as the sum of $(1 + hi)$ over all non-trivial atoms i .
a_heavy	Number of heavy atoms $\#\{Z_i \mid Z_i > 1\}$.
a_ICM	Atom information content (mean). This is the entropy of the element distribution in the molecule (including implicit hydrogens but not lone pair pseudo-atoms). Let ni be the number of occurrences of atomic number i in the molecule. Let $pi = ni / n$ where n is the sum of the ni . The value of a_ICM is the negative of the sum over all i of $pi \log pi$.
a_IC	Atom information content (total). This is calculated to be a_ICM times n .
b_1rotN	Number of rotatable single bonds. Conjugated single bonds are not included (e.g., ester and peptide bonds).
b_1rotR	Fraction of rotatable single bonds: b_1rotN divided by b_heavy.
b_ar	Number of aromatic bonds.
b_count	Number of bonds (including implicit hydrogens). This is calculated as the sum of $(di/2 + hi)$ over all non-trivial atoms i .
b_double	Number of double bonds. Aromatic bonds are not considered to be double bonds.
b_heavy	Number of bonds between heavy atoms.
b_rotN	Number of rotatable bonds. A bond is rotatable if it has order 1, is not in a

	ring, and has at least two heavy neighbors.
b_rotR	Fraction of rotatable bonds: b_rotN divided by b_heavy.
b_single	Number of single bonds (including implicit hydrogens). Aromatic bonds are not considered to be single bonds.
b_triple	Number of triple bonds. Aromatic bonds are not considered to be triple bonds.
rings	The number of rings.
chiral	The number of chiral centers.
chiral_u	The number of unconstrained chiral centers
VAdjMa	Vertex adjacency information (magnitude): $1 + \log_2 m$ where m is the number of heavy-heavy bonds. If m is zero, then zero is returned.
VAdjEq	Vertex adjacency information (equality): $-(1-f)\log_2(1-f) - f\log_2 f$ where $f = (n_2 - m) / n_2$, n is the number of heavy atoms and m is the number of heavy-heavy bonds. If f is not in the open interval (0,1), then 0 is returned.
chi0	Atomic connectivity index (order 0) from [Hall 1991] and [Hall 1977]. This is calculated as the sum of $1/\sqrt{d_i}$ over all heavy atoms i with $d_i > 0$.
chi0_C	Carbon connectivity index (order 0). This is calculated as the sum of $1/\sqrt{d_i}$ over all carbon atoms i with $d_i > 0$.
chi1	Atomic connectivity index (order 1) from [Hall 1991] and [Hall 1977]. This is calculated as the sum of $1/\sqrt{d_i d_j}$ over all bonds between heavy atoms i and j where $i < j$.
chi1_C	Carbon connectivity index (order 1). This is calculated as the sum of $1/\sqrt{d_i d_j}$ over all bonds between carbon atoms i and j where $i < j$.
chi0v	Atomic valence connectivity index (order 0) from [Hall 1991] and [Hall 1977]. This is calculated as the sum of $1/\sqrt{v_i}$ over all heavy atoms i with $v_i > 0$.
chi0v_C	Carbon valence connectivity index (order 0). This is calculated as the sum of $1/\sqrt{v_i}$ over all carbon atoms i with $v_i > 0$.
chi1v	Atomic valence connectivity index (order 1) from [Hall 1991] and [Hall 1977].

	This is calculated as the sum of $1/\sqrt{v_i v_j}$ over all bonds between heavy atoms i and j where $i < j$.
chi1v_C	Carbon valence connectivity index (order 1). This is calculated as the sum of $1/\sqrt{v_i v_j}$ over all bonds between carbon atoms i and j where $i < j$.
Kier1	First kappa shape index: $(n-1)^2 / m^2$ [Hall 1991].
Kier2	Second kappa shape index: $(n-1)^2 / m^2$ [Hall 1991].
Kier3	Third kappa shape index: $(n-1)(n-3)^2 / p^3$ for odd n , and $(n-3)(n-2)^2 / p^3$ for even n [Hall 1991].
KierA1	First alpha modified shape index: $s(s-1)^2 / m^2$ where $s = n + a$ [Hall 1991].
KierA2	Second alpha modified shape index: $s(s-1)^2 / m^2$ where $s = n + a$ [Hall 1991].
KierA3	Third alpha modified shape index: $(n-1)(n-3)^2 / p^3$ for odd n , and $(n-3)(n-2)^2 / p^3$ for even n where $s = n + a$ [Hall 1991].
KierFlex	Kier molecular flexibility index: $(KierA1)(KierA2) / n$ [Hall 1991].
zagreb	Zagreb index: the sum of d_i^2 over all heavy atoms i .
radius	If r_i is the largest matrix entry in row i of the distance matrix D , then the radius is defined as the smallest of the r_i [Petitjean 1992].
VDistEq	If m is the sum of the distance matrix entries then VdistEq is defined to be the sum of $\log_2 m - p_i \log_2 p_i / m$ where p_i is the number of distance matrix entries equal to i .
diameter	Largest value in the distance matrix [Petitjean 1992].
VDistMa	If m is the sum of the distance matrix entries then VDistMa is defined to be the sum of $\log_2 m - D_{ij} \log_2 D_{ij} / m$ over all i and j .
a_acc	Number of hydrogen bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
a_acid	Number of acidic atoms.

a_base	Number of basic atoms.
a_don	Number of hydrogen bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
a_hyd	Number of hydrophobic atoms.
Q_PC+ PEOE_P C+	Total positive partial charge: the sum of the positive q_i . Q_PC+ is identical to PC+ which has been retained for compatibility.
Q_PC- PEOE_P C-	Total negative partial charge: the sum of the negative q_i . Q_PC- is identical to PC- which has been retained for compatibility.

11.2 Prédiction du point azéotrope. Matériel supplémentaire

SUPPORTING INFORMATION

SM1. Experimental and predicted values of normal boiling point temperature ($T_{bp\ az}$) and the composition (X_{Iw}) of 176 two-component liquid azeotropes ^a.

no.	component 1		component 2		$T_{bp\ az}$			X_{Iw}	
	name	$T_{bp, C1}$	name	$T_{bp, C2}$	exp.	pred.1	pred.2 ^b	exp.	pred.
1	methyl ethyl ketone	352.75	benzene	353.25	351.15	347.4	347.00	38	58
2	2-butanol	372.65	H ₂ O	373.15	360.15	361.7	366.57	73	69
3	CCl ₄	349.65	ethyl acetate	350.25	348.15	342.9	343.95	57	89
4	benzene	353.25	cyclohexane	353.85	351.15	340.0	347.49	55	64
5	acetone	329.35	methyl acetate	330.15	328.15	329.5	324.00	48	70
6	nitromethane	374.15	1,4-dioxane	375.15	373.15	373.8	368.08	57	58
7	H ₂ O	373.15	formic acid	374.15	373.15	355.0	367.10	26	18
8	H ₂ O	373.15	nitromethane	374.15	357.15	360.1	367.10	24	18
9	2-propanol	355.55	1,2-dichloroethane	356.65	346.15	355.9	349.80	39	46
10	ethyl acetate	350.25	ethanol	351.45	345.15	345.5	344.59	74	82
11	cyclohexane	353.85	t-butanol	355.15	345.15	345.5	348.14	63	70
12	ethanol	351.45	methyl ethyl ketone	352.75	348.15	342.7	345.78	34	43
13	pyridine	388.75	1-butanol	390.15	392.15	370.2	382.47	30	77
14	benzene	353.25	acetonitrile	354.75	346.15	347.9	347.56	66	81
15	heptane	371.55	H ₂ O	373.15	352.15	358.3	365.57	87	84
16	ethylic ether	307.65	pentane	309.25	306.15	307.2	302.72	68	85
17	cyclohexane	353.85	2-propanol	355.55	342.15	345.6	348.17	68	73
18	ethanol	351.45	benzene	353.25	341.15	346.8	345.82	32	44
19	CCl ₄	349.65	ethanol	351.45	338.15	340.2	344.05	84	87
20	benzene	353.25	t-butanol	355.15	347.15	348.8	347.60	63	76
21	H ₂ O	373.15	1,4-dioxane	375.15	361.15	367.8	367.18	18	25
22	benzene	353.25	2-propanol	355.55	345.15	348.2	347.63	66	80
23	pyridine	388.75	acetic acid	391.15	411.15	371.6	382.55	49	73
24	ethanol	351.45	cyclohexane	353.85	338.15	339.4	345.86	31	41
25	i-butanol	381.15	toluene	383.75	374.15	380.9	375.09	45	55
26	1-propanol	370.55	H ₂ O	373.15	361.15	362.3	364.67	72	63
27	chloroform	334.85	methanol	337.65	326.15	329.9	329.57	87	89
28	cyclohexane	353.85	1,2-dichloroethane	356.65	348.15	347.1	348.26	50	58
29	methyl ethyl ketone	352.75	2-propanol	355.55	351.15	349.5	347.17	68	75
30	THF	339.15	hexane	342.15	336.15	320.8	333.81	54	60
31	acetic acid	391.15	tetrachloroethylene	394.15	380.15	363.3	384.96	39	55
32	CCl ₄	349.65	methyl ethyl ketone	352.75	347.15	343.2	344.15	71	72
33	ethanol	351.45	acetonitrile	354.75	346.15	344.1	345.93	56	41
34	benzene	353.25	1,2-dichloroethane	356.65	353.15	350.1	347.71	85	61
35	1,2-dichloroethane	356.65	trichloroethylene	360.15	355.15	350.7	351.06	61	61
36	1-propanol	370.55	nitromethane	374.15	363.15	362.0	364.74	52	48
37	ethyl acetate	350.25	cyclohexane	353.85	345.15	343.8	344.78	56	65
38	ethyl acetate	350.25	acetonitrile	354.75	348.15	347.3	344.85	77	58

39	methanol	337.65	hexane	342.15	323.15	330.3	332.45	26	44
40	1-propanol	370.55	1,4-dioxane	375.15	368.15	367.3	364.82	55	46
41	2-propanol	355.55	trichloroethylene	360.15	349.15	354.0	350.07	30	45
42	methyl acetate	330.15	chloroform	334.85	338.15	334.1	325.09	23	43
43	toluene	383.75	pyridine	388.75	383.15	374.5	377.83	78	75
44	CCl ₄	349.65	acetonitrile	354.75	338.15	341.9	344.30	83	84
45	ethanol	351.45	1,2-dichloroethane	356.65	344.15	345.6	346.08	27	41
46	ethyl acetate	350.25	2-propanol	355.55	349.15	346.5	344.91	75	79
47	pyridine	388.75	tetrachloroethylene	394.15	386.15	400.1	382.78	49	20
48	acetone	329.35	chloroform	334.85	338.15	330.8	324.37	21	28
49	CCl ₄	349.65	t-butanol	355.15	344.15	343.5	344.33	83	75
50	ethylic ether	307.65	dichloromethane	313.15	314.15	313.5	303.02	30	8
51	CCl ₄	349.65	2-propanol	355.55	342.15	342.2	344.36	82	84
52	dichloromethane	313.15	carbon bisulfide	319.35	309.15	319.5	308.49	65	92
53	toluene	383.75	1-butanol	390.15	379.15	374.9	377.94	72	83
54	cyclopentane	322.45	acetone	329.35	314.15	345.3	317.69	64	59
55	CCl ₄	349.65	1,2-dichloroethane	356.65	348.15	342.4	344.45	83	66
56	chloroform	334.85	hexane	342.15	333.15	325.6	329.92	72	63
57	toluene	383.75	acetic acid	391.15	374.15	381.2	378.02	72	78
58	methyl acetate	330.15	methanol	337.65	327.15	329.4	325.31	82	89
59	H ₂ O	373.15	i-butanol	381.15	363.15	364.4	367.64	33	33
60	acetone	329.35	methanol	337.65	329.15	325.7	324.58	88	82
61	carbon bisulfide	319.35	ethyl formate	327.65	312.15	317.1	314.75	63	60
62	ethanol	351.45	trichloroethylene	360.15	344.15	349.0	346.35	27	35
63	hexane	342.15	ethanol	351.45	331.15	334.4	337.25	79	84
64	formic acid	374.15	toluene	383.75	359.15	356.4	368.75	50	38
65	nitromethane	374.15	toluene	383.75	370.15	369.0	368.75	55	63
66	carbon bisulfide	319.35	acetone	329.35	312.15	317.0	314.88	67	82
67	ethyl formate	327.65	methanol	337.65	324.15	327.2	323.04	84	88
68	pentane	309.25	carbon bisulfide	319.35	309.15	307.9	304.95	89	81
69	trichloroethylene	360.15	1-propanol	370.55	355.15	354.4	355.04	83	88
70	hexane	342.15	methyl ethyl ketone	352.75	337.15	337.3	337.35	70	68
71	H ₂ O	373.15	toluene	383.75	358.15	365.4	367.84	20	26
72	carbon bisulfide	319.35	methyl acetate	330.15	313.15	315.3	314.94	70	84
73	2-butanol	372.65	toluene	383.75	368.15	366.3	367.39	55	62
74	carbon bisulfide	319.35	1,1-dichloroethane	330.45	319.15	317.0	314.96	94	84
75	ethylic ether	307.65	carbon bisulfide	319.35	307.15	307.4	303.50	99	71
76	methyl acetate	330.15	hexane	342.15	325.15	326.4	325.66	61	71
77	methanol	337.65	CCl ₄	349.65	329.15	331.6	333.03	21	59
78	methanol	337.65	ethyl acetate	350.25	335.15	331.1	333.08	49	52
79	acetone	329.35	hexane	342.15	323.15	330.8	324.93	59	69
80	hexane	342.15	t-butanol	355.15	337.15	336.5	337.54	78	75
81	trichloroethylene	360.15	H ₂ O	373.15	346.15	353.3	355.24	95	98
82	1-propanol	370.55	toluene	383.75	366.15	363.3	365.49	49	53
83	hexane	342.15	2-propanol	355.55	336.15	335.3	337.57	77	81
84	acetaldehyde	293.95	ethylic ether	307.65	293.15	318.5	290.18	76	58
85	H ₂ O	373.15	1,1,2-trichloroethane	386.95	359.15	363.8	368.09	16	10
86	trichloroethylene	360.15	nitromethane	374.15	354.15	352.9	355.32	80	86
87	methanol	337.65	benzene	353.25	331.15	331.9	333.31	39	24

88	H ₂ O	373.15	pyridine	388.75	367.15	364.3	368.23	42	20
89	nitromethane	374.15	1-butanol	390.15	371.15	375.9	369.24	71	66
90	methanol	337.65	cyclohexane	353.85	318.15	331.6	333.36	37	35
91	chloroform	334.85	ethanol	351.45	332.15	326.9	330.64	93	83
92	cyclohexane	353.85	1-propanol	370.55	348.15	348.6	349.33	81	82
93	H ₂ O	373.15	1-butanol	390.15	366.15	363.5	368.34	43	24
94	methanol	337.65	acetonitrile	354.75	336.15	331.7	333.43	19	54
95	heptane	371.55	pyridine	388.75	369.15	369.0	366.78	75	86
96	benzene	353.25	1-propanol	370.55	350.15	351.1	348.79	83	87
97	2-propanol	355.55	H ₂ O	373.15	353.15	345.9	351.07	88	73
98	t-butanol	355.15	H ₂ O	373.15	353.15	323.8	350.71	88	90
99	carbon bisulfide	319.35	methanol	337.65	311.15	309.9	315.52	86	97
100	benzene	353.25	heptane	371.55	353.15	346.3	348.87	99	87
101	heptane	371.55	1-butanol	390.15	367.15	360.9	366.89	82	73
102	2-propanol	355.55	nitromethane	374.15	352.15	351.6	351.15	72	62
103	cyclohexane	353.85	2-butanol	372.65	349.15	346.5	349.49	82	79
104	methanol	337.65	1,2-dichloroethane	356.65	333.15	333.4	333.58	35	35
105	cyclohexane	353.85	H ₂ O	373.15	343.15	343.9	349.53	92	89
106	benzene	353.25	2-butanol	372.65	352.15	351.9	348.95	85	85
107	heptane	371.55	acetic acid	391.15	365.15	365.9	366.97	67	87
108	benzene	353.25	H ₂ O	373.15	342.15	349.9	348.99	91	98
109	pentane	309.25	acetone	329.35	306.15	310.7	305.73	80	73
110	ethanol	351.45	heptane	371.55	345.15	343.6	347.23	48	61
111	acetone	329.35	CCl ₄	349.65	329.15	330.7	325.51	88	45
112	cyclohexane	353.85	nitromethane	374.15	343.15	347.7	349.61	73	77
113	methyl ethyl ketone	352.75	H ₂ O	373.15	346.15	350.6	348.54	89	79
114	pentane	309.25	methyl acetate	330.15	307.15	310.6	305.79	88	69
115	benzene	353.25	nitromethane	374.15	352.15	346.5	349.07	87	83
116	CCl ₄	349.65	1-propanol	370.55	346.15	344.6	345.53	89	92
117	1,1-dichloroethane	330.45	ethanol	351.45	328.15	329.4	326.65	86	84
118	trichloroethylene	360.15	i-butanol	381.15	358.15	356.8	355.86	91	92
119	H ₂ O	373.15	tetrachloroethylene	394.15	361.15	360.8	368.65	16	30
120	cyclohexane	353.85	1,4-dioxane	375.15	353.15	347.8	349.69	75	68
121	2-butanol	372.65	tetrachloroethylene	394.15	370.15	362.5	368.19	57	60
122	ethanol	351.45	H ₂ O	373.15	351.15	339.8	347.36	96	69
123	iodomethane	315.55	methanol	337.65	311.15	312.9	312.08	96	90
124	methanol	337.65	trichloroethylene	360.15	342.15	331.9	333.85	38	25
125	ethanol	351.45	nitromethane	374.15	349.15	344.2	347.44	71	59
126	ethyl acetate	350.25	H ₂ O	373.15	344.15	346.2	346.27	92	97
127	CCl ₄	349.65	H ₂ O	373.15	339.15	340.1	345.73	96	98
128	1-propanol	370.55	tetrachloroethylene	394.15	367.15	367.8	366.29	48	50
129	methyl acetate	330.15	cyclohexane	353.85	328.15	324.2	326.56	83	58
130	acetone	329.35	cyclohexane	353.85	326.15	329.3	325.84	67	62
131	dichloromethane	313.15	methanol	337.65	311.15	309.3	309.90	93	82
132	1,2-dichloroethane	356.65	i-butanol	381.15	356.15	349.0	352.69	94	94
133	CCl ₄	349.65	formic acid	374.15	340.15	349.1	345.80	82	92
134	CCl ₄	349.65	nitromethane	374.15	344.15	342.8	345.80	83	87
135	H ₂ O	373.15	octane	398.85	363.15	362.4	369.01	26	35
136	cyclohexane	353.85	i-butanol	381.15	351.15	350.4	350.15	86	86

137	benzene	353.25	i-butanol	381.15	352.15	355.3	349.61	92	93
138	pentane	309.25	methanol	337.65	304.15	306.0	306.37	93	95
139	hexane	342.15	1-propanol	370.55	339.15	337.5	338.73	96	87
140	trichloroethylene	360.15	1-butanol	390.15	360.15	356.3	356.56	97	90
141	carbon bisulfide	319.35	ethyl acetate	350.25	319.15	319.1	316.50	97	72
142	hexane	342.15	H ₂ O	373.15	335.15	333.3	338.93	94	94
143	trichloroethylene	360.15	acetic acid	391.15	359.15	359.4	356.64	96	87
144	CCl ₄	349.65	i-butanol	381.15	349.15	348.3	346.35	95	91
145	hexane	342.15	nitromethane	374.15	335.15	335.2	339.01	79	83
146	carbon bisulfide	319.35	ethanol	351.45	315.15	314.6	316.59	91	89
147	ethanol	351.45	toluene	383.75	350.15	345.9	348.18	68	56
148	carbon bisulfide	319.35	methyl ethyl ketone	352.75	319.15	316.2	316.69	84	69
149	methanol	337.65	heptane	371.55	332.15	332.7	334.73	54	57
150	carbon bisulfide	319.35	t-butanol	355.15	318.15	317.1	316.88	93	81
151	iodomethane	315.55	ethanol	351.45	314.15	313.2	313.15	97	97
152	carbon bisulfide	319.35	2-propanol	355.55	317.15	316.6	316.91	92	87
153	cyclohexane	353.85	1-butanol	390.15	353.15	350.0	350.85	90	86
154	cyclohexane	353.85	acetic acid	391.15	353.15	352.6	350.93	98	84
155	benzene	353.25	acetic acid	391.15	353.15	355.6	350.38	98	89
156	chloroform	334.85	H ₂ O	373.15	329.15	328.0	332.32	97	98
157	2-propanol	355.55	tetrachloroethylene	394.15	355.15	356.4	352.70	81	54
158	iodomethane	315.55	2-propanol	355.55	315.15	318.7	313.46	98	76
159	methyl acetate	330.15	heptane	371.55	330.15	332.4	327.93	96	75
160	CCl ₄	349.65	acetic acid	391.15	349.15	350.5	347.12	98	91
161	pentane	309.25	ethanol	351.45	307.45	307.0	307.44	95	93
162	acetone	329.35	heptane	371.55	329.15	333.1	327.21	90	80
163	methyl acetate	330.15	H ₂ O	373.15	329.15	331.1	328.06	95	97
164	acetone	329.35	H ₂ O	373.15	329.15	329.2	327.33	88	100
165	methanol	337.65	toluene	383.75	337.15	334.6	335.67	71	53
166	hexane	342.15	1-butanol	390.15	341.15	339.1	340.25	97	94
167	H ₂ O	373.15	nonane	423.95	368.15	357.4	370.95	40	42 ^c
168	carbon bisulfide	319.35	1-propanol	370.55	319.15	317.9	318.07	95	92
169	carbon bisulfide	319.35	H ₂ O	373.15	317.15	312.9	318.27	98	100
170	carbon bisulfide	319.35	formic acid	374.15	316.15	321.7	318.35	83	88
171	carbon bisulfide	319.35	nitromethane	374.15	317.15	315.8	318.35	90	91
172	methanol	337.65	tetrachloroethylene	394.15	337.15	336.6	336.48	64	47
173	dichloromethane	313.15	H ₂ O	373.15	312.15	315.0	312.65	98	100
174	methanol	337.65	octane	398.85	336.15	335.2	336.84	68	64
175	pentane	309.25	H ₂ O	373.15	308.15	307.7	309.12	99	100
176	ethylic ether	307.65	H ₂ O	373.15	307.15	310.0	307.67	99	100

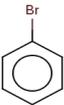
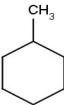
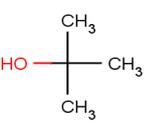
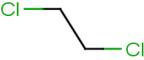
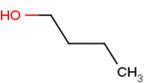
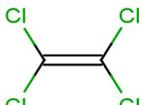
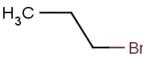
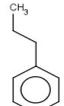
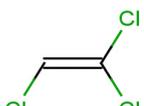
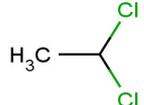
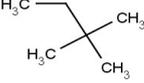
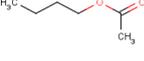
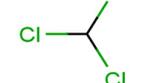
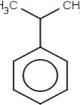
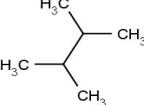
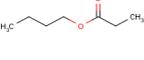
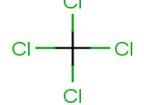
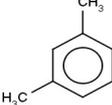
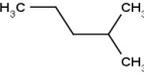
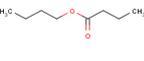
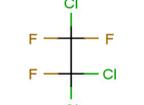
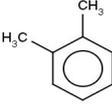
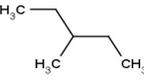
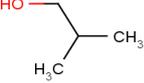
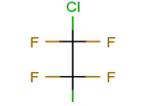
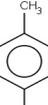
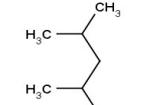
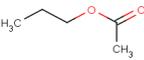
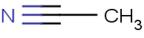
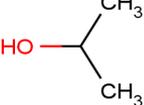
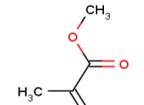
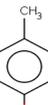
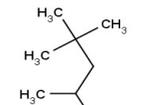
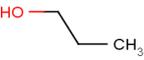
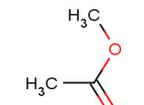
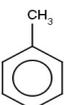
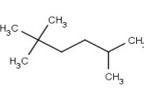
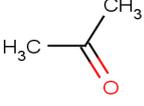
^{a)} Experimental data are collected from the book ²⁶, the values of normal boiling point temperatures of first ($T_{bp, C1}$) and second ($T_{bp, C2}$) components and the azeotropes ($T_{bp, az}$) are given in K; wt. % is the unit of weight fraction for first component (X_{Iw}). The values of $T_{bp, az, pred.1}$ and $X_{Iw, pred}$ are predicted by consensus models and represent combinations of all external test sets of 5-fold cross-validation procedure; ^{b)} $T_{bp, az, pred.2}$ is predicted according to eq. 1; ^{c)} outside of AD of CMs.

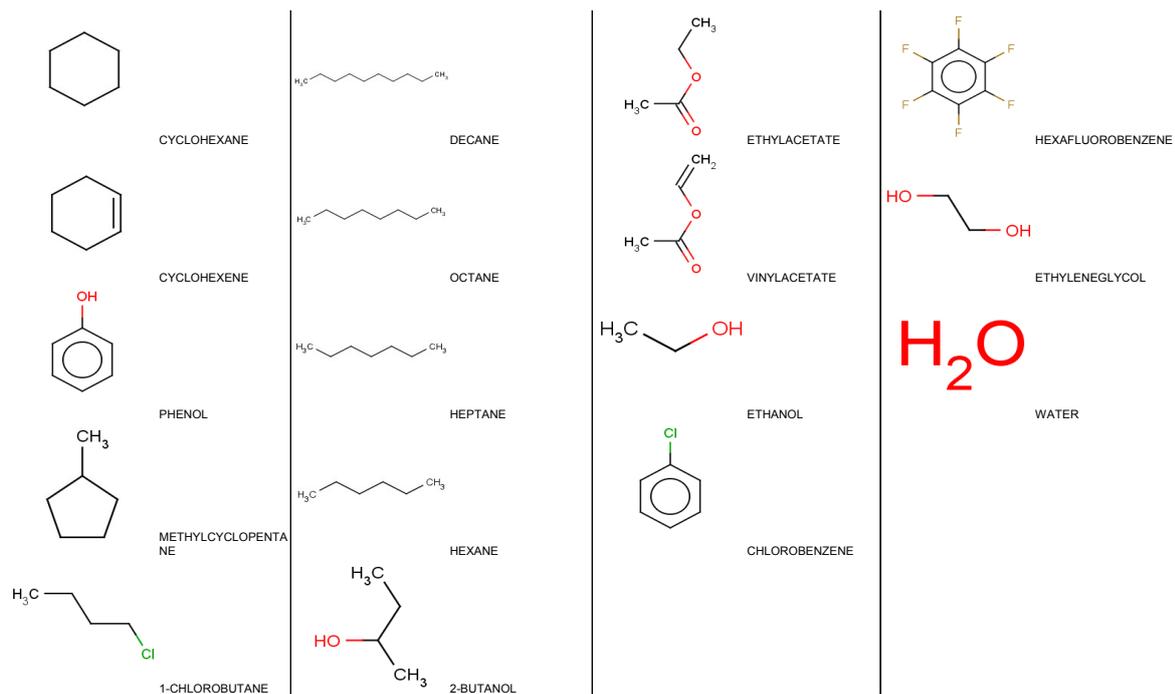
SM2. Experimental and predicted values of normal boiling point temperature ($T_{bp\ az}$) and composition (X_{Iw}) for extra test set of 24 two-component liquid azeotropes ^a.

no.	component 1 name	$T_{bp, C1}$	component 2 name	$T_{bp, C2}$	$T_{bp\ az}$		X_{Iw}		
					<i>exp.</i>	<i>pred.1</i>	<i>pred.2^b</i>	<i>exp.</i>	<i>pred.</i>
1	1,1,2-trichloro- 1,2,2-trifluoroethane	321.00	t-butanol	355.15	319.95	391.5 ^c	318.37	98	76
2	1,1,2-trichloro- 1,2,2-trifluoroethane	321.00	ethanol	351.45	317.75	388.9 ^c	318.09	96	87
3	1,1-dichloroethane	330.37	2-propanol	355.55	329.55	330.0	326.89	93	81
4	1,1-dichloroethane	330.37	hexane	342.15	329.30	323.6	325.86	82	66
5	propyl acetate	374.75	1,4-dioxane	375.15	373.35	373.2 ^c	368.63	60	69
6	2,3-butanedione	361.15	toluene	384.15	362.70	374.6 ^c	357.00	95	64
7	methyl ethyl ketone	353.15	acetonitrile	355.15	352.15	349.0	347.50	79	61
8	t-butanol	355.15	1,2-dichloroethane	356.65	349.45	353.6	349.43	40	55
9	2-propanol	355.55	cyclohexene	356.15	344.65	348.6 ^c	349.76	35	39
10	2-propanol	355.55	fluorobenzene	357.15	347.75	351.6 ^c	349.83	35	46
11	acetone	329.35	1-hexene	336.00	323.35	327.5 ^c	324.46	51	65
12	cyclohexane	353.15	trichloroethylene	360.15	353.40	347.3	347.89	86	50
13	cyclohexene	356.15	1,4-dioxane	375.15	355.75	348.6 ^c	351.77	89	69
14	cyclohexene	356.15	2-butanol	372.00	352.75	347.0 ^c	351.53	81	83
15	dichloromethane	313.15	ethanol	351.45	313.05	314.6	310.97	98	86
16	diisopropylether	332.15	methyl ethyl ketone	353.15	340.55	326.7	328.32	85	79
17	ethanol	351.45	fluorobenzene	357.15	343.85	345.0 ^c	346.12	30	38
18	ethyl acetate	350.25	t-butanol	355.15	349.75	348.5	344.88	81	73
19	heptane	371.55	1,4-dioxane	375.15	364.30	363.8	365.73	55	66
20	heptane	371.55	propyl acetate	374.75	366.75	361.6 ^c	365.70	57	59
21	methanol	338.15	ETBE	342.15	330.95	326.1	332.91	32	36
22	methanol	338.15	fluorobenzene	357.15	333.35	332.6 ^c	334.07	40	34
23	methylcyclopentane	344.15	1-propanol	370.15	340.85	338.6	340.51	87	81
24	THF	339.15	acetonitrile	355.15	338.95	337.8	334.82	95	66

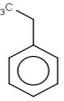
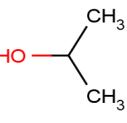
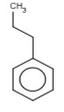
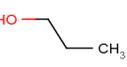
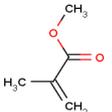
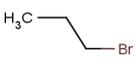
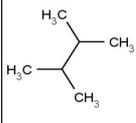
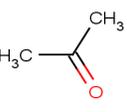
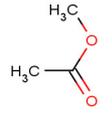
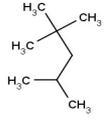
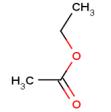
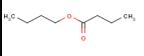
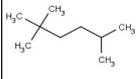
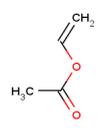
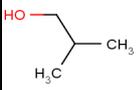
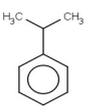
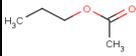
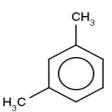
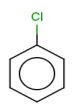
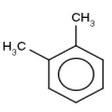
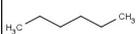
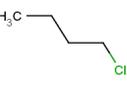
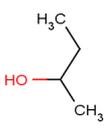
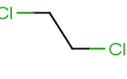
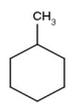
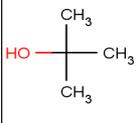
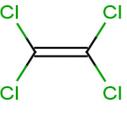
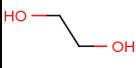
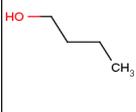
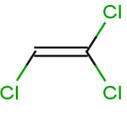
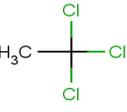
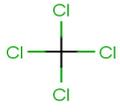
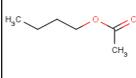
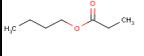
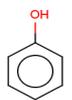
^a) Experimental data are collected from the paper ¹. See the footnote for Table 1. The values of $T_{bp\ az}$ *pred.1* and $X_{Iw, pred}$ are predicted by CMs using AD ($T_{bp\ az}$) and without AD (X_{Iw}); ^b) $T_{bp\ az\ pred.2}$ is predicted according to eq. 1; ^c) outside of AD of CMs.

11.3 Composés formant les mélanges du jeu de modélisation pour la classification

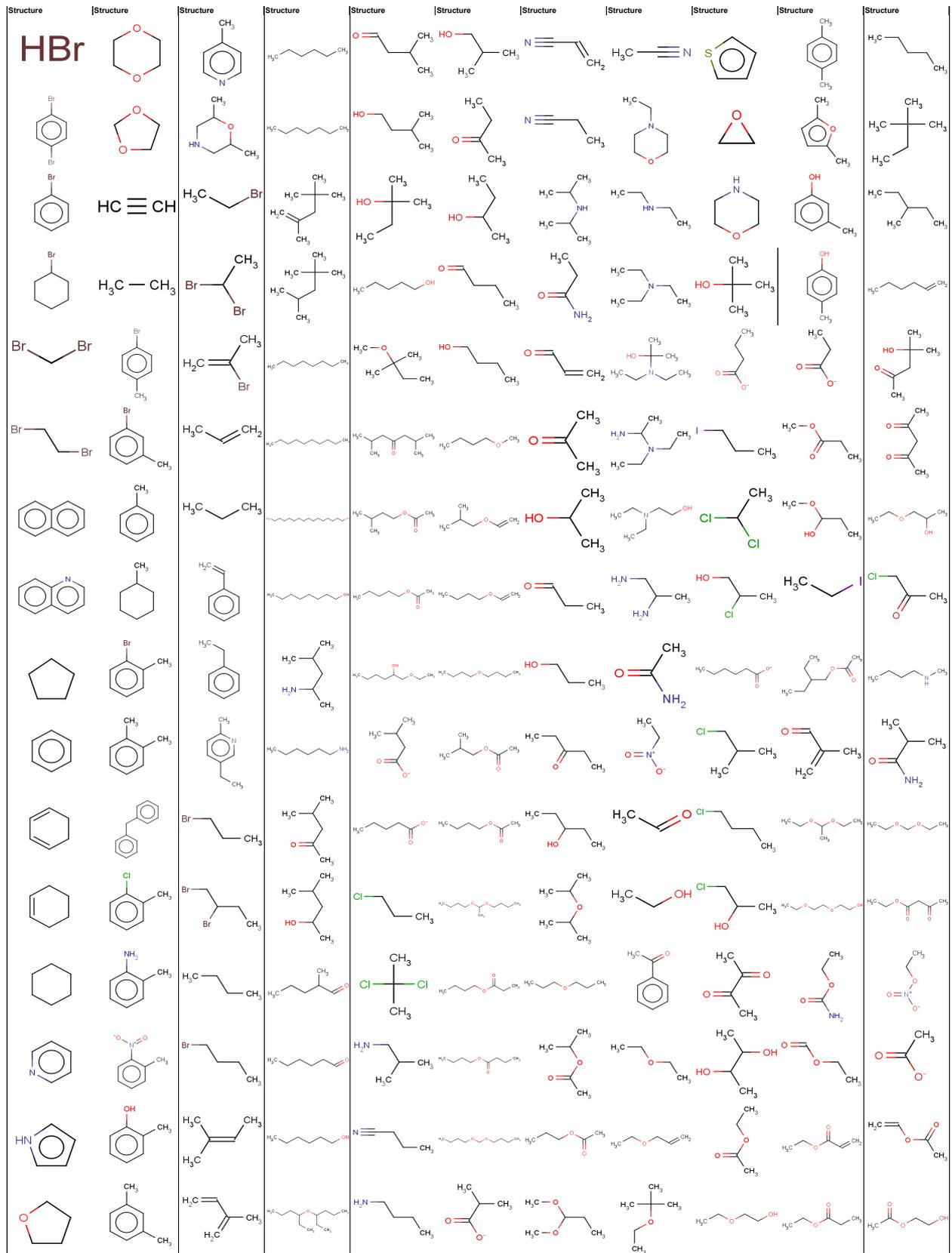
Structure	Nom	Structure	Nom	Structure	Nom	Structure	Nom
	BROMOBENZENE		METHYLCYCLOHEXANE		2-METHYL-2-PROPANOL		1,2-DICHLOROETHANE
	DIBROMOMETHANE		ETHYLBENZENE		BUTANOL		TETRACHLOROETHYLENE
	1-BROMOPROPANE		PROPYLBENZENE		DIBUTYLETHER		TRICHLOROETHYLENE
	CYCLOPENTANE		1-HEXENE		BUTYLFORMATE		1,1-DICHLOROETHANE
	BENZENE		2,2-DIMETHYLBUTANE		BUTYLACETATE		CHLOROFORM
	CUMENE		2,3-DIMETHYLBUTANE		BUTYLPROPIONATE		CARBON TETRACHLORIDE
	m-XYLENE		2-METHYLPENTANE		BUTYLBUTYRATE		1,1,2-TRICHLOROTRIFLUOROETHANE
	o-XYLENE		3-METHYLPENTANE		ISOBUTANOL		1,2-DICHLOROTETRAFLUOROETHANE
	p-XYLENE		2,4-DIMETHYLPENTANE		PROPYLACETATE		ACETONITRILE
	BENZOTRIFLUORIDE		1-OCTENE		ISOPROPANOL		METHYLMETHACRYLATE
	p-CRESOL		2,2,4-TRIMETHYLPENTANE		PROPANOL		METHYLACETATE
	TOLUENE		2,2,5-TRIMETHYLHEXANE		ACETONE		METHANOL

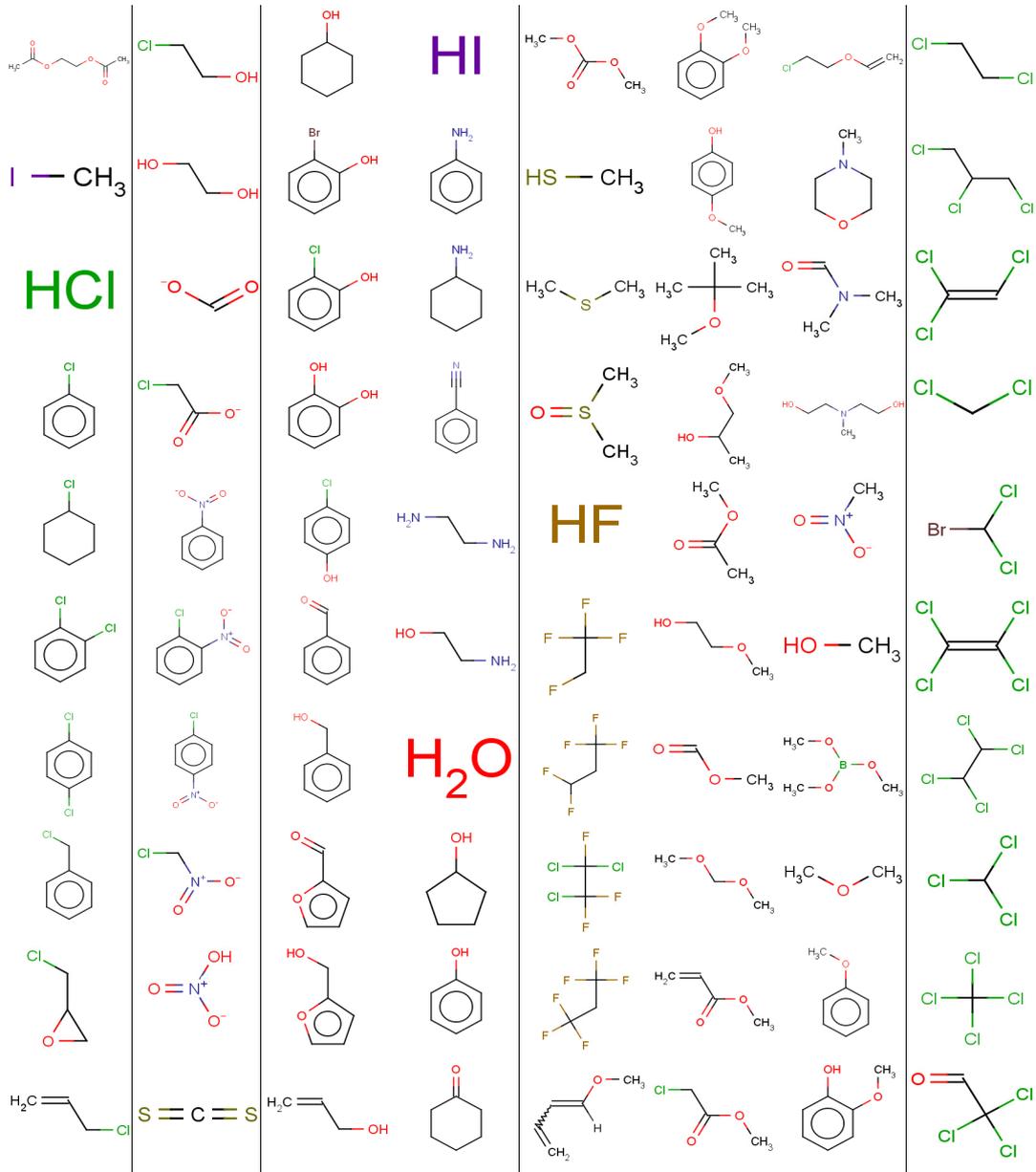


11.4 Composés formant les mélanges du TS1 pour la classification

Structure	Nom	Structure	Nom	Structure	Nom	Structure	Nom
	BROMOBENZENE		ETHYLBENZENE		ISOPROPANOL		ACETONITRILE
	DIBROMOMETHANE		PROPYLBENZENE		PROPANOL		METHYL METHACRYLATE
	1-BROMOPROPANE		2,3-DIMETHYLBUTANE		ACETONE		METHYL ACETATE
	CYCLOPENTANE		2,2,4-TRIMETHYLPENTANE		ETHYL ACETATE		BUTYL BUTYRATE
	BENZENE		2,2,5-TRIMETHYLHEXANE		VINYL ACETATE		ISOBUTANOL
	CUMENE		OCTANE		ETHANOL		PROPYLACETATE
	m-XYLENE		HEPTANE		CHLOROBENZENE	H₂O	WATER
	o-XYLENE		HEXANE		1-CHLOROBUTANE		METHYLCYCLOPENTANE
	p-XYLENE		2-BUTANOL		1,2-DICHLOROETHANE		METHYLCYCLOHEXANE
	p-CRESOL		2-METHYL-2-PROPANOL		TETRACHLOROETHYLENE		ETHYLENEGLYCOL
	TOLUENE		BUTANOL		TRICHLOROETHYLENE	HO - CH₃	METHANOL
	CYCLOHEXANE		DIBUTYLETHER		1,1,1-TRICHLOROETHANE		CARBON TETRACHLORIDE
	CYCLOHEXENE		BUTYLACETATE		CHLOROFORM		BUTYLPROPIONATE
	PHENOL						

11.5 Composés formant les mélanges du TS2 pour la classification





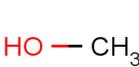
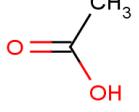
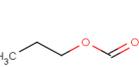
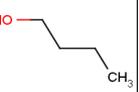
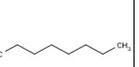
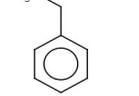
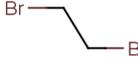
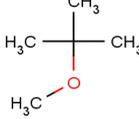
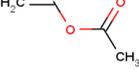
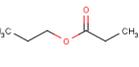
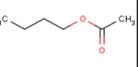
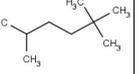
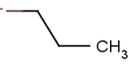
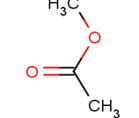
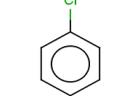
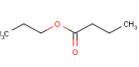
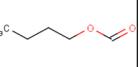
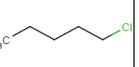
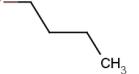
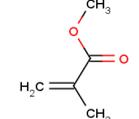
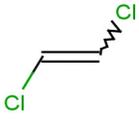
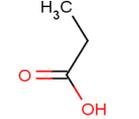
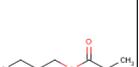
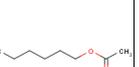
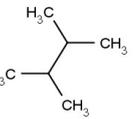
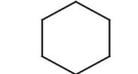
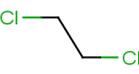
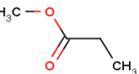
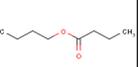
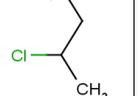
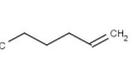
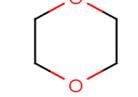
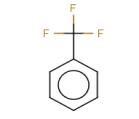
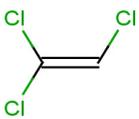
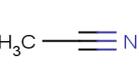
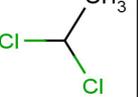
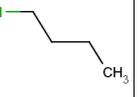
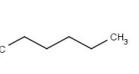
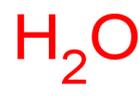
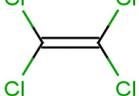
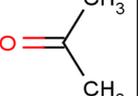
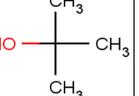
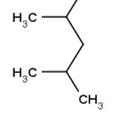
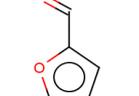
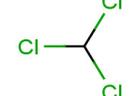
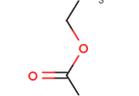
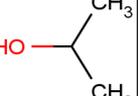
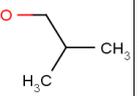
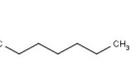
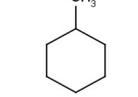
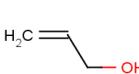
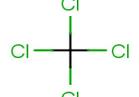
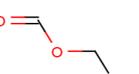
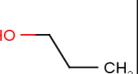
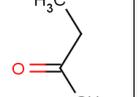
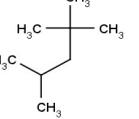
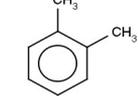
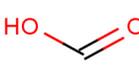
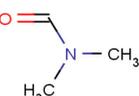
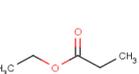
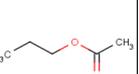
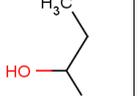
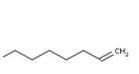
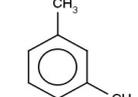
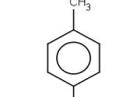
11.6 Composés formant les mélanges du jeu de modélisation pour modéliser la courbe de bulle

| Structure |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

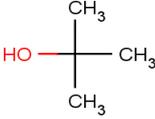
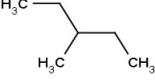
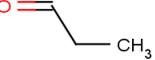
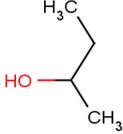
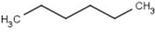
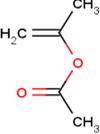
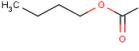
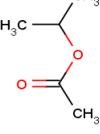
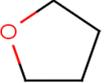
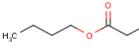
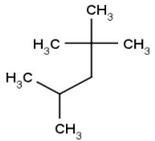
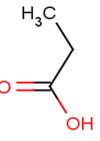
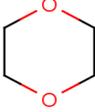
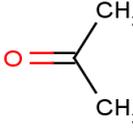
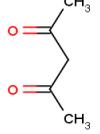
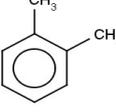
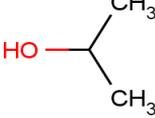
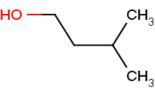
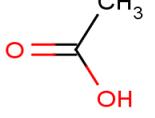
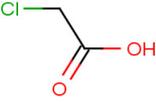
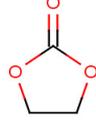
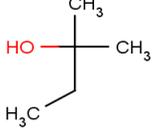
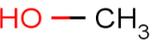
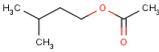
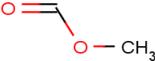
11.7 Composés formant les mélanges du jeu de test pour la pour modéliser la courbe de bulle

Structure	Nom	Structure	Nom	Structure	Nom	Structure	Nom
	BROMOBENZENE		METHYLCYCLOHEXANE		2-METHYL-2-PROPANOL		1,2-DICHLOROETHANE
	DIBROMOMETHANE		ETHYLBENZENE		BUTANOL		TETRACHLOROETHYLENE
	1-BROMOPROPANE		PROPYLBENZENE		DIBUTYL ETHER		TRICHLOROETHYLENE
	CYCLOPENTANE		1-HEXENE		BUTYL FORMATE		1,1-DICHLOROETHANE
	BENZENE		2,2-DIMETHYLBUTANE		BUTYL ACETATE		1,1,1-TRICHLOROETHANE
	CUMENE		2,3-DIMETHYLBUTANE		BUTYL PROPIONATE		CHLOROFORM
	m-XYLENE		2-METHYLPENTANE		BUTYL BUTYRATE		CARBON TETRACHLORIDE
	o-XYLENE		3-METHYLPENTANE		ISOBUTANOL		1,1,2-TRICHLOROTRIFLUOROETHANE
	p-XYLENE		2,4-DIMETHYLPENTANE		PROPYL ACETATE		1,2-DICHLOROTETRAFLUOROETHANE
	BENZOTRIFLUORIDE		1-OCTENE		ISOPROPANOL		ACETONITRILE
	p-CRESOL		2,2,4-TRIMETHYLPENTANE		PROPANOL		METHYL METHACRYLATE
	TOLUENE		2,2,5-TRIMETHYLHEXANE		ACETONE		METHYL ACETATE
	CYCLOHEXANE		DECANE		ETHYL ACETATE		METHANOL
	CYCLOHEXENE		OCTANE		VINYL ACETATE		HEXAFLUOROBENZENE
	PHENOL		HEPTANE		ETHANOL		ETHYLENE GLYCOL
	METHYLCYCLOPENTANE		HEXANE		CHLOROBENZENE		WATER
			2-BUTANOL		1-CHLOROBUTANE		

11.8 Composés formant les mélanges du jeu d'entraînement pour modéliser $y=f(x)$

Structure	Structure	Structure	Structure	Structure	Structure	Structure
						
						
						
						
						
						
						
						
						
						
						

11.9 Composés formant les mélanges du jeu de test pour la modéliser $y=f(x)$

Structure	Structure	Structure	Structure
			
			
			
			
			
			
			
			

Modélisation QSPR de mélanges binaires non-additifs. Application au comportement azéotropique

Généralement les modèles QSPR ne sont utilisés que pour prédire des propriétés des corps purs. Dans cette thèse nous avons développé une approche QSPR permettant de prédire des propriétés non additives de mélanges binaires, plus précisément leur caractère azéotropique/zéotropique. Pour parvenir à ce résultat, plusieurs types de modèles quantitatifs et qualitatifs ont été développés.

L'approche est originale pour deux raisons. Premièrement, peu de travaux de recherche ont été publiés sur des mélanges dont les propriétés sont non-additives. Deuxièmement, plusieurs nouveaux aspects méthodologiques ont été introduits dans ce travail. Tout d'abord des descripteurs "spéciaux", capables de décrire des mélanges ont été proposés. De plus, un protocole robuste d'obtention et de validation des modèles a été utilisé, et un domaine d'applicabilité des modèles fiable a été proposé.

La méthodologie développée pendant cette thèse démontre la fiabilité d'un nouveau concept – les modèles QSPR pour les mélanges. Elle est comparable à d'autres méthodes classiques, quoique n'utilisant qu'un faible nombre de données en comparaison.

Mots clefs : chémoinformatique, QSPR, EVL, mélanges, azéotrope.

Generally, QSPR models are limited to individual compounds. In this thesis we have developed a QSPR approach to predict non-additive properties of binary mixtures, more explicitly their azeotropic behavior. To achieve this, several types of quantitative and qualitative models have been developed.

This approach is original for two reasons. First, little research has been published on mixtures whose properties are no additive. Second, several new methodological aspects have been introduced in this work. First of all "special" descriptors able to describe mixtures have been proposed. In addition, a robust protocol for obtaining and validating models was used, and a reliable models applicability domain was proposed.

The methodology developed during this thesis demonstrates the consistency of a new concept - the QSPR models for mixtures. It is comparable to other conventional methods, though using only limited data.

Keywords: chemoinformatics, QSPR, VLE, mixtures, azeotrope.