

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

THÈSE présentée par :

Grace DELOUIS

soutenue le : **26 septembre 2017**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline / Spécialité : **Chimie / Chémoinformatique**

**Modélisation QSPR de solvants
d'intérêt technologique :
les liquides ioniques et
les électrolytes pour batteries Li-ion**

THÈSE dirigée par :

M. VARNEK Alexandre

Professeur, Université de Strasbourg

RAPPORTEURS :

M. BUREAU Ronan

Professeur, université de Caen Normandie

Mme CAMPROUX Anne-Claude

Professeure, université Paris Diderot

AUTRES MEMBRES DU JURY :

M. MARCOU Gilles

Maître de conférences, HDR, université de Strasbourg

MEMBRE INVITÉ :

M. LACHICHE Nicolas

Maître de conférences, HDR, université de Strasbourg

Qui n'essaye, point saura.

Table des matières

Remerciements	5
Liste des acronymes	8
Introduction générale	10
1 Introduction aux solvants étudiés	15
1.1 Liquides ioniques	15
1.1.1 Généralités	15
1.1.2 Revue de la littérature pour la modélisation QSPR des liquides ioniques	16
1.1.3 Facteurs externes influençant les propriétés des liquides ioniques . . .	17
1.2 Les électrolytes pour batteries Li-ion	21
1.2.1 Généralités	21
1.2.2 Revue de la littérature pour la modélisation des électrolytes	23
1.3 En résumé	25
2 Méthodes de modélisation	27
2.1 Descripteurs	27
2.1.1 Descripteurs fragmentaux ISIDA	28
2.1.2 Descripteurs d'effets électroniques	29
2.1.3 Descripteurs MOE2D	30
2.2 Méthodes d'apprentissage automatiques	31
2.2.1 Régression à vecteurs supports	31
2.2.2 Apprentissage semi-supervisé et transduction	32
2.2.3 Régression ridge transductive	34
2.3 Domaine d'applicabilité	37
2.4 Validation croisée	37
2.5 Paramètres statistiques de validation et de vérification des modèles	38
2.6 Détection des points aberrants	39
2.7 Comparaison des modèles	40
2.8 En résumé	40

3	Étude de la TRR	43
3.1	Implémentation de la TRR	43
3.2	Effet transductif	43
3.3	Jeux de données utilisés	44
3.4	Courbes d'optimisation des paramètres g et gp	44
3.5	Impact de la taille relative du jeu d'entraînement et du paramètre gp sur l'effet transductif	46
3.5.1	Procédure de modélisation	46
3.5.2	Recherche d'un gp optimal	48
3.5.3	Recherche d'un gp par validation croisée	50
3.5.4	Étude des cas pour lesquels l'effet transductif est négatif	56
3.6	En résumé	56
4	Modélisation des liquides ioniques	59
4.1	Présentation des données	59
4.2	Modélisation avec la SVR	61
4.2.1	Procédure de modélisation	61
4.2.2	Points aberrants dans les jeux d'entraînement	63
4.2.3	Résultats obtenus sur les jeux de validation	65
4.2.4	Mise en ligne des modèles	71
4.3	Modélisation avec la TRR	71
4.3.1	Procédure	75
4.3.2	Modèles consensus	76
4.3.3	Comparaison de l'ensemble des modèles individuels	78
4.3.4	Comparaison de la SVR, de la RR et de la TRR pour un espace de descripteurs fixé	79
4.4	En résumé	79
5	Modélisation des électrolytes	83
5.1	Données	83
5.2	Modélisation avec la SVR	86
5.2.1	Procédure	86
5.2.2	Résultats	86
5.2.3	Points aberrants	89
5.2.4	Predictor	89
5.2.5	Criblage et sélection de candidats	89
5.3	Modélisation avec la TRR	93
5.3.1	Procédure	93
5.3.2	Résultats	96
5.4	En résumé	99

Conclusion	99
Bibliographie	120

Remerciements

Je remercie mon directeur de thèse, Alexandre Varnek, pour avoir accepté de diriger ma thèse.

Je remercie Gilles Marcou pour m'avoir encadré pendant ces quatre ans et pour les nombreux conseils qu'il m'a donnés.

Je remercie Nicolas. Lachiche pour sa collaboration sur le projet TRR.

Je remercie mes rapporteurs de thèse Ronan Bureau et Anne-Claude Camproux pour avoir accepté de juger mes travaux.

Je remercie Dragos Horvath pour son aide sur la mise en ligne des modèles et pour son humour.

Je remercie Olga Klimshuk pour son aide sur le projet ILDB et pour sa bienveillance.

Je remercie Guillaume Beck, Olena Mokshyna et Benjamin Flamme pour leur aide plus que précieuse sur le projet DEVEGA.

Je remercie les différentes équipes et sociétés avec lesquelles nous avons collaboré : l'équipe d'Isabelle Billard (IPHC, université de Strasbourg) et la compagnie Solvionic pour le projet ILDB, ainsi que les équipes de Alexandre Chagnes (Chimie Paris-Tech), de Jean-Marie Tarascon (Collège de France) et de Laurent Joubert (COBRA, université de Rouen) pour le projet DEVEGA.

Je remercie Fanny pour sa bonne humeur et son soutien.

Je remercie les doctorants et stagiaires avec qui j'ai passé de bons moments : Pavel, Hélène, Fiorella, Julien, Arcadii, Timur, Martha, Shilva, Kirill, Iuri, Albina, Guillaume, Valentin, Olena, Tetiana.

Je remercie les stagiaires que j'ai eu l'occasion d'encadrer, Charline Fagnen et Baptiste Foucher, et toutes les personnes qui ont eu un rapport, de près ou de loin, avec l'enseignement.

Je remercie Soumia pour sa gentillesse et son aide concernant le domaine administratif.

Je remercie Véronique Bulach, Jean-Serge Rémy, et Petra Hellwig, à la direction de l'école doctorale des sciences chimiques, pour leur soutien et pour m'avoir permis de mener cette thèse jusqu'au bout malgré les retards et les difficultés rencontrées.

Je remercie Marine, Aurore, Mathilde, Alex pour avoir été des compagnons de galère géniaux.

Je remercie Mutti, Vati, Gilles et Carole pour leur soutien, leur aide et leurs encouragements, bien que cela ne soit pas toujours évident quand il y a 800 kilomètres qui nous séparent.

Remerciements

Je remercie mes amis, le clan de la Garde : Adrien, Arnaud, Ben, Chim, Delphine, JP, Manon, Megan, Morgane, Paul, Ritchy. Vous êtes devenus une seconde famille pour moi. Vous avez été formidables, et si je suis arrivée au bout, c'est en grande partie grâce à vous. Je tiens en particulier à exprimer ma gratitude envers Chim et JP qui m'ont aidé à y voir clair dans les moments les plus sombres.

Enfin, le plus important, je remercie mon petit ami, Nicolas Boulay. Tu m'a soutenu de façon inconditionnelle pendant ces quatre années, dans les bons comme dans les mauvais moments, alors que tu es toi aussi en thèse. Je ne te remercierais jamais assez pour tout ce que tu as fait pour moi. J'ai hâte de te voir soutenir aussi, tu le mérites amplement.

Liste des acronymes

- A2AR – récepteur adénosine A2A (*A2A adenosine receptor*)
- AAD – écart absolu moyen (*Absolute Average Deviation*)
- AARD – écart relatif absolu l'écart relatif absolu moyen (*Absolute Average Relative Deviation*)
- AIMD – dynamique moléculaire ab initio (*Ab Initio Molecular Dynamics*)
- ARD – écart absolu relatif (*Absolute Relative Deviation*)
- Co-FTF – algorithme « ajuster les ajustements » (*Fitting-the-Fits based co-training algorithm*)
- CODESSA – descripteurs compréhensifs pour l'analyse structurale et statistique (*COMprehensive DEscriptors for Structural and Statistical Analysis*)
- COREG – régresseur co-entraînés (*Co-training REGressors*)
- COSMO-RS – modèle de criblage similaire à un conducteur pour solvants réels (*CONductor like Screening MOdel for Real Solvents*)
- COSMOtherm – modèle de criblage similaire à un conducteur pour les propriétés thermodynamiques (*CONductor like Screening MOdel for thermodynamic properties*)
- CV – validation croisée (*Cross Validation*)
- DA – domaine d'applicabilité
- DEC – carbonate de diéthyl (*diethyl carbonate*)
- DFT – théorie fonctionnelle de la densité (*Density Functionnal Theory*)
- DMC – carbonate de diméthyl (*dimethyl carbonate*)
- DSC – calorimétrie différentielle à balayage (*Differential Scanning Calorimetry*)
- EC – carbonate d'éthylène (*ethylene carbonate*)
- EED – descripteurs d'effets électroniques (*Electronic effect descriptors*)
- EMC – carbonate d'éthylméthyl (*ethylmethyl carbonate*)
- Eox – Potentiel d'oxydation (*oxydation potential, V*)
- IEE – interface électrolyte-électrode
- IL – liquide ionique (*Ionic Liquid*)
- IP – potentiel d'ionisation (*Ionization Potential*)
- IPHC – Institut Pluridisciplinaire Hubert Curien
- ISIDA – Ensemble des outils développés au laboratoire pour la conception *in silico* et l'analyse de données (*In Silico Design and Data Analysis*)

Liste des acronymes

- k -CV – validation croisée en k paquets (*k folds Cross Validation*)
- kNN – les k voisins les plus proches (*k Nearest Neighbours*)
- LLGC – apprendre avec la cohérence locale et globale (*Learning with Local and Global Consistency*)
- LogS – logarithme de la solubilité aqueuse (*Logarithm of the aqueous Solubility*)
- LOO – validation croisée avec un seul composé écarté (*Leave One Out cross validation*)
- MD – dynamique moléculaire (*Molecular Dynamics*)
- MOE – environnement d’opération moléculaire (*Molecular Operating Environment*)
- MP2 – théorie de la perturbation de Møller-Plesset au second ordre (*second order Møller-Plesset perturbation theory*)
- MPE – pourcentage d’erreur moyen (*Mean Percentage Error*)
- MR – ordonnancement multiple (*Multi Ranking*)
- MSE – erreur quadratique moyenne (*Mean Squared Error*)
- MTL – apprentissage multi-tâches (*Multi-Task Learning*)
- NCI – Institut National du Cancer aux États-Unis (*National Cancer Institute of the United States*)
- $n \times k$ -CV – validation croisée en k paquets répétée n fois (*n times k folds Cross Validation*)
- PC – carbonate de propylène (*propylene carbonate*)
- PCT – arbre de classification prédictif (*Predictive Clustering Tree*)
- pKa – constante d’acidité
- pKi – $-\log_{10}(\text{affinité})$
- PM3 – modèle semi-empirique paramétré numéro 3 (*Parameterized Model number 3*)
- Q^2 – coefficient de corrélation
- QSPR – relation quantitative structure-propriété (*Quantitative Structure-Property Relationship*)
- R^2 – coefficient de détermination
- RAAD – écart absolu moyen relatif (*Relative Average Absolute Deviation*)
- RMSE – racine de l’erreur quadratique moyenne (*Root Mean Squared Error*)
- rNN – les voisins les plus proches selon un rayon (*radius Nearest Neighbours*)
- RR – régression ridge (*Ridge Regression*)
- SVR – régression à vecteurs supports (*Support Vector Regression*)
- TE – effet transductif (*Transductive Effect*)
- Teb. – température d’ébullition (°C)
- Tf₂N – bis(trifluorométhylsulfonyl)imide
- TFSI – bis(trifluorométhylsulfonyl)imide
- Tfus. – température de fusion (°C)
- TRR – régression ridge transductive (*Transductive Ridge Regression*)
- WFT – théorie de la fonction d’onde (*Wave Function Theory*)
- YATSI – algorithme « nouvelle idée en deux étapes » (*Yet Another Two Stage Idea algorithm*)

Introduction générale

Les solvants sont des substances qui permettent de dissoudre des espèces chimiques sans que celles-ci ne soient altérées à leur contact. Ils sont essentiels dans le domaine de la chimie, et sont employés dans de nombreuses applications, parmi lesquelles nous pouvons citer la synthèse, la catalyse, la séparation, ou encore le stockage électrochimique de l'énergie. Si certains solvants sont assez classiques, comme l'eau, l'acétone ou l'éthanol, d'autres peuvent être classés dans les solvants à haute valeur technologique.

La conception de solvants se heurte à plusieurs problèmes. Le premier d'entre eux est qu'ils sont souvent utilisés en grande quantité, ce qui impose une contrainte économique forte sur leur coût de développement et de production [1]. Contrairement au domaine du médicament par exemple, il est déraisonnable d'utiliser des approches de criblage à haut débit car les investissements qui seraient demandés ne seraient pas rentabilisés. Un moyen de réduire le coût de développement de nouveaux solvants est le recours à la simulation et au calcul. La solution que nous avons développée dans cette thèse repose sur le criblage virtuel utilisant des modèles statistiques. Cette approche permet de proposer rapidement des listes de composés susceptibles de présenter des avancées concernant une propriété d'intérêt technologique.

Le but de cette thèse est de modéliser deux catégories de solvants à haute valeur technologique : les liquides ioniques et les électrolytes. La stratégie est de développer des modèles statistiques de relation structure-propriété (*Quantitative Structure-Property Relationship* en anglais, QSPR) [2]. Ces modèles sont rapides et demandent peu d'expertise pour être utilisés, ce qui en font de bons outils pour cribler ou annoter des chimiothèques.

Nous nous sommes d'abord intéressé aux liquides ioniques. Ce sont des composés constitués d'un cation organique et d'un anion, organique ou inorganique, dont la température de fusion est inférieure à 100 °C [3, 4]. Ils connaissent un succès grandissant depuis la fin des années 1990 car ils ont une faible tension de vapeur et sont réputés ininflammables. Ces composés sont donc des solvants très prisés dans le cadre de la chimie verte, domaine visant à limiter au maximum l'impact de la chimie sur l'environnement.

La modélisation des liquides ioniques a été faite dans le cadre d'une collaboration avec l'équipe d'Isabelle Billard à l'IPHC de Strasbourg, ainsi qu'avec la société Solvionic, spécialisée dans la vente de liquides ioniques. Le point de départ était de faciliter le développement de liquides ioniques dédiés à l'extraction liquide-liquide pour le recyclage de terres rares par

exemple [5, 6]. La société Solvionic nous a fourni des échantillons de son catalogue pour tester nos modèles en échange de l'annotation par le calcul des composés pour lesquels les mesures physiques n'étaient pas encore disponibles. Nous nous sommes donc concentrée sur les propriétés de transport et l'état physique du liquide ionique : la conductivité, la température de fusion et la viscosité. Il aurait été souhaitable de modéliser également la miscibilité du liquide ionique avec l'eau, en vue d'applications de procédés de séparation, mais nous n'avons pas pu rassembler suffisamment d'informations à ce sujet. Plusieurs questions se sont posées ici :

- Comment modéliser un sel contenant deux espèces chimiques ? En effet, les approches QSPR sont généralement développées pour des corps purs, et les applications aux mélanges sont, en comparaison, assez rares.
- En quoi la qualité des données utilisées pour la construction du modèle est-elle déterminante ? Quoique la littérature sur les liquides ioniques soit abondante, elle s'illustre par une grande variété de situations expérimentales. Il a donc fallu mener une réflexion sur le contrôle qualité des données auxquelles nous avons accès.
- Est-il possible de construire des modèles utiles avec peu de données disponibles ? Le domaine des liquides ioniques est assez prolifique [7]. Aussi est-il rapidement apparu un déséquilibre important entre les nouveaux liquides ioniques disponibles et ceux pour lesquels des données utiles à la modélisation avaient été collectées.

La seconde catégorie de solvants modélisée concerne les électrolytes pour batteries Li-ion. Nous désignons ici par le terme « électrolyte » le solvant électrolytique utilisé pour dissoudre le sel. Il s'agit d'un raccourci employé ici pour éviter d'alourdir le discours. Il s'agit d'un élément crucial dans la conception d'une batterie : c'est lui qui influe sur la vitesse de libération de l'énergie [8]. À l'heure actuelle, où l'on cherche des alternatives aux énergies fossiles, le stockage de l'énergie électrique est de plus en plus crucial. La recherche de nouveaux électrolytes est donc essentielle.

La modélisation des électrolytes s'est faite dans le cadre de l'ANR DEVEGA [9], en collaboration avec Laurent Joubert de l'université de Rouen, Alexandre Chagnes de Chimie ParisTech et Jean-Marie Tarascon du Collège de France. Nous avons modélisé 6 propriétés différentes (la conductivité, la constante diélectrique, le potentiel d'oxydation, la température d'ébullition, la température de fusion et la viscosité) afin de proposer de nouveaux électrolytes potentiels pour la conception de batteries Li-ion pour véhicules électriques. Nos collaborateurs ont ensuite synthétisé et mesuré la conductivité, le potentiel d'oxydation et la température d'ébullition de certains de ces candidats. Lors de ce projet, plusieurs questions se sont posées :

- Est-il possible de construire des modèles utiles avec peu de données disponibles ? Là encore, nous sommes confrontés à un type de composé pour lequel il y a peu de données dans la littérature. En effet, peu de solvants peuvent répondre aux critères définissant l'électrolyte idéal (tels qu'un point de fusion bas, une température d'ébullition élevée, une faible viscosité, une conductivité élevée, une large fenêtre électrochimique, etc.) Il est en pratique quasi impossible de combiner toutes ces qualités dans un seul solvant,

les électrolytes commerciaux sont donc généralement des cosolvants afin de réussir à combiner toutes les qualités nécessaires à la confection de batteries.

- Comment sélectionner une liste de candidats potentiels ? En effet, nous cherchons ici à identifier les candidats les plus susceptibles de correspondre aux besoins de nos collaborateurs. Pour cela, nous devons faire cette sélection sur 6 propriétés simultanément. Nous devons donc choisir une stratégie permettant, à partir d'un grand nombre de structures générées *in silico*, d'obtenir une liste d'électrolytes potentiels.

La question de la quantité de données s'est posée pour les deux types de solvants étudiés. Nous travaillons ici dans des domaines pointus, pour lesquels l'acquisition de nouvelles données peut être difficile et coûteuse, aussi bien pour la conception de nouvelles substances que pour la mesure de leurs propriétés physico-chimiques. La quantité de données est donc faible, ce qui est un frein potentiel à la conception de modèles prédictifs.

Pour palier à ce problème, une stratégie possible consiste à utiliser des algorithmes d'apprentissage transductif [10] (souvent associés aux méthodes semi-supervisées [11]). Le principe de ces méthodes est d'améliorer les performances des modèles en les spécialisant sur les données dont on cherche à estimer spécifiquement les propriétés. En théorie, nous pouvons donc améliorer les performances d'un modèle en ajoutant, aux données pour lesquelles la propriété d'intérêt est connue (les données étiquetées), de nombreuses structures potentiellement intéressantes mais jamais testées (les données non étiquetées). Étant donné que les outils informatiques actuels nous permettent relativement facilement de générer un grand nombre de structures hypothétiques, cette approche pourrait être très intéressante. Toutefois, cette approche est rarement mise en pratique et son intérêt pour la modélisation QSPR n'a jamais été démontré.

En plus de la SVR (régression à vecteurs supports, *Support Vector Regression* en anglais), une méthode utilisant une fonction de coût ϵ -sensible qui la rend robuste aux points aberrants, nous nous sommes donc intéressé à la RR (Régression Ridge) et son homologue transductif, la TRR (Régression Ridge Transductive). La RR est une méthode de régression linéaire dont la fonction de coût est l'erreur quadratique moyenne. C'est une méthode classique qui sert de référence. La TRR, quant à elle, allie le principe de la transduction et la méthode RR. Il s'agit d'un algorithme type pour explorer la transduction, quantifier ses effets et chercher les facteurs pouvant en moduler l'intensité. L'étude de la TRR nous a amené à nous intéresser aux points suivants :

- Quelles sont les valeurs recommandées pour les paramètres de ces méthodes ? La TRR possède deux paramètres à régler manuellement, nous devons donc mettre en place une procédure pour les déterminer.
- Quel est l'impact de la quantité de données non étiquetées sur les performances du modèle TRR ? Il est en effet intéressant de savoir s'il est nécessaire de rajouter un grand nombre de données non étiquetées, ou si une faible quantité suffit.

Cette thèse contient 5 chapitres. Le premier chapitre, introductif, s'intéresse aux solvants modélisés. Le second chapitre présente les outils chémoinformatiques utilisés, en particulier la

TRR. Le troisième chapitre expose les développements méthodologiques concernant la TRR. Les chapitres 4 et 5 concernent respectivement la modélisation des liquides ioniques et celle des électrolytes.

Chapitre 1

Introduction aux solvants étudiés

Un solvant est une substance permettant la dissolution d'espèces chimiques sans que celles-ci ne soient altérées à son contact. Ils sont couramment utilisés en chimie pour de nombreuses applications telles que la synthèse, la catalyse, la séparation, etc. En raison de notre collaboration avec diverses équipes de recherche et un partenaire industriel, nous nous sommes ici intéressés à deux catégories de solvants à haute valeur technologique : les liquides ioniques et les électrolytes pour batterie Li-ion.

1.1 Liquides ioniques

1.1.1 Généralités

Les liquides ioniques (LI) sont des composés constitués d'un cation organique et d'un anion, organique ou inorganique [3, 4]. D'un point de vue historique, la température de fusion de ces composés est fixée comme étant inférieure à 100 °C et ceux qui ont un point de fusion inférieur à 20 °C sont couramment appelés « liquides ioniques à température ambiante ». Les cations les plus couramment rencontrés sont les N,N-dialkylimidazoliums, les alkylphosphoniums et les alkylpyridiniums, tandis que les anions généralement utilisés sont les halogénures, le bis(trifluorométhylsulfonyl)imide (Tf₂N), l'héxafluorophosphate et les sulfates [12, 13].

Les caractéristiques courantes des liquides ioniques sont les suivantes : une faible pression de vapeur saturante, une conductivité généralement comprise entre 3 mS/cm et 7 mS/cm, une large fenêtre électrochimique, une faible compressibilité et une viscosité relativement élevée par rapport aux solvants organiques usuels [12]. Ce sont des substances polaires hautement hygroscopiques [14] qui sont stables d'un point de vue chimique et thermique [7, 12].

Les propriétés physico-chimiques des liquides ioniques dépendent de leur composition en terme d'ions. Pour cette raison, ce sont des substances hautement personnalisables : leur composition peut être conçue pour coller aux spécifications physico-chimiques correspondant à une application technologique donnée [7]. Par exemple, en synthèse organique, un liquide

ionique idéal aurait une température de fusion inférieure à 20 °C, une viscosité la plus proche possible de celle de l'eau (environ 0,89 cP à 25 °C [3]) et une conductivité supérieure à 5 mS/cm [15].

De nos jours, les liquides ioniques sont utilisés dans de nombreux domaines. Ils sont employés comme solvants [16] et catalyseurs [17] pour la synthèse, et comme électrolytes [15] dans les batteries pour le stockage de l'énergie et la conversion. Ils sont également utilisés dans de nombreux procédés industriels comme la compression de gaz [18] et le recyclage de la cellulose [19]. Ils ont également des applications inattendues en tant que miroirs liquides [20], agents d'embaumement [21], administration transdermique de médicaments [22] et dans la neutralisation de pathogènes [23]. Récemment, un article a rapporté le premier exemple de produit naturel étant un liquide ionique [24].

Cependant, il est difficile de concevoir un liquide ionique optimal pour une application particulière. En effet, le nombre possible de combinaisons cation-anion pouvant potentiellement être un liquide ionique est évalué à 10^{18} , ce qui rend impossible la synthèse et l'analyse de chacun d'entre eux [25]. Pour remédier à ce problème, une alternative intéressante est d'utiliser des modèles QSPR (« Quantitative Structure-Property Relationship », relation quantitative structure-propriété) pour cribler des liquides ioniques virtuels afin de maximiser les chances d'obtenir une substance avec les propriétés voulues.

1.1.2 Revue de la littérature pour la modélisation QSPR des liquides ioniques

De nombreuses études QSPR ont été publiées récemment. Pour cette raison, nous n'avons pas fait une revue exhaustive de la littérature sur l'ensemble des propriétés modélisables des liquides ioniques. Nous avons préféré nous focaliser sur les trois propriétés modélisées au cours de cette thèse.

Parmi les trois propriétés qui vont nous intéresser, la température de fusion est celle qui a été le plus étudiée. Le tableau 1.1 consigne les différents résultats obtenus à ce sujet. La première étude a été menée par A.R. Katritzky *et al.* en 2002 [25]. Comme pour le travail de Katritzky *et al.*, la grande partie de ces études concerne des jeux de données de bromure, contenant des sels fondus plus que des liquides ioniques car leur température de fusion est généralement supérieure à 25 °C. De plus, la plupart de ces études travaillent avec un anion constant (Br^- , NO_3^- ou nitrocyanamide). Ce sont donc des études QSPR classiques, puisqu'elles ne prennent pas en compte le fait que les liquides ioniques soient des mélanges. Les études restantes rapportent généralement une erreur de prédiction variant entre 25 °C et 50 °C, ou bien n'ont pas de protocole de validation externe clair, ce qui signifie que leurs performances sont optimistes.

La viscosité a elle aussi été largement étudiée. Le tableau 1.2 liste les différentes études recensées. Le premier article retrouvé, à notre connaissance, a été écrit par K. Tochigi and

H. Yamamoto en 2007 [26]. Trois grandes approches ont été utilisées dans les diverses études : calculer les contributions de groupe, utiliser l'approche COSMO-RS [27] ou calculer des descripteurs moléculaires (généralement des descripteurs CODESSA [28] ou Dragon [29]). Peu importe l'outil d'apprentissage par machine choisi, les performances des modèles sont extrêmement élevées. De notre point de vue, ceci est dû à la procédure de validation utilisée dans ces études. Le jeu d'entraînement et le jeu de validation partagent des liquides ioniques identiques qui diffèrent par un seul paramètre, généralement la température de la mesure. Considérant cela, les statistiques reportées dans ces articles doivent être comprises comme la capacité du modèle à pratiquer une interpolation sur la variation de la viscosité par rapport à ce paramètre. En plus de cela, certains modèles utilisent la méthode de contribution de groupe, pour lesquels les groupes considérés sont grands : la contribution de l'anion complet est considérée. Ceci limite l'utilité d'une telle approche puisqu'un nouveau liquide ionique ne pourra pas être prédit s'il contient un nouvel anion. Cette remarque s'applique à toutes les études travaillant avec cette méthode. La seule étude qui fait varier à la fois le cation et l'anion et présentant un protocole de validation rigoureux est Billard [30]. De cette analyse, la précision attendue sur les modèles de la viscosité est d'environ 80 cP.

La conductivité est la propriété la moins étudiée, et la plupart des études en rapport avec cette propriété ont également travaillé sur la viscosité, et sont résumées dans le tableau 1.3. Les premiers ayant travaillé avec cette propriété, à notre connaissance, sont K. Tochigi et H. Yamamoto en 2007 [26]. Bien que les performances généralement reportées soient prometteuses, elles doivent être interprétées comme celles de la viscosité. En effet, le jeu de validation est généralement composé des mêmes liquides ioniques que ceux présents dans le jeu d'entraînement, mais dont la mesure a été obtenue à une température différente. Les performances rapportées sont reliées à leur capacité à interpoler la conductivité d'un liquide ionique à une température donnée. Elles sont donc optimistes. La seule exception concerne le travail de Matsuda *et al.* [31], pour lequel une source additionnelle de données a été utilisée pour tester le modèle. Cependant, les données du jeu d'entraînement et du jeu de validation ne sont pas fournies, nous ne pouvons donc pas vérifier qu'il n'y a pas de recouvrement entre ces deux jeux de données, et nous ne pouvons pas conclure avec certitude que celui-ci est nul. Les performances rapportées ici sont donc probablement optimistes elles aussi. De ces travaux, la précision attendue des modèles est d'environ 2,25 mS/cm.

1.1.3 Facteurs externes influençant les propriétés des liquides ioniques

Les propriétés des liquides ioniques sont sensibles à des facteurs externes qui sont rarement pris en compte lors de leur modélisation. Nous nous sommes intéressés aux facteurs principaux suivants : la pureté des liquides ioniques, leur teneur en eau (%w), l'identification du point de fusion et la température à laquelle la mesure est effectuée.

Auteurs	Nombre de IL	Nombre/Nature de l'anion	Performances rapportées	Réf.
Aguirre <i>et al.</i>	136	14 anions	Dataset entier : ARD = 7,8 % et AAD = 22,6 K	[32]
Bini <i>et al.</i>	126	Br ⁻	Test : RMSE = 23,78 K, R ² = 0,8725	[33]
Carrera <i>et al.</i>	135	Br ⁻	CV : RMS entre 17,11 °C et 31,50 °C ; R ² entre 0,578 et 0,877	[34]
Carrera <i>et al.</i>	101	Cl ⁻ , BPh ₄ ⁻ , Br ⁻ , and I ⁻	Test : RMSE entre 20,30 °C et 30,87 °C ; R ² entre 0,670 et 0,909	[35]
Eike <i>et al.</i>	173	Br ⁻	R ² entre 0,716 et 0,790 selon la famille de cations	[36]
Farahani <i>et al.</i>	705	62 anions	Test : Q ² = 0,689	[37]
Fatemi <i>et al.</i>	62	Cl ⁻ , Br ⁻ , I ⁻ , BF ₄ ⁻ , CF ₃ SO ₃ ⁻ , SO ₄ ⁻ , Tf ₂ N ⁻	Test : R ² entre 0,79 et 0,85	[38]
Gharagheizi <i>et al.</i>	799	60 anions	Test : RMSE = 24,86 ; R ² = 0,817	[39]
Hada <i>et al.</i>	23	8 anions	Dataset entier : R ² entre 0,541 et 0,609	[40]
Huo <i>et al.</i>	190	14 anions	Dataset entier : RMSE = 28,20 K ; R ² = 0,8984	[41]
Katritzky <i>et al.</i>	149	Br ⁻	CV : R ² entre 0,5961 et 0,9166	[42]
Katritzky <i>et al.</i>	126	Br ⁻	CV : R ² entre 0,7002 et 0,7624	[25]
Kireeva <i>et al.</i>	717	Br ⁻	Test : précision entre 0,78 et 0,85	[43]
Lazzus	400	36 anions	Test : AARD(pred) = 6,16 %	[44]
Lopez-Martin <i>et al.</i>	84	22 anions	Test : R ² entre 0,869 et 0,955	[45]
Preiss <i>et al.</i>	57	10 anions	LOO : erreur moyenne = 26,4 °C	[46]
Preiss <i>et al.</i>	520	57 anions	Dataset entier : R ² = 0,537	[47]
Ren <i>et al.</i>	288	Br ⁻	Test : R ² entre 0,712 et 0,810	[48]
Sun <i>et al.</i>	38	BF ₄ ⁻ , PF ₆ ⁻	CV : R ² entre 0,7763 et 0,8423	[49]
Torrecilla <i>et al.</i>	97	29 anions	Train : R ² entre 0,95 et 0,99	[50]
Trohalaki et Patcher	33	Br ⁻ , NO ₃ ⁻ , nitro-cyanamides	LOO : R ² entre 0,839 et 0,960	[51]
Trohalaki <i>et al.</i>	26	Br ⁻ , NO ₃ ⁻	LOO : R ² entre 0,914 et 0,933	[52]
Valderrama <i>et al.</i>	671	Cl ⁻ , B ⁻ , I ⁻ , BF ₄ ⁻ , ClO ₄ ⁻ , NO ₃ ⁻ , Tf ₂ N ⁻	R ² entre 0,935 et 0,984 selon l'anion	[53]
Varnek <i>et al.</i>	717	Br ⁻	Dataset entier : RMSE entre 37,5 et 46,4 °C ; R ² entre 0,52 et 0,63	[54]
Yan <i>et al.</i>	394	25 anions	Test : R ² = 0,753	[55]

Tableau 1.1 – Compilation de différentes études QSPR pour la température de fusion. Les différents indicateurs de performances utilisés sont l'écart absolu relatif (Absolute Relative Deviation, ARD), l'écart absolu moyen (Absolute Average Deviation, AAD), la précision, l'erreur moyenne, l'écart relatif absolu moyen (Absolute Average Relative Deviation, AARD), la racine de l'erreur quadratique moyenne (Root Mean Squared Error, RMSE), le coefficient de détermination R² et le coefficient de corrélation Q².

Auteurs	Nombre de IL	Nombre de données	Performances rapportées	Réf.
Alcalde et al	27	Plus de 146000	CV : $R^2 = 0,99$	[56]
Barycki <i>et al.</i>	23	138	Test : RMSE entre 0,232 et 0,244 log(cP) ; R^2 entre 0,795 et 0,830	[57]
Billard <i>et al.</i>	122	122	CV : $R^2 = 0,73$; RMSE = 67,5 cP ; Test : RMSE=73 cP	[30]
Bini <i>et al.</i>	33	66	CV : R^2 entre 0,4107 et 0,9287	[33]
Daniel <i>et al.</i>	4	20	Dataset entier : moyenne des écarts en pourcentage de 7,64 % ; $R^2=0,98$	[58]
de Riva	134	1860	Dataset entier : $R^2 > 0,99$ pour toutes les équations	[59]
Diaz-Rodriguez <i>et al.</i>	2 mélanges, 3 IL différents	156	CV : $R^2 > 0,99$; MPE = 3,24 %	[60]
Diaz-Rodriguez <i>et al.</i>	4 mix	639	Test : MPE entre 2,1 % and 2,5 % ; R^2 entre 0,98 et 0,99	[61]
Gharagheizi <i>et al.</i>	443	1672	Test : RMSE = 0,22 log(cP) ; $R^2 = 0,854$	[39]
Hada <i>et al.</i>	23	23	Dataset entier : $R^2 = 0,609$	[40]
Han <i>et al.</i>	255	1731	CV : R^2 entre 0,7679 et 0,9113	[62]
Matsuda <i>et al.</i>	Non précisé dans l'article	341	Test : $R^2 = 0,6226$	[31]
Mirkhani et Gharagheizi	293	435	Test : $Q^2 = 0,8502$	[63]
Paduszynski et Domanska	1484	13470	Test : MSE = 0,0603 log(cP) ² ; $R^2 = 0,972$	[64]
Tochigi et Yamamoto	162	335	Train : R^2 entre 0,9127 et 0,9308	[26]
Yu <i>et al.</i>	42	42	CV : R^2 entre 0,9113 et 0,9353	[65]
Zhao <i>et al.</i>	89	1502	Test : MSE entre 0,025 et 0,187 log(cP) ² ; R^2 entre 0,800 et 0,930	[66]
Zhao <i>et al.</i>	25 IL purs, 4 mélanges binaires	154 pour les IL purs, 450 pour les mélanges binaires	Dataset entier : RAAD entre 0,08 % et 1,61 % pour les IL purs et entre 3,2 % et 8,2 % pour les mélanges binaires	[67]

Tableau 1.2 – Compilation de différentes études QSPR pour la viscosité. Les différents indicateurs de performances utilisés sont le coefficient de détermination R^2 , la racine de l'erreur quadratique moyenne (Root Mean Squared Error, RMSE), la moyenne des écarts en pourcentage, le pourcentage d'erreur moyen (Mean Percentage Error, MPE), le coefficient de corrélation Q^2 , l'erreur quadratique moyenne (Mean Squared Error, MSE), et l'écart absolu moyen relatif (Relative Average Absolute Deviation, RAAD).

Chapitre 1. Introduction aux solvants étudiés

Auteurs	Nombre de IL	Nombre de données	Performances rapportées	Réf.
Bini et al	33	66	LOO : R^2 entre 0.6317 et 0.9112	[33]
Cao <i>et al.</i>	35	364	Test : RMSE entre 0,544 et 1,380 mS/cm ; R^2 entre 0,5264 et 0,9703	[68]
Diaz-Rodriguez et al	3	108	6f-CV : $R^2 = 0,991$ et MPE = 6,42 %	[69]
Gharagheizi et al	54	977	Test : RMSE = 0,07 S/m ; $R^2=0,999$	[70]
Gharagheizi et al	54	1077	Test : RMSE = 0,2 mS/cm ; $R^2=0,994$	[71]
Matsuda <i>et al.</i>	Non précisé dans l'article	225	Test : $R^2 = 0,7664$	[31]
Tochigi and Yamamoto	79	150	Train : R^2 entre 0,9089 et 0,9745	[26]

Tableau 1.3 – Compilation de différentes études QSPR pour la conductivité. Les différents indicateurs de performances utilisés sont le coefficient de déterminant R^2 , la racine de l'erreur quadratique moyenne (Root Mean Squared Error, RMSE), et le pourcentage d'erreur moyen (Mean Percentage Error, MPE).

Considérons d'abord la pureté des liquides ioniques. Selon la voie de synthèse choisie, il est courant que des ions halogénures restent dissous dans le produit. Par exemple, à 20 °C, augmenter la quantité de Cl^- de 0,01 mol/kg à 2,2 mol/kg peut faire augmenter la viscosité du nitrate de 1-octyl-3-méthylimidazolium de 1 238 cP à 8 465 cP ($\Delta=7\ 227$ cP) [16]. Cependant, cette information n'est généralement disponible que dans les articles de recherche en synthèse organique. Dans les articles physico-chimiques analysant les propriétés des liquides ioniques, ce paramètre est souvent omis.

Les liquides ioniques sont généralement hygroscopiques. Une petite quantité d'eau peut être présente dans l'échantillon simplement à cause de son exposition à l'air ambiant [72]. Or, la quantité d'eau présente dans un liquide ionique peut fortement influencer ses propriétés. 1 % d'eau dissoute peut faire chuter la viscosité d'une valeur comprise entre 10 et 300 cP [73], et 0,01 % d'eau dissoute peut augmenter la conductivité de 1,2 mS/cm [74]. Heureusement, ce paramètre est souvent donné dans les articles de recherche, en synthèse comme en analyse physico-chimique.

La température de fusion n'est pas toujours facile à mesurer [47]. En effet, celle-ci dépend de la structure cristalline de la substance. Si cette substance est amorphe, on peut observer une transition vitreuse qui peut changer en fonction de la façon dont l'échantillon a été préparé. Cependant, expérimentalement, seule la calorimétrie différentielle à balayage (*Differential Scanning Calorimetry* en anglais, DSC) peut permettre de distinguer une transition vitreuse d'un point de fusion, et encore, avec difficulté car la différence entre ces deux changements de phase est mince [47, 75].

Enfin, la température de mesure peut fortement affecter la viscosité et la conductivité. Par exemple, faire varier la température de 1 °C peut faire augmenter la conductivité de

2 mS/cm [74], tandis que la viscosité du diméthylphosphate de diméthylimidazolium chute de 320,09 cP (20 °C) à 222,70 cP (25 °C) [76]. Certaines approches telles que les méthodes de contribution de groupe ou COSMO-RS peuvent le prendre en compte. Cependant on ne sait pas encore si ces approches conduisent à des améliorations concrètes. D'autres part, les modèles QSPR peuvent être construits à une température fixe.

1.2 Les électrolytes pour batteries Li-ion

1.2.1 Généralités

Une batterie, ou accumulateur électrique, est un appareil dont la fonction est de stocker de l'énergie électrique afin de pouvoir la restituer ultérieurement. Ceci est généralement fait en convertissant l'énergie électrique en énergie chimique (lors de la charge), et inversement (lors de la décharge) [77]. Une batterie est généralement constituée de plusieurs cellules électrochimiques. Ce sont ces cellules qui stockent et délivrent l'énergie. Dans le langage courant, le mot batterie est couramment utilisé pour désigner indifféremment l'appareil entier ou une seule cellule électrochimique. [77].

Les cellules électrochimiques sont constituées de trois éléments majeurs [77] : une anode, une cathode et un électrolyte (voir la figure 1.1). L'anode est l'électrode délivrant des électrons dans le circuit électrique. La cathode, quant à elle, est l'électrode acceptant les électrons provenant du circuit externe. Enfin, l'électrolyte est le milieu permettant la circulation des ions mais pas celle des électrons. Il est généralement constitué d'un ou plusieurs solvants et d'un sel [8]. Dans les batteries Li-ion, le cation circulant dans le milieu électrolytique est le Li^+ . Dans le cadre de cette thèse, on appellera électrolyte le solvant électrolytique utilisé pour dissoudre le sel.

Il existe différents types d'électrolytes [78]. Ils peuvent être de type liquide (organiques, IL, aqueux), solide (polymères solides, solides inorganiques) ou gel. Le choix du type d'électrolyte va dépendre de l'application de la batterie. Si l'on souhaite concevoir des batteries ayant d'excellentes propriétés mécaniques et étant utilisées dans les appareils de stockage d'énergie multifonctionnels, il sera plus intéressant de choisir un électrolyte de type solide. Pour les batteries demandant une haute densité d'énergie à destination d'appareils à forte consommation électrique (tels que des batteries pour ordinateurs portables ou pour voitures électriques), il faudra se tourner vers un électrolyte liquide. Enfin, les électrolytes de type gel constituent un compromis entre les deux précédents électrolytes décrits.

Dans le cadre de cette thèse, nous allons nous intéresser aux électrolytes liquides. Ils possèdent une haute conductivité ionique et sont capables de former un contact stable et homogène avec les électrodes [78]. Leur point faible est la sûreté de ces substances, souvent instables chimiquement et thermiquement. Ils se séparent en trois catégories distinctes : les électrolytes organiques, les électrolytes aqueux et les liquides ioniques. Ici, nous ne nous intéresserons qu'aux électrolytes liquides organiques.

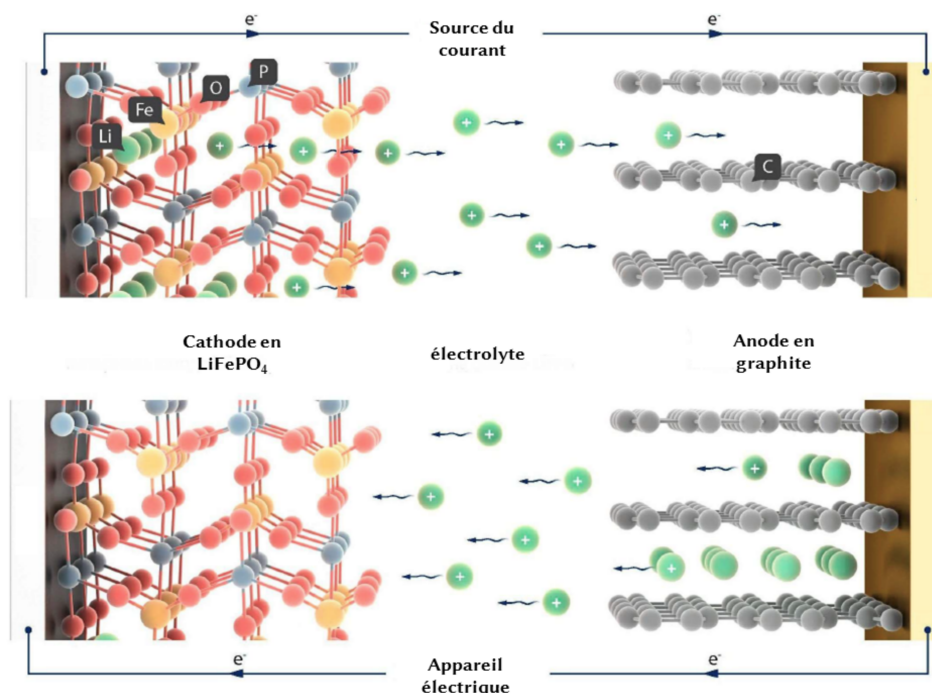


FIGURE 1.1 – Fonctionnement d’une cellule électrochimique. Lors de la charge (en haut sur la figure), les ions Li^+ migrent de la cathode vers l’anode en circulant à travers l’électrolyte. Lors de la décharge (en bas), c’est l’inverse : les ions migrent de l’anode vers la cathode. *Source : <http://liotech.ru/principles>*

Dans le cadre des batteries Li-ion, un bon électrolyte doit être un bon conducteur ionique, un bon isolant électronique, avoir une large fenêtre électrochimique, être inerte vis-à-vis des autres composants de la batterie, résister à diverses contraintes d’ordre chimique, thermodynamique, mécanique, et enfin être respectueux de l’environnement [8]. Il doit être capable de dissoudre des sels en concentration suffisante (c’est-à-dire posséder une constante diélectrique élevée), avoir une viscosité faible pour faciliter le transport des ions, et être liquide sur une large plage de températures. En pratique, le mélange de plusieurs solvants permet d’obtenir des caractéristiques difficiles à combiner avec un seul solvant, par exemple une grande fluidité et une haute constante diélectrique. Pour éviter la formation de dihydrogène, les solvants utilisés doivent être aprotiques [79, 80]. Enfin, ils doivent être polaires afin de permettre la solvataion des sels de lithium.

Les électrolytes les plus couramment employés sont le carbonate de propylène (PC), le carbonate d’éthylène (EC), le carbonate d’éthylméthyl (EMC), le carbonate de diéthyl (DEC) et le carbonate de diméthyl (DMC) [8]. Les sels de lithium qui leur sont généralement associés sont le perchlorate de lithium (LiClO_4), l’hexafluorophosphate de lithium (LiPF_6) et le bis(trifluorométhylsulfonyl)imide de lithium (LiTFSI). Il est difficile de trouver une bonne combinaison entre les sels et les solvants. En effet, les sels doivent être solubles, stables d’un point de vue électrochimique, et avoir un faible impact environnemental. Pour les solvants, il faut faire attention à diverses propriétés : la constante diélectrique, la viscosité, la température de

fusion, la température d'ébullition, et la sûreté du solvant en terme d'inflammabilité, de toxicité, et de stabilité thermique. Enfin, il faut que les deux soient compatibles avec les électrodes et les collecteurs de courant, et qu'ils permettent la formation d'une interface électrolyte-électrode (IEE) stable [78]. En pratique, les batteries Li-ion commerciales sont donc constituées de LiPF_6 , de carbonate d'éthylène et d'autres co-solvants tel que le carbonate de diméthyl [8, 81].

Les deux principaux problèmes de ces solvants sont leur stabilité (thermique et électrochimique) et leur sûreté. En théorie, les électrodes ne devraient subir aucun changement chimique pendant le fonctionnement de la batterie [8]. En pratique, la stabilité de l'électrode est cinétique et non statique, et elle est mise à mal par la nature hautement oxydante de l'anode, et par la nature fortement réductrice de la cathode. Ceci peut conduire à la formation de dendrites et de sels de lithium en suspension dans le solvant électrolytique, pouvant générer des dysfonctionnements et des problèmes de sécurité tels que des explosions et des court-circuits [8]. En ce qui concerne la stabilité thermique des batteries Li-ion, elle est comprise entre $-20\text{ }^\circ\text{C}$ et $50\text{ }^\circ\text{C}$. En dessous de $-20\text{ }^\circ\text{C}$, la capacité et la vitesse de délivrance de l'énergie sont en chute libre, mais le processus reste réversible si on place la batterie à $20\text{ }^\circ\text{C}$ ou plus. En revanche, au dessus de $50\text{ }^\circ\text{C}$, une détérioration irréversible de la batterie a lieu, ce qui crée des risques d'explosion de la batterie dus à la formation de PF_5 sous forme gazeuse.

Pour remédier à ces problèmes, diverses solutions ont été mises en œuvre : ajout d'additifs, de navettes redox (*redox shuttle* en anglais) pour stabiliser le voltage, de retardants et d'inhibiteurs thermiques pour augmenter la stabilité thermique, ou encore changer d'électrolytes. Cependant, à l'heure actuelle, les batteries commercialisées présentent le meilleur compromis.

1.2.2 Revue de la littérature pour la modélisation des électrolytes

Il existe de nombreux travaux concernant la modélisation des divers composants des cellules électrochimiques. Dans le cadre de cette thèse, nous allons uniquement nous intéresser aux travaux concernant l'IEE ainsi que ceux impliquant les électrolytes. Les études décrites dans cette section sont listées dans le tableau 1.4.

De nombreux travaux se sont intéressés à la modélisation de l'IEE. Les principales questions étudiées concernent la composition chimique, l'organisation de l'IEE et les réactions qui s'y déroulent. Les méthodes employées sont surtout la dynamique moléculaire [82], et l'AIMD [83–85]. En effet, les phénomènes se déroulant dans l'IEE sont essentiels pour rationaliser les performances d'une batterie. Par exemple, Park *et al.* [86] ont estimé l'affinité de 32 espèces potentiellement présentes dans des IEE pour le cation Li^+ . Ils en déduisent des recommandations sur les propriétés des molécules organiques qui sont susceptibles de perturber les comportements de la cellule électrochimique.

Les électrolytes bénéficient également de travaux de recherche dans les domaines de la chimie quantique et de la modélisation moléculaire. Certaines études s'intéressent à la stabilité chimique des électrolytes. Ces calculs emploient généralement la DFT. On peut citer Xing *et*

al. [87, 88] qui se sont intéressés au mécanisme de décomposition oxydative de PC, EC, DMC, DEC et EMC. Ou encore, ceux de Bedrov *et al.* [89] qui se sont intéressés au mécanisme de décomposition oxydative de l'EC. Les résultats identifient les produits de réaction, les cinétiques et les différents mécanismes réactionnels afférents. Dans le projet DEVEGA dans lequel s'inscrit cette partie de la thèse, des études similaires sont menées au laboratoire COBRA de l'université de Rouen, sous la direction de L. Joubert. D'autres études visent à estimer par le calcul certaines caractéristiques essentielles à l'ingénierie des électrolytes : le potentiel d'oxydation (qui est identifié en général au potentiel d'ionisation) et le potentiel de réduction (identifié à l'affinité électronique). Par exemple, Han *et al.* [90] ont estimé le potentiel d'oxydation de 108 composés. Il s'agit de la seule étude qui se compare à des données expérimentales. Mais les performances des modèles issus des calculs DFT (RMSE de 0,08 V) sont données sur les mêmes molécules que celles qui ont servi à construire le modèle. Les autres études sont corrélées à des valeurs expérimentales publiées et reposent sur un petit nombre de structures, à l'instar de celle de Ong *et al.* [91] qui se sont intéressés à la fenêtre électrochimique de 6 liquides ioniques. Une approximation supplémentaire est généralement faite, qui consiste à identifier l'affinité électronique à $-E_{LUMO}$ et le potentiel d'ionisation à $-E_{HOMO}$. Les potentiels calculés sont donc assez différents de ceux que l'on peut obtenir expérimentalement (et qui dépendent entre autres, de la nature de l'électrode).

Enfin, il existe, à notre connaissance, peu d'études concernant le criblage virtuel à grande échelle des électrolytes. Halls et Tasaki ont criblé 7381 électrolytes potentiels dérivés de l'EC [92]. Ce sont les seuls pour lesquels la méthode COSMOtherm [93] n'a pas été utilisée. Ils ont estimé l'affinité électronique et le potentiel d'ionisation, à l'aide de méthodes semi-empiriques et du module Material Studio de Pipeline Pilot [94]. Cette étude se focalise sur les moyens de prioriser les touches, mais les détails du calcul des potentiels sont laissés à la seule responsabilité de l'éditeur du logiciel Pipeline Pilot : en d'autres termes, il s'agit d'une boîte noire. Les auteurs notent que leur protocole de criblage a une vitesse de 20 s par structure. Les autres études de criblages proviennent de l'équipe de Korth, et utilisent l'approche COSMOtherm. Le logiciel COSMO est alimenté par des calculs quantiques (DFT, WFT ou semi-empirique) qui génèrent des descripteurs pour les modèles disponibles dans COSMOtherm. Les auteurs ont ainsi pu cribler de grandes chimiothèques contenant de 5 000 structures [95, 96] à plusieurs millions de structures [97]. Les performances de ces modèles sur des électrolytes sont données dans [96] sans toutefois qu'il soit possible de savoir si les composés testés faisaient partie des données utilisées pour l'entraînement des modèles. Ces performances sont reportées dans le tableau 1.4. Les moyens de calcul nécessaires à ce criblage sont vraisemblablement importants puisqu'ils ont nécessité un déploiement sur de grandes grilles de calculs [96, 98]. Ces calculs ont mené à la sélection, la synthèse et la caractérisation de 4 structures [95, 97], mais les résultats des calculs ne sont pas comparés rétrospectivement aux résultats expérimentaux. Toutefois, les électrolytes choisis possèdent des caractéristiques intéressantes pour le développement de super-condensateurs.

Auteurs	Composant modélisé	Type d'étude	Notes	Réf.
Bedrov <i>et al.</i>	électrolytes	MP2, DFT, et MD (ReaxFF)		[89]
Ganesh <i>et al.</i>	IEE	AIMD		[83]
Halls et Tassaki	électrolytes	PM3	criblage virtuel de 7 381 composés, pas comparé à l'expérience	[92]
Han <i>et al.</i>	électrolytes	DFT	RMSE(Eox)= 0,08 V, R ² (Eox)=0,98, R ² (IP)=0,84 (ce sont probablement les statistiques liées à l'ajustement du modèle aux données)	[90]
Korth	électrolytes	COSMOtherm	criblage virtuel de 11 000 molécules	[98]
Leung et Budzien	IEE	AIMD		[84]
Leung	IEE	AIMD		[85]
Ong <i>et al.</i>	électrolytes	MD, DFT		[91]
Park <i>et al.</i>	IEE/électrolytes	DFT		[86]
Schutter <i>et al.</i>	électrolytes	COSMOtherm	criblage virtuel de 5000 nitriles (super-condensateurs)	[95]
Schutter <i>et al.</i>	électrolytes	COSMOtherm	criblage virtuel de plus de 60 000 composés (super-condensateurs)	[97]
Xing <i>et al.</i>	électrolytes	DFT		[87]
Xing <i>et al.</i>	électrolytes	DFT		[88]
Xing <i>et al.</i>	IEE	MD	(super-condensateurs)	[82]

Tableau 1.4 – Compilation de différentes études QSPR pour la modélisation des divers composants d'une cellule électrochimique (IEE : interface électrode-électrolyte). Sauf précision contraire, les études concernent les batteries Li-ion.

1.3 En résumé

Les *liquides ioniques* sont des sels liquides en dessous de 100 °C qui sont de plus en plus populaires avec le développement de la chimie verte. Ils ont été modélisés à de nombreuses reprises à l'aide de la chémoinformatique. Cependant, les études menées souffrent de plusieurs problèmes. Certaines d'entre elles fixent l'anion à modéliser, ce qui revient à faire une modélisation classique de composés qui nécessiteraient pourtant une adaptation. La validation des modèles n'est pas toujours rigoureuse, notamment lorsque la température est utilisée comme descripteur : les performances des modèles ainsi préparés sont donc très optimistes et illustrent plus la capacité d'un modèle à faire une interpolation que sa capacité à faire une prédiction pour un nouveau liquide ionique. Les modèles basés sur les contributions de groupe sont très difficilement généralisables car les groupes sélectionnés sont assez volumineux pour les cations, et considèrent l'anion complet comme un seul groupe : un nouvel anion ne pourra donc pas être prédit. Enfin, la qualité des données est rarement prise en compte lors de la construction des jeux d'entraînement. Lorsque nous modéliserons les liquides ioniques, nous devons donc prendre plusieurs choses en compte :

- utiliser des descripteurs permettant de prédire par la suite des liquides ioniques constitués d'ions jamais vus dans le jeu d'entraînement,

- mettre en place une stratégie permettant de faire varier les cations et les anions dans le jeu d'entraînement,
- tester les modèles avec des liquides ioniques complètement nouveaux,
- veiller à la qualité des données sélectionnées pour la construction des jeux d'entraînement.

Enfin, il est important de noter que les modèles décrits dans les différents articles ne sont pas mis à disposition par les auteurs. Nous allons donc mettre les modèles développés en ligne.

Dans le cadre de cette thèse, nous appelons *électrolyte* le ou les solvants permettant la dissolution et la migration du sel de lithium lors du fonctionnement d'une batterie. C'est un élément crucial, puisqu'il détermine la vitesse à laquelle l'énergie va être délivrée. Ils sont couramment étudiés en chimie quantique et en dynamique moléculaire, mais sont très peu modélisés en chémoinformatique. Pourtant, les méthodes issues de la chémoinformatique ont l'avantage de pouvoir construire des modèles permettant d'obtenir directement une estimation de la propriété choisie, ce qui n'est que rarement le cas lors d'études quantiques ou de dynamique moléculaire. Ces méthodes permettent également de pouvoir faire des prédictions pour plusieurs propriétés simultanément. Dans le cadre de cette thèse, nous allons donc nous concentrer sur la construction de modèles permettant la modélisation des 6 propriétés suivantes : la conductivité, la constante diélectrique, le potentiel d'oxydation, la température de fusion, la température d'ébullition et la viscosité . Par la suite, nous implémenterons les modèles les plus prédictifs dans un outil facilement utilisable par tous.

Chapitre 2

Méthodes de modélisation utilisées

Nous avons modélisé les solvants par le biais du QSPR (*Quantitative Structure-Property Relationship*, relation quantitative structure-propriété). C'est un domaine visant à formuler une relation quantitative entre les structures chimiques d'un ensemble de molécules et une propriété de celles-ci. La structure de ladite molécule est généralement décrite avec des vecteurs de nombres appelés descripteurs. La mise en relation de la structure avec la propriété se fait à l'aide de méthodes d'apprentissage automatique. Enfin, il faut évaluer les paramètres statistiques mesurant les performances des modèles.

2.1 Descripteurs

Les descripteurs sont définis par Todeschini et Consonni [99] de la façon suivante :

« Les descripteurs moléculaires sont le résultat final d'une procédure mathématique et logique qui transforme l'information encodée à travers une représentation symbolique de la molécule en un nombre utile ou le résultat de quelque expérience standardisée. »

D'après cette définition, ils peuvent être séparés en deux groupes distincts : les mesures expérimentales (c'est le cas par exemple des propriétés physico-chimiques, des activités biologiques et des propriétés environnementales) et les descripteurs moléculaires théoriques, qui sont dérivés d'une représentation symbolique de la molécule.

Les descripteurs moléculaires théoriques peuvent eux aussi être séparés en divers groupes :

- descripteurs 1D, ou descripteurs constitutionnels (comptes d'atomes, nombre de groupes fonctionnels, masse molaire de la molécule, etc.),
- descripteurs 2D, basés sur la connectivité, la topologie de la molécule (descripteurs fragmentaux, indice de Wiener, indices de Zagreb, etc.)

- descripteurs 3D, basés sur la représentation tridimensionnelle de la molécule (surfaces, paramètres stériques, descripteurs quantiques, etc.)

Selon Todeschini et Consonni [99], pour qu'un descripteur puisse être considéré comme tel, il doit posséder les caractéristiques suivantes :

- il ne doit pas dépendre de l'étiquetage et de la numérotation des atomes,
- il doit être invariant par rapport à une rotation ou une translation de la molécule,
- il doit avoir une définition non ambiguë et calculable de façon algorithmique,
- et posséder des valeurs dans un intervalle de valeurs adéquat pour le jeu de molécules auquel il doit être appliqué.

Idéalement, un descripteur devrait avoir une interprétation structurale, une bonne corrélation avec au moins une propriété, ne pas avoir de corrélations triviales avec d'autres descripteurs moléculaires, avoir une évolution graduelle de ses valeurs lorsque la structure chimique est graduellement modifiée, ne pas être inclus dans la définition pour le cas des propriétés expérimentales, et ne pas être restreint à un ensemble trop restreint de structures [99]. Il devrait pouvoir faire une discrimination en fonction des isomères considérés, ne pas être inclus dans la définition d'autres descripteurs et permettre un décodage réversible de la molécule (c'est-à-dire pouvoir remonter à la molécule à partir des descripteurs).

Au cours de cette thèse, nous avons utilisé trois types de descripteurs différents : les fragments ISIDA, les descripteurs EED et les descripteurs MOE2D.

2.1.1 Descripteurs fragmentaux ISIDA

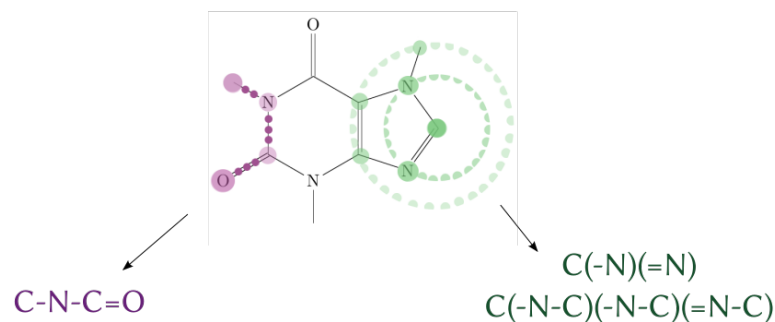
Les descripteurs ISIDA (*In Silico Design and Data Analysis*) sont des descripteurs 2D de type topologique ayant été développés au laboratoire de Chémoinformatique de Strasbourg [100–105]. Ils sont construits sur les données 2D de la molécule, c'est-à-dire uniquement les atomes, les charges formelles et leurs connexions, mais pas leur coordonnées dans l'espace. La conformation de la molécule n'est donc pas prise en compte ici. Ils se basent sur le décompte du nombre d'occurrences d'un fragment dans une molécule donnée. Ces fragments peuvent être de deux types (voir figure 2.1) :

- séquences (*sequences*, type I) : le fragmenteur se base sur des séquences d'atomes (A), de liaisons (B) ou d'atomes et de liaisons (AB) de taille fixée par le type de fragments ;
- atomes unis augmentés (*augmented atoms*, type II) : le fragmenteur part d'un atome donné ; à partir de ce point, le fragmenteur va déterminer toutes les séquences d'atomes, de liaisons ou d'atomes et de liaison de taille fixée par la fragmentation voulue.

Quel que soit le type choisi, la construction des fragments est définie par rapport à une taille fixe (c'est-à-dire un nombre d'atomes ou de liaisons invariables) ou par rapport à un intervalle donné $[n_{min}-n_{max}]$. On peut également demander au fragmenteur de réutiliser des descripteurs déjà calculés auparavant sur de nouvelles données. Ceci est utile lorsque l'on a

construit un modèle avec un jeu d'entraînement et que l'on souhaite l'appliquer à un jeu de validation.

Lorsque qu'on a calculé tous les descripteurs d'un ensemble de molécules, on peut ensuite compter le nombre d'occurrences de chaque fragment dans la molécule. Chaque molécule aura donc un vecteur de nombre la décrivant.



	C-N-C=O	...	C(-N)(=N)	C(-N-C)(-N-C)(=N-C)
mol 1	0	...	1	3
mol 2	2	...	6	0
...
mol n	1	...	0	1

FIGURE 2.1 – Principe de création des fragments ISIDA. Le fragmenteur se base sur le graphe de la structure de la molécule. Des fragments, sous forme de séquences (en violet) ou d'environnements autour d'un atome (en vert) sont énumérés et codés par une chaîne de caractère. Ce code enregistre des informations sur les atomes (A), les liaisons (B) ou les atomes et les liaisons (AB). Les fragments ont une taille maximale fixe, mesurée en distance topologique. Enfin, le nombre d'occurrences de chaque fragment est enregistré. Ce nombre est la valeur du descripteur. Le tableau des valeurs de tous les descripteurs pour toutes les molécules est enregistré dans une matrice de descripteurs illustrée dans la moitié basse de la figure.

2.1.2 Descripteurs d'effets électroniques

Les descripteurs d'effets électroniques (*Electronic effect descriptors*, EED) sont des descripteurs générés à l'aide de fonctions de la librairie ChemAxon [106–108]. Ils servent à caractériser l'impact global de l'environnement chimique sur la densité électronique d'un ou plusieurs atomes clés K dans une molécule (par exemple un atome porteur d'un atome d'hydrogène labile dans le cas du pKa). Ces descripteurs sont simplement conçus comme des contributions additives topologiques pondérées par la distance de chacun des N atomes de la molécule.

Les atomes clés sur lesquels les descripteurs EED sont calculés doivent être définis par l'utilisateur afin de correspondre aux atomes pertinents. Les atomes voisins de ces atomes clés

contribuent à différents termes de descripteurs proportionnellement à leurs propriétés (charges partielles, σ et π , polarisabilité, électronégativité, indice d'énergie nucléophile, indice d'énergie électrophile, indice de densité de charge, indice d'hybridation, et charges formelles) associées à chaque terme. La contribution d'un atome donné j dépend d'un facteur $\frac{1}{d}$ ou d'un facteur $\frac{1}{d^2}$, où d correspond à une distance topologique entre l'atome clé K et l'atome j . Certains termes de ces descripteurs EED découlent uniquement de contributeurs à travers une séquence de liaisons aromatiques (c'est-à-dire une alternance entre liaisons simples et liaisons doubles). Ils servent à capturer l'influence possible d'effets de résonance potentiels. Chaque atome clé dénombre 704 descripteurs EED.

2.1.3 Descripteurs MOE2D

Les descripteurs MOE2D sont développés par le *Chemical Computing Group* [109]. Ils proviennent du logiciel QuaSAR-descriptors de la librairie MOE (*Molecular Operating Environment*), qui permet le calcul d'une grande variété de descripteurs 2D et 3D. Dans le cadre de cette thèse, seuls les descripteurs 2D ont été utilisés.

Les descripteurs MOE2D peuvent être divisés en 7 catégories :

- propriétés physiques (par exemple la masse molaire moléculaire, la charge totale de la molécule, la somme des polarisabilités atomiques, le logarithme du coefficient de partage octanol/eau, ou encore la réfractivité moléculaire) [110–113],
- zones de surfaces subdivisées (*Subdivided Surface Areas*) [112],
- comptes d'atomes et de liaisons (tels que le nombre d'atomes aromatiques, le nombre d'atomes de carbone, de fluor, d'hydrogène, d'azote, ou encore le nombre de liaisons simples, doubles, triples ou aromatiques, pour n'en citer que quelques-uns) [109],
- indices de connectivité de Kier et Hall et indices de forme Kappa (*Kier&Hall Connectivity and Kappa Shape Indices*) [114, 115],
- descripteurs basés sur les matrices de distance et d'adjacence (tels que la valeur la plus grande dans la matrice d'adjacence ou l'indice topologique de connectivité de Balaban) [116–119],
- descripteurs basés sur des éléments pharmacophoriques (atomes acides, basiques, hydrophobes, accepteurs ou donneurs de liaisons hydrogène, surfaces de van der Waals des atomes acides, basiques, etc.) [109],
- descripteurs de charge partielle [120] (par exemple la charge partielle positive relative, les surfaces de van der Waals négatives totales ou encore les surfaces de van der Waals polaires totales).

Si on regroupe tous les descripteurs MOE2D générés, nous obtenons 186 descripteurs au total.

2.2 Méthodes d'apprentissage automatiques

Les méthodes d'apprentissage automatique permettent de faire le lien entre la propriété à modéliser et les descripteurs utilisés pour décrire les molécules. Il en existe une grande variété, en fonction de ce que l'on cherche à faire : régressions linéaires simples, regroupement de molécules ou classification.

Au cours de cette thèse, nous avons utilisé trois méthodes : la SVR, la RR et la TRR.

2.2.1 Régression à vecteurs supports

La SVR (*Support Vector Regression*, régression à vecteurs supports en français) est une méthode développée par Drucker *et al.* en 1997 [121], qui s'inspire de la SVM (*Support Vector Machines*, machine à vecteurs de support ou séparateurs à vaste marge en français), méthode développée par Vapnik [122].

Pour effectuer une prédiction, le modèle SVR fait la somme pondérée de mesures de similarités entre certains objets du jeu d'entraînement (appelés vecteurs de support) et la molécule à prédire. L'entraînement consiste à ajuster ces poids pour minimiser une fonction de perte. Initialement, chaque molécule du jeu de données est potentiellement un vecteur support, mais la procédure d'optimisation conduit à mettre à zéro la plupart des poids : les molécules correspondantes sont donc retirées de l'ensemble des vecteurs supports. Le nombre de vecteurs supports conservés est intimement liée au terme de régularisation de la fonction de perte dont l'importance, C , reste à déterminer. Par ailleurs cette fonction de perte est ϵ -sensible. Le terme de cette fonction qui correspond à la qualité de l'ajustement augmente linéairement avec l'erreur d'ajustement du modèle, mais s'annule si cette erreur est inférieure à ϵ , une valeur seuil.

La méthode dépend donc de deux paramètres : un terme de régularisation C et ϵ . La valeur de ϵ reflète un niveau de bruit estimé dans les données expérimentales. Il est fixé *a priori* pour chaque propriété. Le paramètre C quant à lui équilibre l'importance accordée à l'erreur de l'ajustement du modèle, en comparaison de la complexité de ce dernier mesurée par le carré de la norme euclidienne du vecteur de poids du modèle. Plus le paramètre C est élevé, et plus le modèle essaiera de coller aux données : c'est le phénomène de sur-apprentissage. Un C trop faible conduira au contraire à un modèle simplifié à l'extrême, produisant toujours la même valeur quelle que soit la molécule considérée. Il s'agit d'un phénomène de sous-apprentissage. Ce paramètre est à optimiser par l'utilisateur. Dans le cadre de cette thèse, nous utilisons le logiciel LibSVM (version 3.17, 5) [123] pour réaliser les calculs, et nous avons systématiquement utilisé des noyaux linéaires.

2.2.2 Apprentissage semi-supervisé et transduction

En apprentissage supervisé, les données qui servent à la construction du modèle sont des données pour lesquelles la propriété modélisée est connue (voir figure 2.2). On parlera par la suite de données étiquetées pour les désigner. C'est le type d'apprentissage couramment utilisé dans le domaine du QSPR. On peut également choisir d'apprendre à partir de données non étiquetées (c'est-à-dire de données pour lesquelles la propriété à modéliser n'est pas connue) : c'est l'apprentissage non supervisé. Les méthodes utilisant à la fois des données étiquetées et des données non étiquetées appartiennent au domaine de l'apprentissage semi-supervisé [124–126]. La notion d'apprentissage semi-supervisé est souvent couplée avec la notion de transduction [124–126].

La transduction, illustrée en figure 2.3, consiste en un processus de modélisation qui utilise à la fois les molécules étiquetées et un ensemble pré-déterminé de molécules dont on cherche à déterminer les étiquettes. La construction du modèle débouche sur l'estimation des étiquettes pour ces dernières. Le modèle lui-même ne se généralise pas nécessairement à des molécules additionnelles par rapport aux données utilisées par le modèle transductif : il peut même être impossible pour un tel modèle de proposer une étiquette pour ces molécules supplémentaires. Ce processus est différent de celui couramment utilisé en QSPR, le processus d'induction-déduction, qui consiste à apprendre en construisant un modèle potentiellement généralisable au préalable (phase d'induction), puis à utiliser le modèle généré pour obtenir les prédictions sur les composés souhaités (phase de déduction) [10].

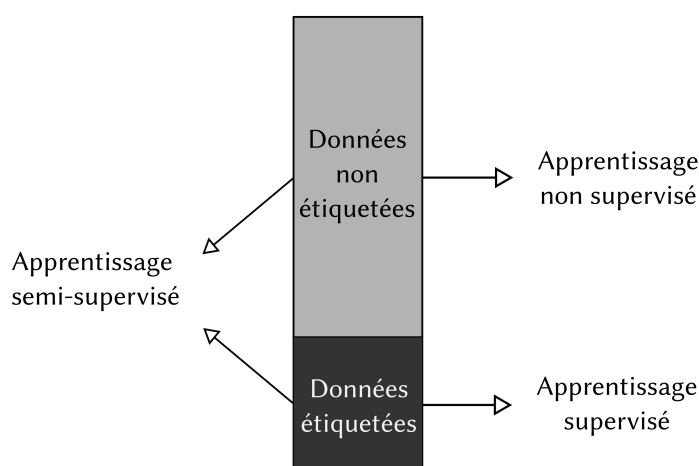


FIGURE 2.2 – Principe de l'apprentissage semi-supervisé. C'est un mode d'apprentissage qui utilise à la fois des données étiquetées et des données non étiquetées.

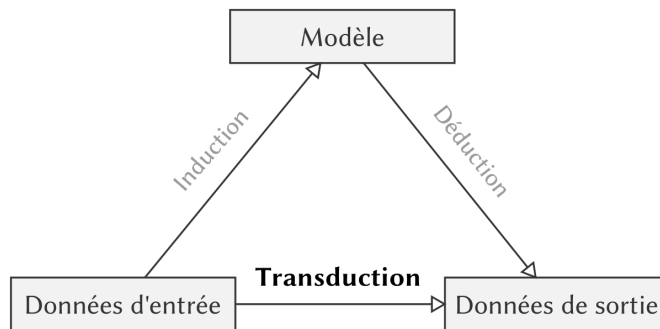


FIGURE 2.3 – Principe de la transduction. Le but ici est de créer un modèle adapté aux données sur lesquelles on veut appliquer ledit modèle. Ce principe diffère du schéma classique d'induction-dédution (trouver un modèle général à partir des données d'entrée, puis appliquer ce modèle à des données de sortie) utilisé généralement dans le domaine du QSPR.

Si elles sont assez connues dans le domaine de la fouille de données [126], les méthodes d'apprentissage semi-supervisé ne sont pas très courantes en chémoinformatique. Elles sont majoritairement utilisées, à notre connaissance, pour de la classification. Li *et al.* [127] ont adapté l'algorithme COREG [128], qui fonctionne de la façon suivante : un ensemble de modèles de classification de base est construit sur un jeu de données étiquetées. Des données non étiquetées sont ensuite ajoutées après avoir été étiquetées par le modèle. Les modèles de base sont ensuite entraînés à nouveau en prenant en compte l'ensemble de données fraîchement étiquetées. Les données non étiquetées sont ajoutées en fonction de la confiance accordée aux étiquettes estimées. La confiance est mesurée au niveau de la concordance qu'il y a entre les différents modèles de base. Cette approche a été appliquée sur un jeu de données de mutagenicité publié par Young *et al.* [129]. Ning *et al.* [130] ont testé trois méthodes différentes. La première est une méthode semi-supervisée basée sur de la propagation d'étiquettes : à l'aide d'un graphe de voisins les plus proches (*nearest neighbour graph*), des étiquettes pour les données non étiquetées sont estimées à partir de celles qui sont déjà étiquetées. Une SVM est ensuite entraînée en utilisant des poids permettant de différencier l'importance des données étiquetées et celle des données dont les étiquettes sont extrapolées. Les deux autres méthodes sont une méthode d'apprentissage par tâches multiples (*Multi-Task Learning*, MTL) et une méthode d'ordonnancement multiple (*Multi-Ranking*, MR). Le but de ce travail était de comparer ces méthodes à la chémogénomique. La méthode semi-supervisée, la technique de MTL et

l'approche MR surpassent toutes l'approche chémogénomique. Ce résultat a été confirmé par Horvath et al [131]. Levatić *et al.* ont publié deux articles reliés au semi-supervisé. Dans le premier article [132], quatre méthodes de classification semi-supervisées sont testées : ce sont les algorithmes YATSI [133], Co-FTF [134], LLGC [135] et la SVM transductive [136]. Ces méthodes sont comparées à des méthodes de classification supervisée. Pour ces tests, trois jeux de données sont utilisés : les lignes cellulaires cancéreuses chez l'homme du département du développement de traitements thérapeutiques du *National Cancer Institute* (NCI) [137], la mutagenicité [138] et les composés organoleptiques liés à l'odeur de musc [139]. Dans cette étude, la méthode semi-supervisée fonctionne généralement mieux que les autres algorithmes d'apprentissage supervisé, bien que ça ne soit pas systématique. Dans le second article [140], les auteurs ont travaillé sur les propriétés cytostatiques de composés potentiellement efficaces contre le paludisme. Ils ont appliqué une procédure semi-supervisée basée sur des arbres de clustering prédictifs (*Predictive Clustering Tree*, PCT) pour de la régression multi-cibles. Les auteurs travaillent ici avec une forêt d'arbres de régression. Ils commencent par construire un premier modèle avec les données étiquetées, et font une prédiction des données non étiquetées. Ils ajoutent les prédictions les plus fiables à leur jeu d'entraînement, puis ils reconstruisent un modèle et ainsi de suite jusqu'à ce qu'on ne puisse plus ajouter de données au jeu d'entraînement. Ils ont observé que l'addition de molécules non étiquetées provenant de la base de données ChEMBL [141] a permis d'améliorer les performances des modèles comparées à celles des modèles construits uniquement avec les données étiquetées. Enfin, Kondratovitch *et al.* [142] se sont intéressés à la SVM transductive pour la classification. Ici, l'étude est spécialement conçue pour le jeu de validation, qui contient les données non étiquetées. La qualité des estimations est meilleure que celle obtenue avec un algorithme supervisé équivalent, en particulier pour les petits jeux de données non équilibrés.

2.2.3 Régression ridge transductive

Dans le cadre de cette thèse, nous nous sommes intéressée à la Régression Ridge Transductive (*Transductive Ridge Regression*, TRR) [143]. C'est la version transductive de la régression ridge (*Ridge Regression*, RR), autrement appelée régression de Tikhonov [144, 145].

Les modèles construits par la méthode RR sont relativement faciles à interpréter : chaque descripteur utilisé pour décrire l'ensemble des molécules étudiées est pondéré en fonction de son importance. Comparée à d'autres méthodes de régression, la RR a pour avantage de pouvoir travailler avec des descripteurs colinéaires sans qu'il n'y ait de problème lors du calcul du poids des descripteurs.

L'algorithme de la RR minimise $\epsilon_{\mathbf{Y},\mathbf{X}}(\mathbf{W})$ (équation 2.1), une version modifiée de la fonction de perte de la régression multilinéaire qui fait une régularisation des poids de la régression linéaire :

$$\epsilon_{\mathbf{Y},\mathbf{X}}(\mathbf{W}) = g\|\mathbf{Y} - \mathbf{X} \cdot \mathbf{W}\|^2 + \|\mathbf{W}\|^2 \quad (2.1)$$

Avec :

- \mathbf{Y} un vecteur colonne contenant les propriétés de N données,
- \mathbf{X} une matrice de D colonnes et de N lignes contenant les descripteurs de chacune des données,
- \mathbf{W} un vecteur colonne qui contient le poids attribué à chacun des D descripteurs utilisés au cours de la régression linéaire,
- g le paramètre de régularisation.

La forme de la fonction de perte 2.1 est inhabituelle car le paramètre de régularisation est mis en facteur de l'erreur empirique plutôt que sur la contrainte imposée aux poids du modèle. La fonction de perte usuelle est retrouvée si on multiplie les deux membres de l'égalité 2.1 par $\frac{1}{\lambda}$, puis qu'on substitue $\frac{1}{\lambda}$ par g , λ étant alors le paramètre de Tikhonov.

Le terme de régulation permet de favoriser des modèles utilisant les quelques descripteurs les plus utiles pour reproduire le jeu d'entraînement. Le paramètre g doit être optimisé. Plus g sera grand, et plus la solution sera proche de la solution multilinéaire classique, qui souffre souvent de sur-apprentissage. Un g petit, au contraire, fera tendre les poids de la régression vers zéro : on sera dans un cas de sous-apprentissage.

La transduction exploite des données non étiquetées dans la construction du modèle. La TRR inclut donc une méthode pour étiqueter des molécules à partir des seules molécules étiquetées du jeu d'entraînement. Dans notre cas, nous utilisons un algorithme de rNN (voir figure 2.4), qui est assez proche du kNN excepté que les voisins les plus proches doivent être compris dans un rayon donné : la propriété d'une molécule est estimée par la moyenne des étiquettes associées aux molécules voisines pondérées par leur similarité à la molécule cible jusqu'à un rayon r fixé. Ce rayon est fixé au premier pourcentile des similarités observées dans les données. Dans notre cas, la similarité entre deux molécules est calculée à l'aide du coefficient de Tanimoto.

Une fois que cela est fait, un modèle est construit en utilisant d'une part les données étiquetées, et d'autre part les données non étiquetées estimées par la rNN. Cependant, il semble préférable de séparer les contributions des données estimées par rNN des autres puisque ce ne sont pas des étiquettes réelles qui leur sont attribuées, mais seulement une estimation.

Un nouveau terme est introduit dans la fonction de perte pour prendre en compte cette première estimation des données non étiquetées :

$$\epsilon_{\mathbf{Y},\mathbf{X},\mathbf{Y}_p,\mathbf{X}_p}(\mathbf{W}) = g\|\mathbf{Y} - \mathbf{X} \cdot \mathbf{W}\|^2 + gp\|\mathbf{Y}_p - \mathbf{X}_p \cdot \mathbf{W}\|^2 + \|\mathbf{W}\|^2 \quad (2.2)$$

Avec :

- \mathbf{Y} et \mathbf{Y}_p un vecteur colonne contenant les propriétés de N données pour les données étiquetées et non étiquetées respectivement,

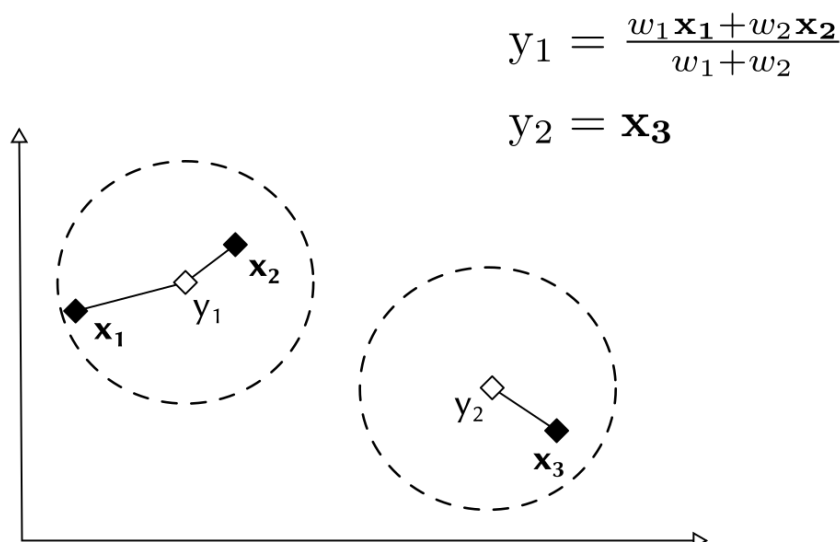


FIGURE 2.4 – Fonctionnement de l’algorithme rNN. La propriété d’une molécule y est estimée par la moyenne pondérée des étiquettes \mathbf{x}_i associées aux molécules similaires x_i jusqu’à un rayon de similarité r . La similarité est mesurée, par exemple, par le coefficient de Tanimoto entre la molécule y et une molécule x_i . Les poids w_i sont identifiés au coefficient de similarité.

- \mathbf{X} et \mathbf{X}_p les matrices de descripteurs pour les données étiquetées et non étiquetées respectivement,
- \mathbf{W} un vecteur colonne qui contient le poids attribué à chacun des D descripteurs utilisés au cours de la régression linéaire,
- g et gp les paramètres à optimiser pour les données étiquetées et non étiquetées respectivement,

Dans ce formalisme, le sur-apprentissage apparaît quand g est grand et le sous-apprentissage quand il est petit. Il n’y a pas de transduction si $gp = 0$. Sinon, gp permet de régler l’équilibre entre les erreurs du jeu d’entraînement et les erreurs sur les valeurs estimées par rNN. Un gp élevé par rapport à g signifiera que l’on accorde plus d’importance aux étiquettes de la rNN qu’aux étiquettes réelles. Au contraire, une petite valeur de gp par rapport à g indiquera que le terme de transduction doit être considéré comme un terme de correction de la RR.

Pour la RR, il faut optimiser le paramètre g relié aux données étiquetées. Pour la TRR, il faut également optimiser le paramètre gp relié aux données non étiquetées. Il va donc falloir mettre en place une stratégie pour les optimiser.

2.3 Domaine d'applicabilité

Le domaine d'applicabilité d'un modèle QSPR est la zone théorique de l'espace chimique définie par les données du jeu d'entraînement et les descripteurs du modèle dans laquelle le modèle peut être raisonnablement appliqué [146, 147]. Si on applique ce modèle sur une nouvelle molécule faisant partie du domaine d'applicabilité, la prédiction obtenue pour celle-ci pourra être considérée comme fiable, ce qui ne sera pas le cas si la molécule considérée est hors du domaine d'applicabilité. Il existe diverses manières de définir un domaine d'applicabilité. Dans le cadre de cette thèse, nous avons utilisé deux types de domaines d'applicabilité : le contrôle par fragments et la boîte bornée [147].

Le contrôle par fragments (*fragment control* en anglais) est un domaine d'applicabilité qui est défini pour les descripteurs ISIDA. Il consiste à contrôler l'apparition de nouveaux fragments dans une molécule, par rapport à ceux définis pour le jeu d'entraînement. Si un fragment exclusif à la molécule est détecté par rapport à ceux énumérés sur le jeu d'entraînement, la molécule est hors du domaine d'applicabilité et les estimations du modèles ne doivent pas être prises en compte.

La boîte bornée (*bounding box en anglais*) consiste à comparer, pour une molécule et pour chaque descripteur, si la valeur du descripteur se situe entre une valeur minimale et une valeur maximale déterminées sur le jeu d'entraînement. Si ce n'est pas le cas, alors la molécule est hors du domaine d'applicabilité et la valeur estimée par le modèle QSPR doit être ignorée. Ce domaine d'applicabilité est plus strict que le précédent.

En pratique, lorsque l'on parlera de boîte bornée, on considérera que le contrôle par fragments a aussi été appliqué. De plus, le principe de boîte bornée aura été appliqué aux descripteurs, mais également aux valeurs de propriété mesurée.

2.4 Validation croisée

Pour tester la précision de nos modèles, nous avons employé la technique de validation croisée (*k-fold cross-validation, k-CV*). La figure 2.5 illustre le fonctionnement d'une CV pour $k = 5$. Le jeu de données est d'abord mélangé, puis divisé en 5 paquets. Un des paquets (en gris foncé sur la figure 2.5) est mis de côté, et un modèle est entraîné sur les 4 autres. Une fois le modèle construit sur ces 4 paquets, on l'applique sur celui mis de côté : il fait office de jeu de validation. Cette procédure est ensuite répétée de façon à ce que tous les paquets construits au départ servent de jeu de validation. À la fin de la procédure, 5 modèles ont été construits, l'ensemble des données a été utilisé à la fois pour l'entraînement et la validation et chaque donnée a été prédite une fois seulement.

Il est possible d'appliquer plusieurs fois la procédure de validation croisée sur un même jeu de données. Pour cela, il suffit de veiller à ce que chaque étape de partage du jeu de données en k paquets génère des paquets différents. On notera $n \times k$ -CV une validation croisée en k paquets

qui est réitérée n fois. De même, la validation peut être interne, c'est-à-dire être utilisée pour l'optimisation du modèle, ou bien externe. Dans le cas d'une validation externe, les données de validation sont isolées du processus de construction et d'optimisation des modèles. La validation externe cherche donc à reproduire fidèlement les conditions de fonctionnement du modèle, tel qu'il est utilisé en production. Aussi, la standardisation des structures moléculaires et le calcul des descripteurs moléculaires des molécules tests sont isolés : la chaîne entière conduisant à l'estimation d'une propriété pour une structure moléculaire complètement nouvelle est testée.

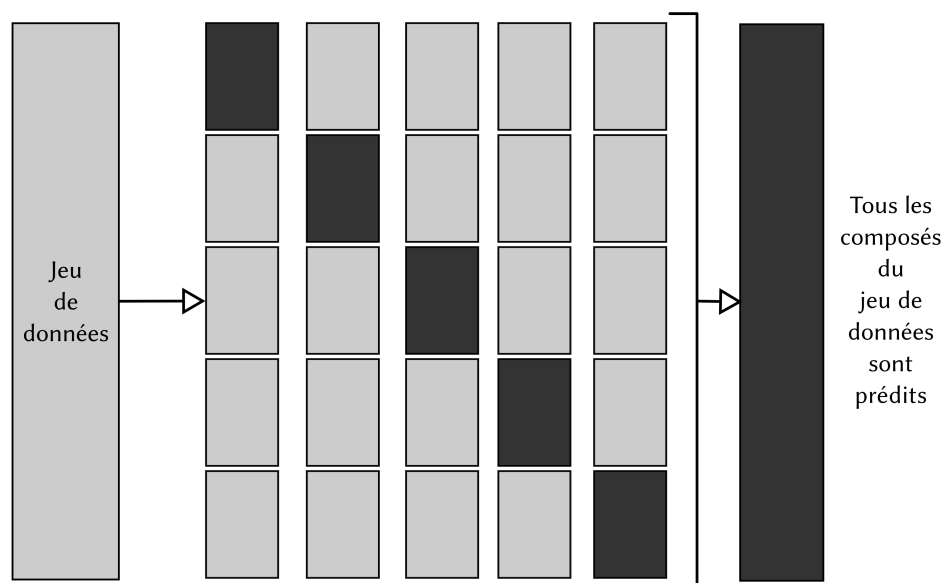


FIGURE 2.5 – Principe de fonctionnement d'une validation croisée. On découpe un jeu de données en k parties égales (ici, $k = 5$). On met un paquet de côté (en gris foncé), et on utilise les 4 autres en tant que jeu d'entraînement. On construit un modèle, et on le teste sur le dernier paquet. On répète l'opération pour que chaque paquet soit utilisé à un moment donné en paquet de test. À la fin, on a construit 5 modèles différents, et l'ensemble des composés a été testé une fois.

2.5 Paramètres statistiques de validation et de vérification des modèles

Pour évaluer la qualité de nos modèles, nous avons utilisé la RMSE, la MAE et le coefficient de détermination R^2 . Notons i la i -ème donnée du jeu de données, N le nombre total de données, $y_{exp,i}$ les valeurs expérimentales, $y_{pred,i}$ les valeurs prédites et y_{moy} la valeur moyenne calculée sur l'ensemble des valeurs expérimentales. Les formules de la RMSE, de la MAE et du R^2 peuvent être définies ainsi :

$$MAE = \frac{\sum_{i=1}^N |y_{pred,i} - y_{exp,i}|}{N} \quad (2.3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_{pred,i} - y_{exp,i})^2}{N}} \quad (2.4)$$

$$R^2 = \frac{\sum_{i=1}^N (y_{pred,i} - y_{moy})^2}{\sum_{i=1}^N (y_{exp,i} - y_{moy})^2} = 1 - \frac{\sum_{i=1}^N (y_{exp,i} - y_{pred,i})^2}{\sum_{i=1}^N (y_{exp,i} - y_{moy})^2}, \in [0, 1] \quad (2.5)$$

La MAE (*Mean Absolute Error*, erreur absolue moyenne en français, voir équation 2.3) correspond à la moyenne arithmétique des valeurs absolues des écarts. Elle fut développée par Sir Eddington [148]. Si on compare deux MAE, celle qui correspondra au meilleur modèle sera celle pour laquelle la MAE est la plus basse. Elle a l'avantage d'être insensible aux très grandes erreurs d'estimation.

La RMSE (*Root Mean Squared Error*, racine de l'erreur quadratique moyenne, voir équation 2.4) correspond à la racine carrée de la moyenne arithmétique des carrés des écarts entre les prévisions et les observations. Cette mesure provient de la découverte de la régression linéaire par K. Pearson [149, 150] à la suite des travaux de F. Gallon [151]. De nombreux algorithmes de régression sont fondés sur la minimisation de cette quantité dans le cadre d'une régression simple ou multiple. Si l'on compare deux RMSE, celle liée au modèle le plus performant sera la RMSE la plus basse. Enfin, il est important de noter que cet estimateur est plus sensible aux points aberrants que la MAE.

Le coefficient de détermination, ou R^2 (équation 2.5), correspond à la proportion de la variance totale de la propriété qui est expliquée par le modèle. Cette équation provient également des travaux de K. Pearson [149, 150]. Il permet de déterminer à quel point le modèle est adapté pour décrire la distribution des points. Il varie généralement entre 0 et 1. Si le R^2 est inférieur à zéro, attribuer la moyenne à chaque point donnerait un meilleur modèle que celui évalué : le modèle n'explique pas du tout la distribution des points. S'il est égal à un, le modèle est parfait.

2.6 Détection des points aberrants

Les points aberrants (*outlier* en anglais) sont, dans un jeu de données, les points qui apparaissent comme hors norme [152]. Ces données sont issues de situations échappant aux conditions contrôlées d'une expérience et devraient pour cette raison être écartées des analyses des données. La détection et le traitement adéquat de ces points aberrants sont cruciaux pour la construction de modèles fiables. Toutefois, ces définitions des points aberrants sont ambiguës : il est impossible d'en faire une traduction formelle unique.

Il existe plusieurs raisons pour lesquelles un point peut être considéré comme étant un point aberrant. Ces raisons peuvent découler de problèmes expérimentaux ou de problèmes de modélisation [153]. Parmi les problèmes expérimentaux, on peut citer une erreur dans la mesure, des conditions expérimentales qui ne sont pas respectées, ou encore la manifestation d'éléments provoquant une perturbation dans la mesure de la propriété (par exemple la réaction du composé analysé).

Du côté des problèmes de modélisation, on peut être confronté à un problème de standardisation. On peut également avoir une seule structure d'un type donné (par exemple, une sulfone isolée au milieu d'esters). Enfin, on peut être en présence d'une falaise d'activité (*activity cliff* en anglais), c'est-à-dire en présence d'une paire de structures chimiques similaires dont l'activité (ou la propriété) diffère fortement [154].

2.7 Comparaison des modèles

Pour comparer nos modèles, nous avons utilisé le test de Student apparié. Il est utilisé lorsque l'on souhaite comparer deux moyennes observées reliées à deux groupes d'échantillons qui ont un lien entre eux. Ce test permet de déterminer si les deux moyennes comparés sont significativement différents d'un point de vue statistique. Pour comparer les moyennes de deux séries appariées, on calcule tout d'abord la différence des deux mesures pour chaque paire.

Soit d la série des valeurs correspondant aux différences des mesures entre les paires de valeurs. La moyenne de la différence d est comparée à la valeur 0. Si la moyenne sur les distances est significativement différente de 0, alors on conclut à la différence entre les deux séries appariées.

La valeur t de Student est donnée par la formule :

$$t = \frac{m}{Sd} \times \sqrt{n} \quad (2.6)$$

Avec m et s qui représentent la moyenne et l'écart-type de la différence d . n est la taille de la série d .

Pour savoir si la différence est significative, on compare t à une valeur critique t_c . Les valeurs critiques dépendent d'un seuil de risque α et du nombre de degré de liberté $n - 1$, où n est le nombre de paires. Les valeurs critiques sont tabulées.

2.8 En résumé

Nous avons décrit les différents outils utilisés dans le cadre de cette thèse. Les modèles ont été construits en utilisant les méthodes « classiques » ϵ -SVR à noyaux linéaires (noté SVR par la suite) et RR. Nous avons également décidé d'étudier la TRR, algorithme qui devrait nous permettre de tirer parti des données non étiquetées à notre disposition, données qui

pourraient être nombreuses pour les solvants étudiés ici. En effet, certains de nos jeux de données concernant les liquides ioniques et les électrolytes sont petits et structurellement divers, ce qui peut poser un problème pour développer un modèle QSPR performant. La transduction pourrait donc potentiellement aider à améliorer les modèles. Cependant, cette méthode n'a encore jamais été étudiée en chémoinformatique, et en pratique il n'existe pas de recommandations pour l'optimisation des paramètres g et gp de la méthode, ni sur la quantité de données non étiquetées à utiliser. Pour savoir si cette méthode est adaptée à la modélisation des solvants, nous allons donc commencer par faire une étude méthodologique de celle-ci.

Chapitre 3

Étude méthodologique de la TRR

Avant de modéliser les solvants qui nous intéressent, nous avons commencé par mener une étude méthodologique. L'objectif ici est de comprendre comment fonctionne la TRR et comment l'utiliser au mieux afin de tirer parti des données non étiquetées. Nous avons d'abord implémenté la méthode, puis nous avons étudié comment optimiser ses paramètres g et gp . Après cela, nous nous sommes intéressée à l'impact de la taille relative du jeu d'entraînement par rapport au jeu de données non étiquetées. Enfin, nous avons fait une étude pour essayer de comprendre les cas problématiques.

3.1 Implémentation de la TRR

Afin de pouvoir contrôler entièrement le fonctionnement de la méthode TRR, nous l'avons implémenté dans un logiciel fonctionnant en ligne de commandes : le TRR-software. Ce logiciel est codé en Free Pascal [155, 156]. Nous avons implémenté la régression ridge, ainsi que l'algorithme TRR proposé par Cortès et Mohri [143]. Les paramètres g et gp peuvent être optimisés automatiquement dans un intervalle de valeurs définies par l'utilisateur. Cette optimisation se fait à l'aide de la méthode du nombre d'or (*golden section* en anglais) [157]. Enfin, il est possible de faire une $n \times k$ -CV.

3.2 Effet transductif

Pour pouvoir comparer les performances de la RR et de la TRR, nous avons défini l'effet transductif (*Transductive Effect*, TE). Cette valeur, exprimée en pourcentage, compare la RMSE du modèle RR non transductif ($RMSE_{RR}$) avec la RMSE du modèle transductif correspondant ($RMSE_{TRR}$). On considère ici que le modèle TRR est comparable au modèle RR si le paramètre g pour construire le modèle TRR est le même que celui qui a été utilisé pour le modèle RR. L'effet transductif peut être formulé de la façon suivante :

$$TE = 100 \times \frac{RMSE_{RR} - RMSE_{TRR}}{RMSE_{RR}} \quad (3.1)$$

Si l'effet transductif est positif, cela signifie que la $RMSE_{TRR}$ est plus basse que la $RMSE_{RR}$. Étant donné que le meilleur modèle est celui qui a la RMSE la plus basse, un effet transductif positif signifie que les performances du modèle TRR sont meilleures que celles du modèle RR. Par conséquent, l'effet transductif correspond au pourcentage d'amélioration (ou de détérioration) de la qualité des prédictions lorsque l'on applique la TRR au lieu de la RR.

3.3 Jeux de données utilisés

Pour cette étude méthodologique, trois jeux de données ont été utilisés. Le tableau 3.1 résume les informations importantes à connaître sur ces jeux de données.

Le jeu de la solubilité aqueuse (LogS) contient 1 635 données publiées précédemment [100]. Ce jeu, créé à partir des références [158–161], a notamment été utilisé lors de la *2nd Chemoinformatics Strasbourg Summer School* [162]. L'étendue des valeurs couverte par ce jeu de données est de -11,62 à 1,58 log(M). Chaque composé est représenté par 437 descripteurs fragmentaux ISIDA de type IAB(2-4). Ceci correspond en pratique à des séquences d'atomes et de liaisons comprenant entre 2 et 4 atomes.

Le jeu de données de la constante d'acidité (pKa) contient 924 molécules extraites de la base de données PHYSPROP [163]. Les pKa relevés sont compris entre -5,16 et 2,51 unités logarithmiques. Les 704 descripteurs utilisés sont des descripteurs EED (electronic effects descriptors) [164].

Le jeu de données A2AR contient 767 ligands du récepteur adénosine A2A. Ce jeu de données a été présenté à la *3rd Chemoinformatics Strasbourg Summer School* [165]. Ces données ont été sélectionnées à partir de trois sources distinctes : IUPHAR-DB [166], ChEMBL [141] et PubChem BioAssay [167]. La propriété modélisée ici est l'affinité (pKi) de ces composés pour le récepteur A2A. L'étendue des pKi relevés pour ce jeu de données est située entre 3,53 et 9,92 unités logarithmiques. Deux jeux de descripteurs distincts ont été utilisés. Nous avons ainsi un jeu de données noté A2AR ISIDA qui contient 483 descripteurs fragmentaux ISIDA de type IAB(2-5) (séquences d'atomes et de liaisons comprenant entre 2 et 5 atomes), et un jeu de données A2AR MOE2D comprenant 186 descripteurs MOE2D.

3.4 Courbes d'optimisation des paramètres g et gp

Nous nous sommes d'abord intéressée à l'optimisation des paramètres des méthodes RR et TRR. Nous avons choisi, pour cela, une approche séquentielle (voir figure 3.1) : nous commen-

3.4. Courbes d'optimisation des paramètres g et gp

Jeu de données	Intervalle de valeurs (unités log)	Nb de molécules	Nb de descripteurs	Type de descripteurs	Source biblio
A2AR ISIDA	[3.53 ; 9.92]	767	483	fragments ISIDA, type IAB(2-5)	[141, 165-167]
A2AR MOE2D	[3.53 ; 9.92]	767	186	MOE 2D	[141, 165-167]
LogS	[-11.62 ; 1.58]	1635	437	fragments ISIDA, type IAB(2-4)	[158-162]
pKa	[-5.16 ; 2.51]	924	704	EED	[163]

Tableau 3.1 – Jeux de données utilisés pour l'étude méthodologique de la TRR.

çons par optimiser le paramètre g pour construire un modèle RR, puis nous optimisons gp , à g constant, pour construire le modèle TRR associé.

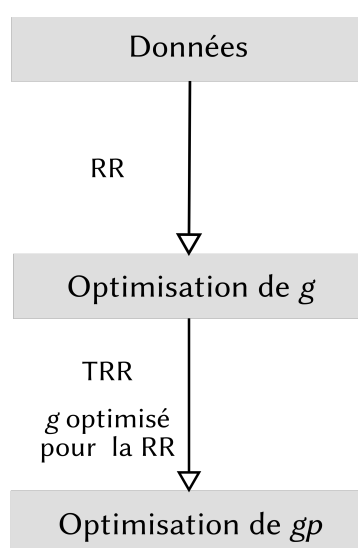


FIGURE 3.1 – Optimisation séquentielle des paramètres g et gp de la TRR. Nous commençons par optimiser le paramètre g pour construire un modèle RR, puis nous optimisons gp , à g constant, pour construire le modèle TRR associé.

Pour comprendre comment évolue la RMSE lorsque l'on fait varier g , nous avons fait une 5×2 -CV au cours de laquelle nous l'avons fait varier entre 10^{-12} et 10^{12} . La figure 3.2 correspond à une courbe d'optimisation caractéristique obtenue pour ce paramètre g . Cette courbe caractéristique comporte un premier plateau pour les faibles valeurs de g . Ce plateau correspond au phénomène de sous-apprentissage. Nous observons ensuite une zone caractérisée par un puits de RMSE minimales. Ce puits correspond à la zone qui nous intéresse : le minimum correspond au g optimal. Enfin, nous observons une remontée de la RMSE et un nouveau plateau : c'est une situation de sur-apprentissage. À l'aide de cette courbe, nous pouvons donc voir que le g optimal est facile à localiser, car il est situé dans un puits assez large et profond correspondant à des RMSE basses.

Une fois que nous avons identifié le g optimal, nous nous sommes intéressés au gp optimal. Nous avons fixé g , puis nous avons refait exactement la même procédure : nous avons effectué une 5×2 -CV au cours de laquelle le paramètre gp a varié entre 10^{-12} et 10^{12} . La figure 3.3 correspond à une courbe d'optimisation caractéristique du paramètre gp . Nous pouvons voir que l'allure de cette courbe est similaire à celle obtenue pour l'optimisation de g : observation d'un plateau pour les faibles valeurs de gp , suivi d'un puits de valeurs minimales, et enfin apparition d'un nouveau plateau pour des gp élevés. On remarque cependant que le puits observé ici est moins large et moins profond. Étant donné que l'on travaille à jeux de données et descripteurs constants, ceci n'est pas étonnant, car le g obtenu correspond au meilleur modèle RR que l'on puisse construire, il n'est donc pas choquant qu'il n'y ait qu'une faible marge de manœuvre pour obtenir un modèle TRR qui surpasse en qualité de prédiction le modèle RR associé.

Les figures 3.2 et 3.3 sont caractéristiques des optimisations des paramètres. Ces courbes montrent que ce problème d'optimisation est convexe. Aussi la procédure d'optimisation est efficacement automatisée par la méthode du nombre d'or [157].

Lors de l'optimisation des paramètres, nous avons observé que, dans la majorité des cas, le gp obtenu est inférieur à g et conduisait à un effet transductif positif, bien qu'il soit faible (entre 0,87 % pour pKa et 2,06 % pour A2AR MOE2D). Ceci est cohérent avec l'algorithme TRR : un poids moindre sur les données non étiquetées indique que celles-ci servent à apporter une correction au modèle RR initial.

Compte tenu que la transduction peut être considérée comme une correction de la RR (voir la section 2.2.3 du chapitre 2 et l'équation 2.2 associée), le paramètre gp doit être inférieur au paramètre g . Par la suite, il faudra donc veiller à chercher un gp dans un intervalle de valeurs ayant pour borne supérieure g . Par la suite, nous optimiserons donc nos paramètres dans l'intervalle $[10^{-4}; 10^4]$ pour g et dans l'intervalle $[g \times 10^{-4}; g]$ pour gp .

3.5 Impact de la taille relative du jeu d'entraînement et du paramètre gp sur l'effet transductif

Nous avons observé les allures caractéristiques des courbes d'optimisation, et nous avons déterminé que le paramètre gp devait être inférieur au paramètre g . Nous allons maintenant voir si nous pouvons observer un effet transductif, et dans quels cas il est le plus important.

3.5.1 Procédure de modélisation

Afin d'étudier les variations de l'effet transductif en fonction de la taille du jeu d'entraînement, nous avons mis en place une procédure de modélisation, schématisée en figure 3.4. Nous commençons par partager le jeu de données entier en deux parts égales. Une moitié de ces données sert de jeu de données non étiquetées : c'est notre jeu de validation. L'autre moitié sert

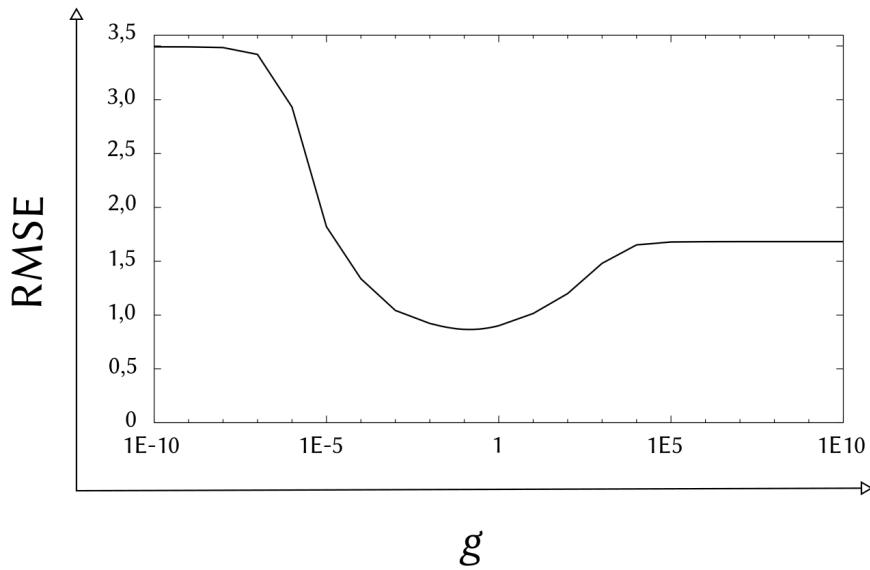


FIGURE 3.2 – Courbe d'optimisation typique du paramètre g obtenue pour le jeu de données LogS. On observe un premier plateau correspondant au phénomène de sous-apprentissage, puis un puits de valeurs optimales assez large, et enfin une remontée suivie d'un plateau illustrant le phénomène de sur-apprentissage.

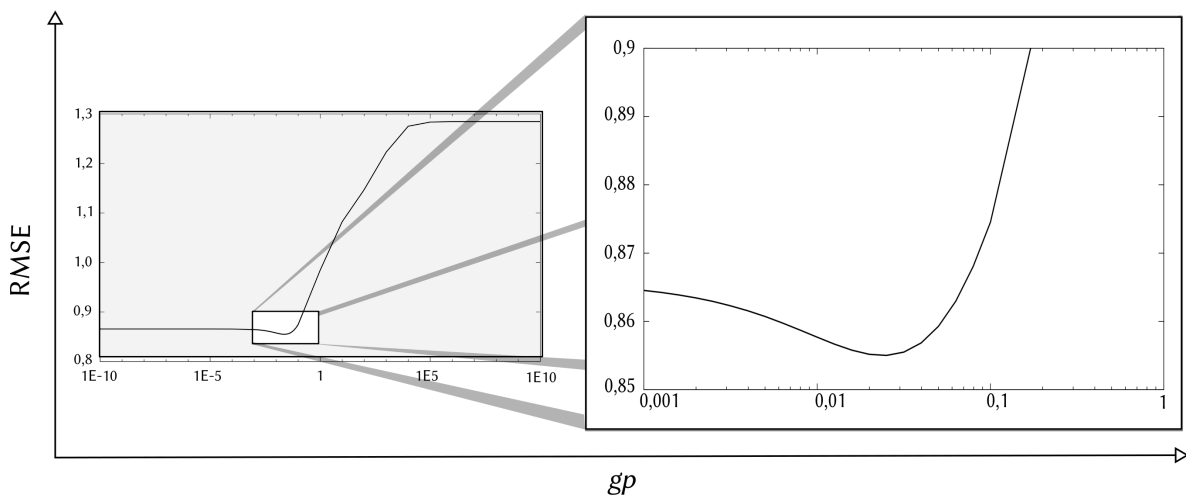


FIGURE 3.3 – Courbe d'optimisation typique du paramètre gp obtenue pour le jeu de données pKa. On observe une allure similaire à la courbe d'optimisation de g : premier plateau correspondant au phénomène de sous-apprentissage, puis un puits de valeurs optimales assez large, et enfin une remontée suivie d'un plateau illustrant le phénomène de sur-apprentissage. On remarque cependant que le puits de valeurs optimales est plus étroit et moins profond que dans le cas de l'optimisation de g .

de réservoir de données étiquetées. De ce réservoir, on construit différents jeux d'entraînement de taille variable en prélevant un pourcentage x variable de données, x étant compris entre 5 % et 100 % de la quantité de données du réservoir. Une fois que nous avons notre jeu d'entraînement de taille x % et notre jeu de données non étiquetées, nous pouvons construire des modèles.

Nous construisons d'abord un modèle RR pour lequel le paramètre g est optimisé à l'aide d'une 5×5-CV. Cette optimisation-ci utilise uniquement les données étiquetées. Nous appliquons ce modèle au jeu de données non étiquetées, puis nous calculons la $RMSE_{RR}$. Le jeu de validation n'est donc utilisé qu'une seule fois pour valider le modèle RR final. Ensuite, nous construisons un modèle TRR, qui aura pour paramètre g le même que celui du modèle RR optimisé sur les données étiquetées uniquement.

L'optimisation du paramètre gp suit deux protocoles différents. Dans un premier temps, le paramètre gp optimal est obtenu rétrospectivement. Cette expérience sert à démontrer qu'un effet transductif existe bien étant donné un jeu d'entraînement, un jeu de validation et un modèle non transductif. Dans un second temps, une estimation de la valeur optimale de gp est recherchée en utilisant à nouveau une procédure de 5 × 5-CV sur les seules données étiquetées. Dans ce protocole, le jeu de validation n'est donc utilisé qu'une seule fois pour valider le modèle. Dans chacun de ces protocoles, nous calculons la $RMSE_{TRR}$ et l'effet transductif afin d'évaluer les performances relatives des modèles RR et TRR dans les différents cas de figure.

En pratique, puisque l'on applique une procédure de 5 × 5-CV, chaque estimation de paramètre repose sur 25 optimisations individuelles. La valeur retenue est alors la médiane des 25 valeurs obtenues. Toutefois, la valeur optimale peut ne pas exister ou alors se situer en dehors de l'intervalle de recherche initialement défini. Dans ce cas, l'optimisation échoue et elle ne contribue pas au calcul de la médiane.

Pour finir, pour chaque taille x % de jeu d'entraînement, nous avons fait 25 expériences, ce qui nous permet d'avoir 25 estimations de l'effet transductif pour une taille donnée. En effet, la sélection de x % de données parmi les données étiquetées est répétée 5 fois, et il en est de même pour le partage initial du jeu de données en deux parts égales. Cela garantit des statistiques suffisantes aux observations que nous faisons.

3.5.2 Recherche d'un gp optimal

Pour commencer nous avons recherché une valeur optimale du paramètre gp *a posteriori* suivant le premier protocole. L'objectif est de vérifier si pour un jeu d'entraînement, un jeu de validation et un modèle non transductif donnés, il existe un effet transductif indépendamment de notre capacité à trouver cette valeur optimale de gp permettant d'en profiter.

L'effet transductif a été tracé en fonction de la taille relative du jeu d'entraînement. La figure 3.5 présente ces graphes pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa respectivement. La première chose que l'on peut noter, c'est qu'aucun effet transductif négatif n'a été relevé, peu importe le jeu de données considéré. Cela signifie que l'application de la

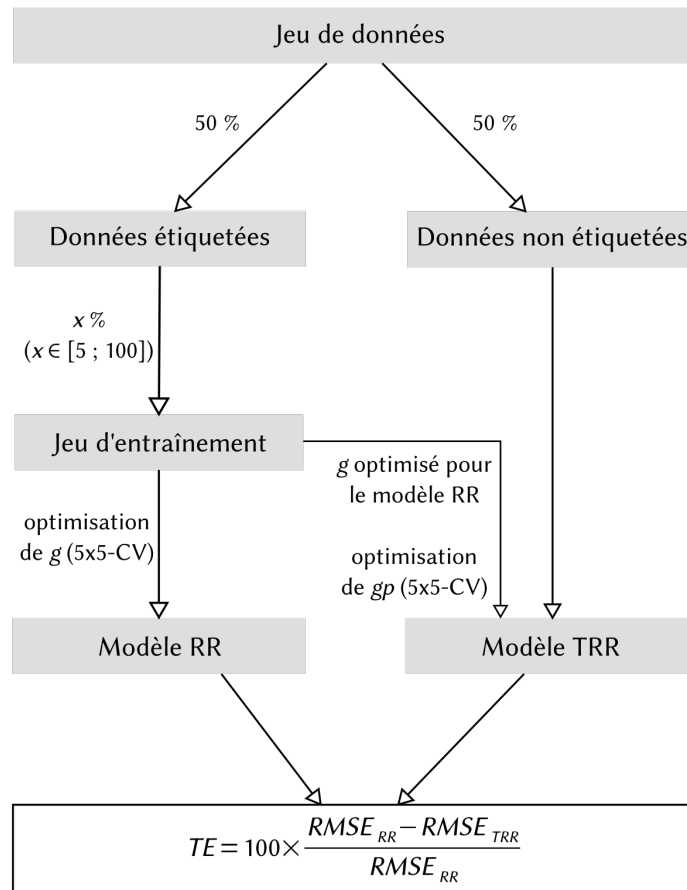


FIGURE 3.4 – Procédure suivie pour voir l'impact de la taille relative du jeu d'entraînement sur le TE. Nous avons partagé le jeu de données en un jeu de données considérées comme étant non étiquetées et un jeu de données étiquetées. De ce dernier, on prélève un pourcentage x de données (x étant compris entre 5 % et 100 %) pour obtenir notre jeu d'entraînement. Ce jeu d'entraînement est d'abord utilisé pour optimiser le paramètre g et construire le modèle RR, puis on combine le jeu d'entraînement et le jeu de données non étiquetées pour optimiser gp et construire le modèle TRR. À la fin de la procédure, on calcule les RMSE des modèles RR et TRR pour obtenir la mesure de l'effet transductif.

TRR n'a pas provoqué ici de dégradation des modèles par rapport à ce que l'on avait avec la RR. Nous observons que cet effet transductif peut être significatif puisque l'on relève des effets transductifs supérieur à 40 % pour le jeu de données pKa. Cependant, ils restent généralement en dessous des 5 %. Majoritairement, l'impact de la transduction sur le modèle est donc modeste. Nous observons également que certains cas sont caractérisés par un effet transductif proche de zéro. Pour ces cas-là, le *gp* obtenu est très faible : il n'y a pas de transduction pour ces ensembles jeu d'entraînement, jeu de validation et modèle non transductif.

En ce qui concerne l'impact de la taille du jeu d'entraînement sur les performances de la TRR, on constate que plus le jeu de données étiquetées est petit par rapport au jeu de données non étiquetées, et plus l'effet transductif a des chances d'être élevé. Cette tendance est clairement visible avec le jeu de données pKa, mais reste observée pour les autres jeux de données. Cela signifie qu'il semble préférable, pour augmenter les chances d'obtenir une amélioration significative des prédictions, de travailler avec un jeu d'entraînement d'une taille significativement plus petite (entre 5 % et 20 %) que celle du jeu de données non étiquetées.

3.5.3 Recherche d'un *gp* par validation croisée

Étant donné que nous avons observé une majorité de cas pour lesquels l'effet transductif était visible avec un *gp* optimal, nous avons essayé de retrouver ce *gp* sans prendre en compte les étiquettes du jeu de données non étiquetées. Ceci correspond au second protocole d'optimisation. Les résultats obtenus sont présentés dans la figure 3.6 pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa.

On observe toujours des cas pour lesquels l'effet transductif est élevé, ce qui signifie qu'il est possible de retrouver le *gp* optimal dans certains cas. Cependant, cette fois-ci, les figures obtenues montrent l'apparition de cas pour lesquels l'effet transductif est négatif. Nous ne sommes donc pas toujours capable de trouver un paramètre *gp* satisfaisant. Tout d'abord, quand il n'existe pas de transduction, notre protocole parvient quand même à proposer une valeur pour *gp*, mais celle-ci conduit systématiquement à une dégradation du modèle. Il existe aussi de nombreux cas où un effet transductif existe, mais la valeur de *gp* proposée par notre protocole conduit à une dégradation du modèle. Pour finir, nous observons à nouveau que plus le nombre de données étiquetées est faible par rapport au nombre de données non étiquetées, plus l'effet transductif est important, qu'il soit positif ou négatif.

Pour comprendre pourquoi, dans certains cas, des effets transductifs négatifs étaient observés, nous avons comparé les *gp* optimaux avec les *gp* obtenus par validation croisée. Les comparaisons sont illustrées en figure 3.7. Sur ces figures, nous pouvons voir que les *gp* obtenus par validation croisée sont souvent surestimés, bien que cela ne soit pas systématique.

Comme on peut le constater sur la figure 3.3, surestimer la valeur de *gp* peut rapidement conduire à une dégradation importante des performances du modèle. Au contraire, sous-estimer

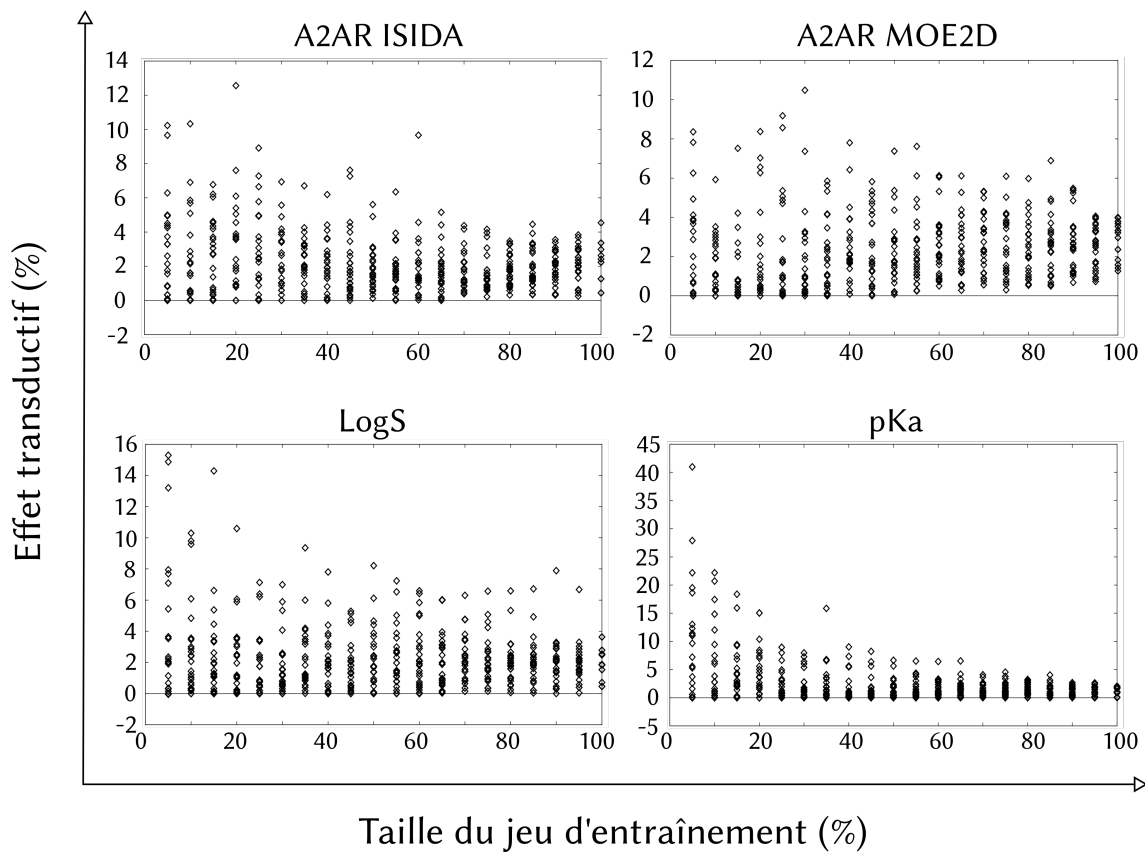


FIGURE 3.5 – TE en fonction de la taille relative du jeu d'entraînement, pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa, cas du gp optimal. Aucun TE négatif n'est relevé, ce qui signifie que la TRR n'a pas provoqué une chute de performance par rapport à la RR. Les TE positifs restent relativement modestes. Les cas pour lesquels le TE est proche de zéro s'explique par le fait que la courbe d'optimisation de gp était continue, linéaire et croissante : aucun puits de valeurs minimum n'a pu être trouvé. Enfin, plus la taille du jeu d'entraînement est petite, et plus les chances d'obtenir un TE positif élevé sont grandes.

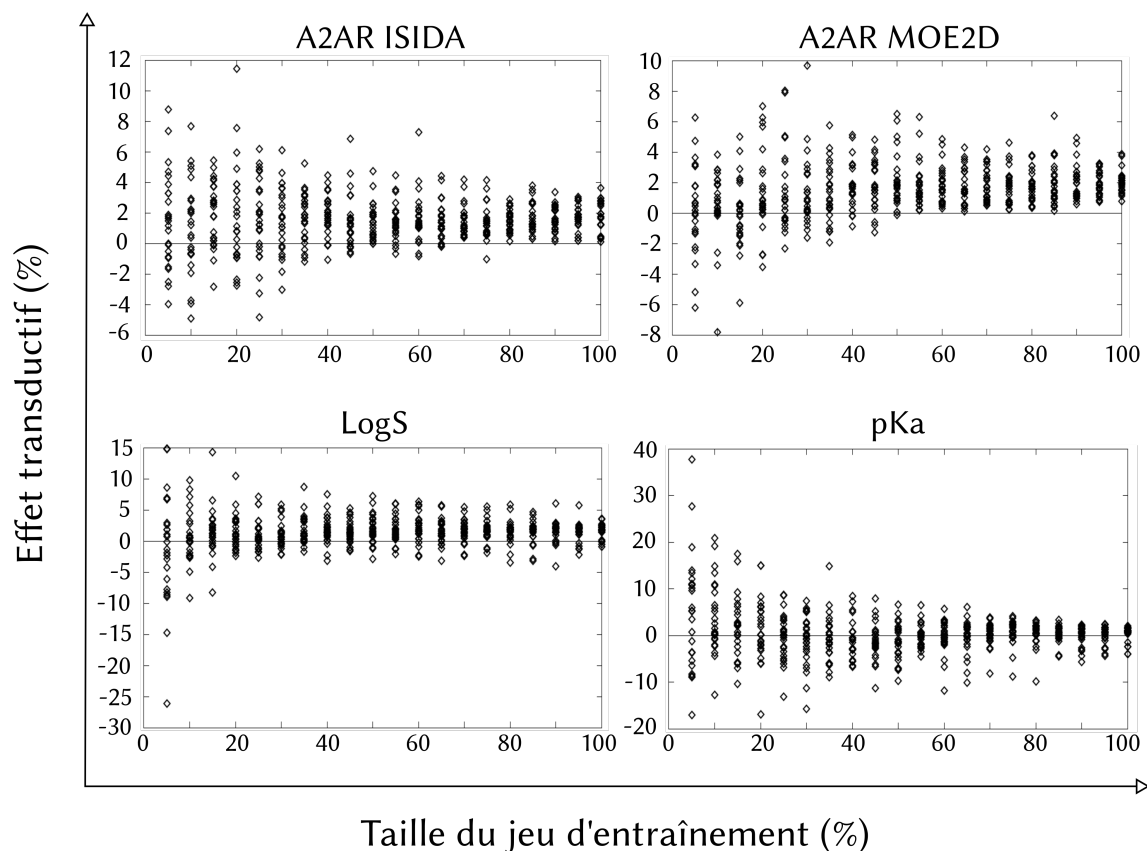


FIGURE 3.6 – TE en fonction de la taille relative du jeu d’entraînement, pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa, cas du gp obtenu par CV. Pour certains cas, le TE est élevé, ce qui signifie qu’il est possible, pour ces cas-ci de s’approcher du modèle construit avec un gp optimal. Il y a cependant des cas pour lesquels le TE est négatif. Nous ne sommes donc pas toujours capable de trouver un paramètre gp satisfaisant. Enfin, nous observons toujours une amplitude d’effet transductif très grande pour les petites tailles et plus faible pour les grandes tailles de jeu d’entraînement.

3.5. Impact de la taille relative du jeu d'entraînement et du paramètre gp sur l'effet transductif

cette valeur ne peut, au pire, que conduire à amener l'effet transductif à 0. Par conséquent, un choix conservatif et prudent est de tenter de sous-estimer la valeur de gp .

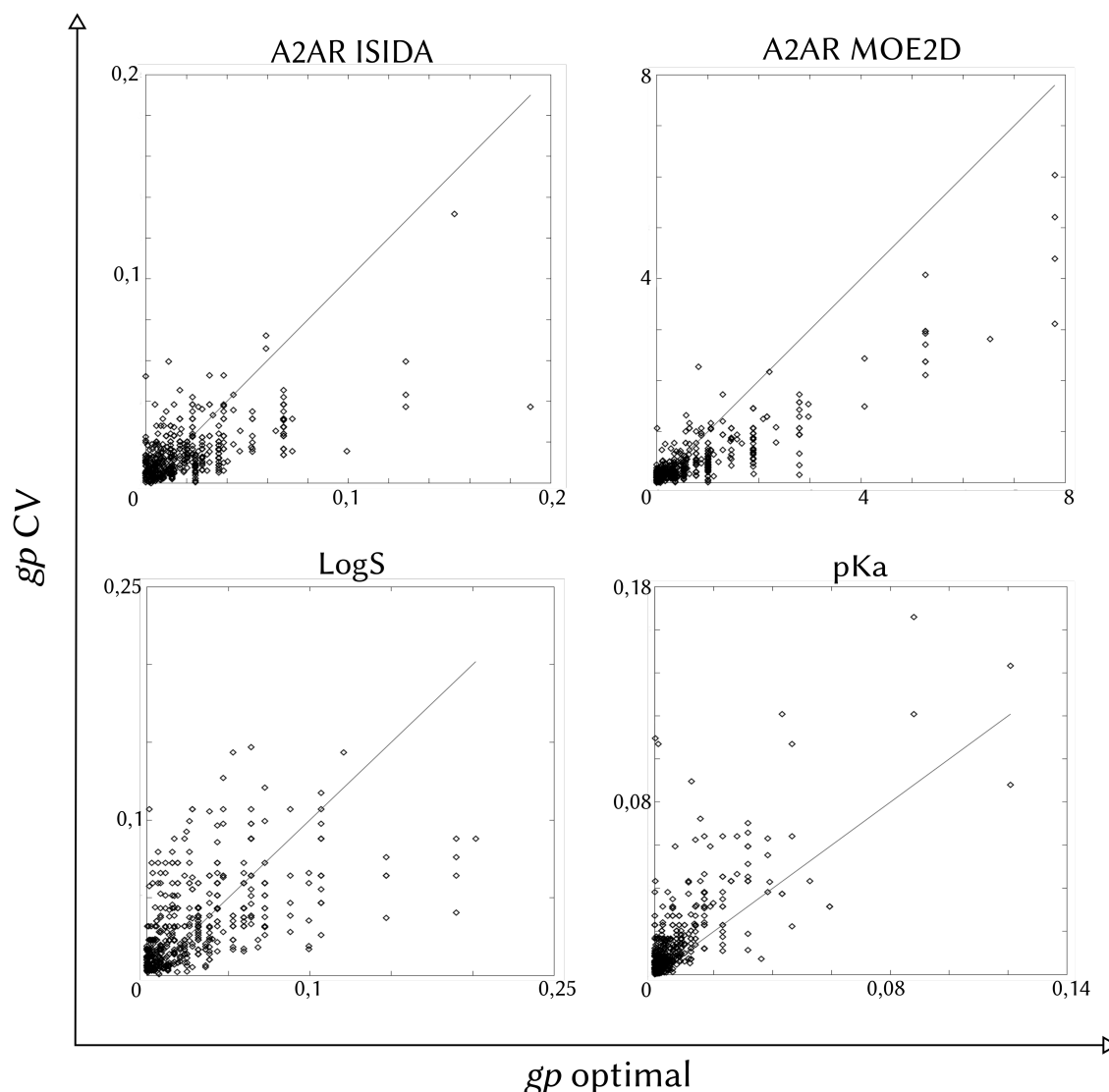


FIGURE 3.7 – gp obtenu par CV en fonction du gp optimal pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa. On observe que l'on obtient rarement le gp optimal par CV, il est soit surestimé (pKa et dans une moindre mesure LogS) soit sous-estimé (A2AR ISIDA et A2AR MOE2D).

Pour cette raison, nous avons divisé par 10 les gp obtenus à la fin de la validation croisée. Ce gp est appelé gp heuristique. Le but ici était de guérir les cas pour lesquels l'effet transductif était négatif, tout en essayant de conserver les cas pour lesquels l'effet transductif était positif. Les résultats sont présentés dans la figure 3.8. Nous observons une nette diminution du nombre de cas négatifs (voir le tableau 3.2) : entre 6,6 % (jeu de données LogS) et 13,2 % (jeu de données pKa) des cas sont maintenant négatifs, alors qu'auparavant on pouvait monter à 42,4 % de cas négatifs pour le jeu de données pKa. Le pourcentage de cas pour lesquels l'effet transductif est

négatif ou nul observé pour le *gp* heuristique tend vers le nombre de cas à TE nul observé avec le *gp* optimal. De plus, l'amplitude des cas négatifs est nettement réduit. Néanmoins, ceci a également pour conséquence de réduire l'amplitude des effets transductifs positifs, qui restent majoritairement en dessous de 5 %. L'allure globale de ces graphes ne change pas : plus la taille du jeu d'entraînement est petite par rapport à celle du jeu de données non étiquetées, et plus l'amplitude des effets transductifs est importante.

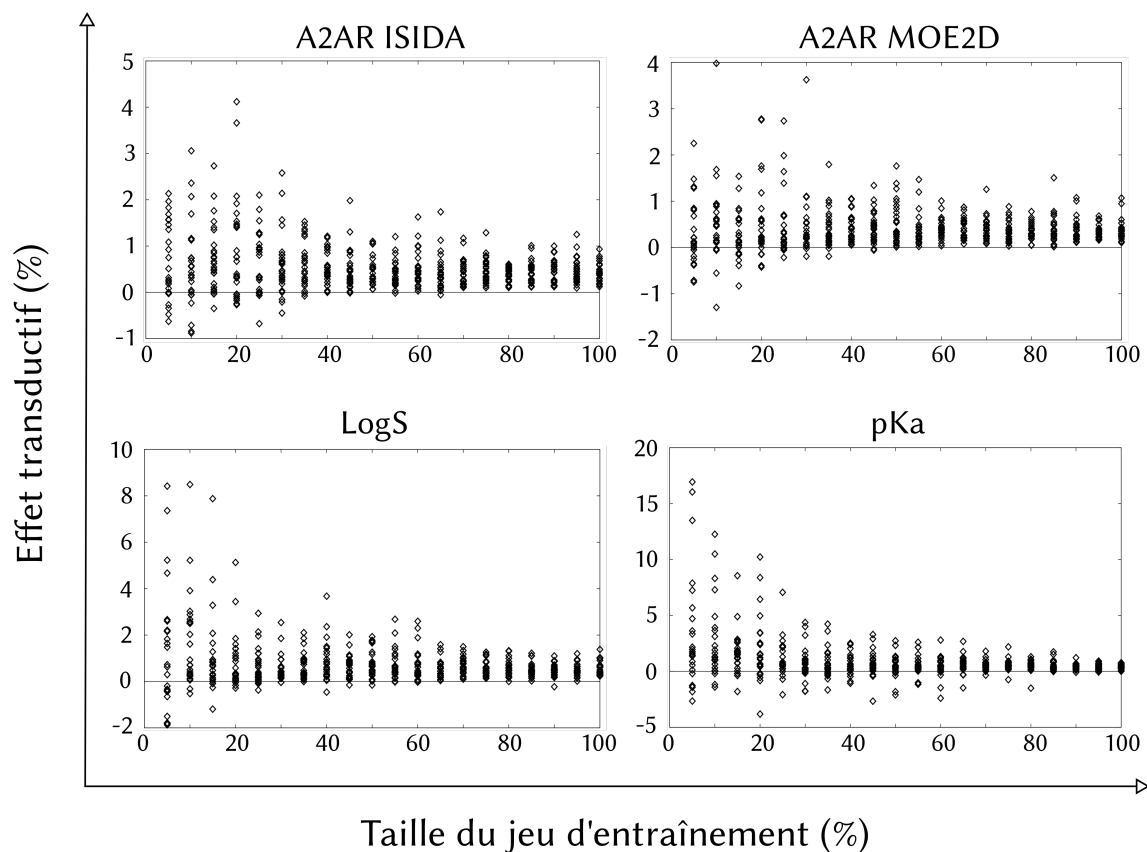


FIGURE 3.8 – TE en fonction de la taille relative du jeu d'entraînement, pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa, cas du *gp* heuristique. Nous observons une nette régression du nombre de cas négatifs, ainsi qu'une diminution des valeurs de TE positives. Ici encore, plus la taille du jeu d'entraînement est petite, plus l'amplitude des TE observés est grande.

La comparaison de la RMSE obtenue avec la TRR en fonction de celle calculée pour la RR (figure 3.9) est également intéressante. Elle montre que plus la RMSE du modèle non transductif est importante, plus l'effet transductif est important lui aussi. Il est donc plus fort là où il est le plus utile.

3.5. Impact de la taille relative du jeu d'entraînement et du paramètre gp sur l'effet transductif

Jeu de données	gp optimal	gp CV	gp heuristique
A2AR ISIDA	5,0	15,0	7,4
A2AR MOE2D	6,0	12,4	7,0
LogS	3,6	22,4	6,6
pKa	10,2	42,4	13,2

Tableau 3.2 – Pourcentage de cas pour lesquels le TE est inférieur ou égal à zéro. On constate que lorsque l'on divise par 10 le gp obtenu par validation croisée (gp heuristique), le nombre de cas ≤ 0 tend vers le nombre de cas $\simeq 0$ relevés dans le cas du gp optimal.

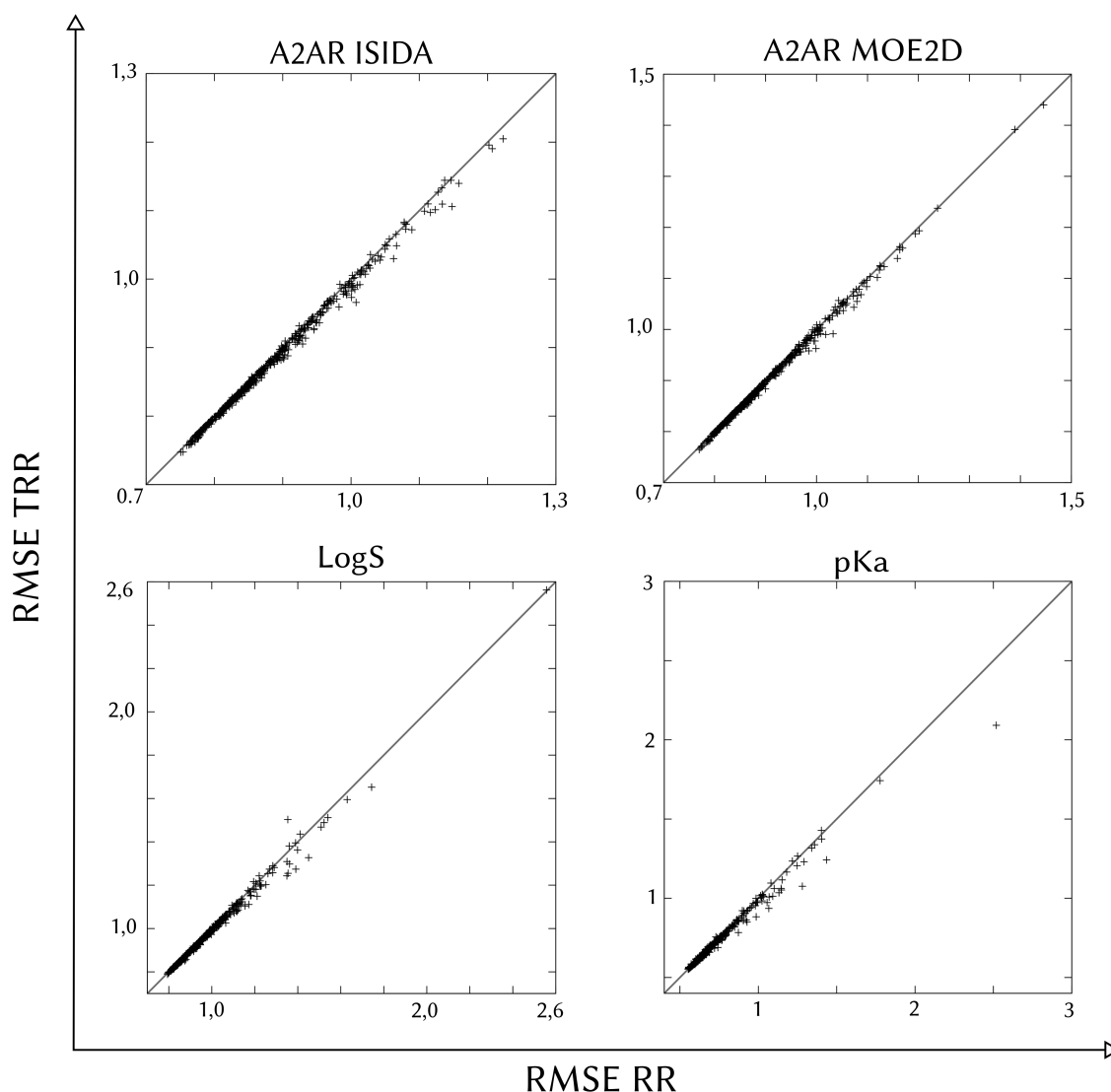


FIGURE 3.9 – $RMSE_{TRR}$ en fonction de la $RMSE_{RR}$ pour les jeux de données A2AR ISIDA, A2AR MOE2D, LogS et pKa, cas du gp heuristique. Le TE est positif si les points sont en dessous de la droite d'équation $y = f(x)$. On observe ici que la majorité des points sont en dessous de cette droite, ce qui signifie que pour la majorité des cas le TE obtenu est positif.

3.5.4 Étude des cas pour lesquels l'effet transductif est négatif

Pour comprendre pourquoi certains cas étaient caractérisés par des effets transductifs en validation croisée, nous avons commencé par regarder ce que nous avons obtenu comme résultats lors de la recherche d'un gp optimal. Il s'est avéré que les cas négatifs correspondent aux cas où le gp optimal était associé à un effet transductif nul. Pour ces cas, les courbes d'optimisation ne permettent pas de trouver un minimum pour le gp ce qui signifie qu'il n'y a pas de transduction dans les situations correspondantes. Utiliser la transduction dans ces cas-là ne peut donc conduire qu'à une détérioration du modèle.

Nous avons cherché à comprendre pourquoi, dans ces cas-là, la transduction était néfaste. Nous nous sommes d'abord intéressée aux courbes d'ajustement du jeu d'entraînement dans les situations où l'ensemble jeu d'entraînement, jeu de validation et modèle non transductif ne bénéficie d'aucune transduction. Cependant, aucune corrélation n'a été établie entre le R^2 de la courbe d'ajustement et l'effet transductif.

Une seconde hypothèse est que le jeu de validation soit trop différent du jeu d'entraînement pour bénéficier de la transduction. Pour vérifier cette idée, nous avons également testé l'application d'un domaine d'applicabilité de type boîte bornée. Nous avons consigné dans l'histogramme 3.10 les évolutions des effets transductifs pour les 4 jeux de données. On observe que, majoritairement, on arrive à guérir un grand nombre de cas pour lesquels l'effet transductif était négatif (avec une exception pour le jeu de données LogS pour lequel nous ne guérissons que la moitié des cas environ). Cependant, nous constatons également que de nouveaux cas d'effet transductifs négatifs apparaissent. Il est donc difficile de conclure sur l'intérêt du domaine d'applicabilité dans le cadre de la transduction, d'autant plus que les molécules bénéficiant le plus de cet effet transductif semblent être les molécules situées hors du domaine d'applicabilité.

3.6 En résumé

Nous avons implémenté la TRR, puis nous l'avons étudié. Nous avons déterminé qu'en théorie une approche séquentielle de l'optimisation des paramètres (c'est-à-dire déterminer g , puis déterminer gp à g fixé) nous permettait d'obtenir des modèles TRR plus performants que les modèles RR correspondants. Nous avons ensuite appliqué cette méthode d'optimisation pour déterminer l'impact de la taille relative du jeu d'entraînement sur l'effet transductif. Nous avons constaté que l'effet transductif était le plus intéressant pour des jeux d'entraînement de petite taille (entre 5 % et 20 %) par rapport au jeu de validation. Nous avons également observé que la transduction n'existe pas systématiquement pour un jeu d'entraînement, un jeu de validation et un modèle non transductif fixé. Dans ce type de situations et dans le cas où le paramètre est surestimé, la méthode transductive aboutit à une dégradation des performances du modèle. Pour limiter le nombre de ces cas, nous avons utilisé le gp heuristique consistant à diviser par 10 le gp sélectionné en validation croisée. En effet, il est prudent de sous-estimer

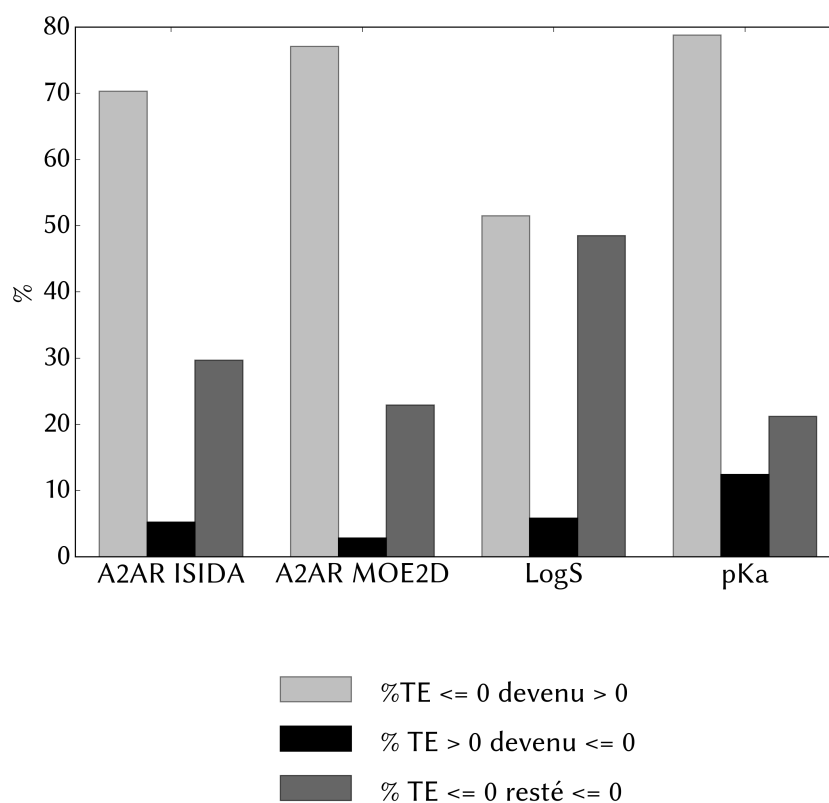


FIGURE 3.10 – Histogrammes présentant l'évolution des TE lorsque l'on applique un domaine d'applicabilité de type boîte bornée. On observe que, majoritairement, on arrive à guérir un grand nombre de cas pour lesquels l'effet transductif était négatif (avec une exception pour le jeu de données LogS pour lequel nous ne guérissons que la moitié des cas). Cependant, nous constatons également que de nouveaux cas d'effet transductifs négatifs apparaissent. Il est donc difficile de conclure sur l'intérêt du domaine d'applicabilité dans le cadre de la transduction, d'autant plus que les molécules bénéficiant le plus de cet effet transductif semblent être les molécules situées hors du domaine d'applicabilité.

la valeur de gp car, au pire, cela ne fait qu'éteindre la transduction, et donc ne détériore pas le modèle. Ceci nous permet de limiter le nombre de cas négatifs, mais provoque également une baisse de l'amplitude des effets transductifs positifs. Nous avons finalement cherché à comprendre pourquoi certains cas ne semblaient pas permettre l'application de la transduction. Nous n'avons pas trouvé de réponse générale à ce problème ; néanmoins la qualité du modèle RR de départ ainsi que la nature des composés présents dans le jeu de test semblent influencer sur cette absence de transduction.

Nous savons maintenant comment optimiser la méthode TRR pour maximiser les chances d'obtenir des modèles plus prédictifs que leurs homologues RR. Nous pouvons donc maintenant nous intéresser aux données principales de cette thèse : les liquides ioniques et les électrolytes pour batteries Li-ion.

Chapitre 4

Modélisation des liquides ioniques

Les premières modélisations de données réelles se sont focalisées sur les liquides ioniques. C'est un travail qui a été fait en collaboration avec l'équipe d'Isabelle Billard à l'IPHC de Strasbourg, ainsi qu'avec la société Solvionic, spécialisée dans la production de liquides ioniques. Nous nous sommes concentrés ici sur trois propriétés : la conductivité, la température de fusion et la viscosité.

Nous avons préparé des modèles SVR, RR et TRR. Les modèles SVR ont les meilleures performances ; nous avons donc fait une analyse poussée de nos données avec cette méthode, et nous avons mis les modèles SVR finaux en ligne à disposition de tous sur le serveur du laboratoire : infochim.u-strasbg.fr/webserv/VSEngine.html

4.1 Présentation des données

Pour chaque propriété, nous avons travaillé avec trois jeux de données : un jeu d'entraînement et deux jeux de validation. Le premier jeu de validation est constitué de données issues de la littérature, le second est constitué de données obtenues expérimentalement durant ce projet. Les mesures de la conductivité et de la viscosité ont été faites à 25 °C. La composition de chacun des jeux de données pour chacune des propriétés est donnée dans le tableau 4.1, et leur répartition est illustrée dans la figure 4.1.

Les jeux d'entraînement proviennent d'une base de données interne au laboratoire. Celle-ci contient des données issues de la littérature, collectées jusque dans la première moitié de 2015, pour plus de 2 000 liquides ioniques. De cette base, nous avons extrait 85 liquides ioniques pour la conductivité à 25 °C, 228 liquides ioniques pour la température de fusion et 189 liquides ioniques pour la viscosité à 25 °C. Suite à une première étape de modélisation et à l'analyse des points aberrants découverts dans les différents jeux d'entraînement, nous avons veillé à ce que

les données sélectionnées ici concernent les échantillons contenant le moins d'eau possible (les points dont la teneur en eau était supérieure à 500 ppm ont été retirés).

Les premiers jeux de validation contiennent des données issues de la littérature récente, c'est-à-dire publiées depuis la seconde moitié de 2015. Aucune précaution n'a été prise lors de la sélection de ces données. Ceci nous permettra de vérifier la qualité de nos modèles, et de voir si ces derniers peuvent être utiles dans la détection de points aberrants.

Les derniers jeux de validation contiennent des données issues du catalogue de la compagnie Solvionic. La conductivité à 25 °C, la température de fusion et la viscosité à 25 °C ont été mesurées par nos collègues Isabelle Billard et Olga Klimchuk. Cet ensemble de données est le plus homogène puisque les données ont toutes été mesurées par la même équipe et dans les mêmes conditions.

Pour faire ces mesures, les échantillons reçus ont été placés sous pompe à vide à 70 °C pendant 24 heures, puis placés sous argon. La présence d'eau dans l'échantillon a été contrôlée par la technique de Karl-Fischer. Toutes les mesures de conductivité et de viscosité ont été faites à température ambiante (25 °C).

Les mesures de conductivité ont été menées à l'aide de l'appareil de mesure de conductivité *SevenMulti S47 Mettler TOLEDO* avec des capteurs *InLab751-4 mm* et avec la compensation automatique de la température. Les capteurs ont été calibrés avec des solutions standard (1,41 mS/cm et 12,88 mS/cm) avant chaque utilisation. La mesure de la conductivité a été effectuée trois fois par échantillon et la valeur moyenne a été calculée. Après chaque mesure, les capteurs ont été nettoyés avec de l'eau déminéralisée et de l'éthanol.

Pour la viscosité, trois échantillons (le Tf₂N de N-propyl-N-méthylpyrrolidinium, le Tf₂N de 1-éthanol-3-méthylimidazolium et le Tf₂N de 1-butyl-1-méthylpyrrolidinium) ont été mesurés à l'aide du viscosimètre ARES (*Rheometric Scientific*) avec une géométrie cône/assiette (L 40 mm, ϕ 0,0436 rad, écart 5/100) nécessitant 1,5 ml de l'échantillon. La température a été contrôlée à 25±0,1 °C. Les autres valeurs ont été données par la société Solvionic.

Les températures de fusion et de transition vitreuse ont été mesurées avec un appareil de calorimétrie différentielle à balayage *Thermal Analysis DSC 2920 Modulated DSC*. Les échantillons d'analyses ont été préparés en ajoutant 6 mg de liquide ionique dans réceptacle en aluminium. Les échantillons ont été chauffés dans le calorimètre à 125 °C pendant 10 minutes pour enlever toute trace d'eau. Ils ont ensuite été refroidis à -100 °C. Les courbes de chauffe et de refroidissement ont été enregistrées à une allure de 10 ° par minute. Les températures de fusion, de cristallisation et de transition vitreuse ont été mesurées à partir de la seconde courbe de chauffe. Les résultats obtenus ont une précision de ±1 °C.

On notera que pour un certain nombre de liquides ioniques, nous n'avons qu'une seule ou deux des trois propriétés étudiées, et ce, y compris pour les données Solvionic. Ceci s'explique car tous les liquides ioniques ne sont pas forcément liquides à 25 °C.

Pour une propriété donnée, les différents jeux de données peuvent avoir des ions en commun, mais pas les liquides ioniques. Les structures des ions ont été standardisées : nous avons

positionné la charge positive sur l'atome substitué avec la chaîne la plus courte, pour les guanidiniums nous avons placé la charge sur le carbone central, et les anions ont tous été préparés en utilisant le même protocole.

Pendant la préparation des jeux d'entraînement, nous avons pu observer de larges écarts de valeurs de propriétés. Par exemple, la température de fusion du chlorure de 1-butyl-3-méthylimidazolium a été détectée à 41,0 °C par Huddleston *et al.* [168] (2 200 ppm d'eau pour l'échantillon sec) alors que Han *et al.* [169] ont reporté une température de fusion de 89,0 °C pour le même composé (quantité d'eau inconnue). Autre exemple, la conductivité à 25 °C du triflate de 1-butyl-3-méthylimidazolium est de 9 mS/cm pour l'équipe de Zech *et al.* [170] (avec une fraction massique d'eau inférieure à $8 \cdot 10^{-5}$ g) et de 0,3 mS/cm pour l'équipe de Mbondo Tsamba *et al.* [171] (« Typical water quantities between 50 and 100 ppm were measured », des quantités d'eau entre 50 et 100 ppm ont typiquement été mesurées). Finalement, la viscosité du même liquide ionique a été mesurée à 179,15 cP par nos collègues tandis que Mbondo Tsamba *et al.* [171] ont mesuré une valeur de 86,6 cP. Les écart-types relevées dans notre base de données sont d'environ 1 mS/cm pour la conductivité à 25 °C, 6 °C pour la température de fusion et 19 cP à 25 °C pour la viscosité.

	Conductivité	Température de fusion	Viscosité
Jeu d'entraînement	78	183	139
Jeu de validation Littérature	15	41	30
Jeu de validation Solvionic	19	22	14

Tableau 4.1 – Nombre de IL contenu dans chaque set. Les données du jeu d'entraînement et du jeu de validation Littérature viennent de la littérature, et les données du jeu de validation Solvionic ont été obtenues par nos collaborateurs sur des échantillons fournis par la compagnie Solvionic. Les données présentées pour les jeux d'entraînement correspondent aux jeux de données obtenus après le retrait des points aberrants.

4.2 Modélisation avec la SVR

Nous avons commencé par modéliser les liquides ioniques avec la SVR. Nous avons pu, lors de cette phase, identifier des points aberrants et construire des modèles prédictifs. Ces modèles sont accessibles par tous à l'adresse suivante : infochim.u-strasbg.fr/webserv/VSEngine.html

4.2.1 Procédure de modélisation

La modélisation des liquides ioniques s'est faite selon la procédure décrite en figure 4.2. Nous avons généré 272 fragmentations ISIDA différentes. Pour générer les descripteurs des liquides ioniques au complet, nous avons d'abord énuméré les fragments pour les cations d'une part et pour les anions d'autre part, puis nous avons concaténé les deux fichiers de descripteurs ainsi générés. Pour chacune de ces fragmentations, nous avons ensuite fixé la largeur de la

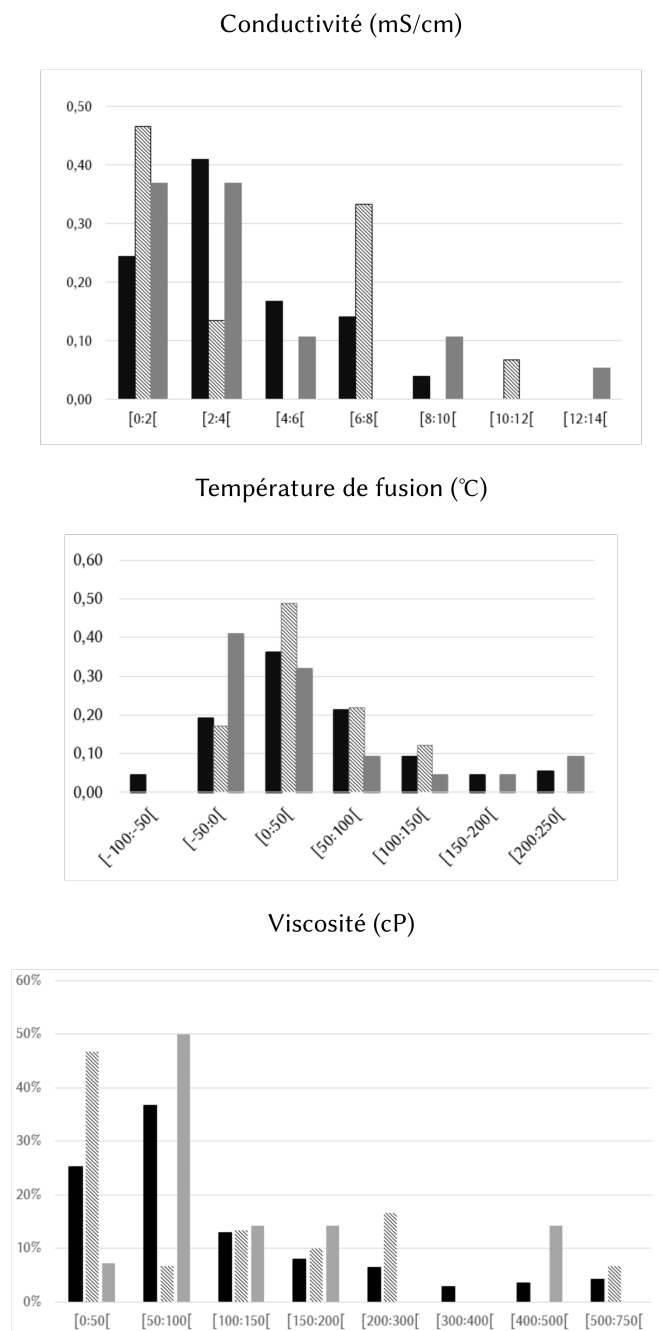


FIGURE 4.1 – Histogramme de répartition des données de la conductivité (en haut), de la température de fusion (au milieu) et de la viscosité pour les trois jeux de données (jeu d'entraînement en noir, jeu de validation contenant les données de la littérature récente en hachuré, et jeu de validation contenant les données Solvionic en gris). On peut voir que les trois jeux de données ne sont pas équilibrés. Pour la conductivité, la majorité des données se situe entre 0 mS/cm et 4 mS/cm, pour la température de fusion entre -50 °C et 100 °C, et pour la viscosité entre 0 cP et 100 cP

fonction de perte ϵ -sensible à 0,5 pour la conductivité, à 3 pour la température de fusion et à 9,5 pour la viscosité. Ces valeurs ont été choisies car elles correspondent à notre estimation initiale du bruit expérimental observé dans notre base de données. Le paramètre de coût a été optimisé pour chacune des fragmentations à l'aide d'une 5×5-CV. Nous avons ensuite classé les modèles individuels par rapport à leur RMSE moyenne de validation croisée. Nous avons conservé le meilleur modèle, ainsi que les modèles ayant des performances équivalentes selon un test de Student à 95 %. Nous avons finalement construit un modèle consensus à l'aide des modèles sélectionnés. Ce modèle consensus correspond à la moyenne arithmétique de l'ensemble des prédictions des différents modèles.

Nous avons également appliqué un domaine d'applicabilité de type contrôle de fragments. Dans le cas du modèle consensus, le composé est considéré comme hors du domaine d'applicabilité s'il est hors du domaine d'applicabilité d'au moins 50 % des modèles du consensus.

4.2.2 Points aberrants dans les jeux d'entraînement

Suite à une première série de modélisation, nous avons fait une recherche de points aberrants dans les différents jeux d'entraînement. Le but ici était d'avoir des jeux d'entraînement avec des données de bonne qualité pour pouvoir construire des modèles les plus fiables possibles. Pour cela, nous avons appliqué de façon itérative la procédure suivante, illustrée en figure 4.3 [172]. Le liquide ionique avec la plus grande erreur de prédiction en ajustement consensus est analysé. Si c'est un point aberrant confirmé, il est retiré du jeu d'entraînement, et un nouveau modèle consensus est construit. Les paramètres ne sont pas réoptimisés. Nous répétons cette procédure jusqu'à ce qu'il n'y ait plus de points aberrants confirmés. Étant donné que nous travaillons ici avec des petits jeux de données, nous avons limité le processus à 10 points suspects analysés. Les données présentées dans le tableau 4.1 correspondent, pour les jeux d'entraînement, aux jeux de données après retrait des points aberrants.

Pendant la détection de points aberrants, nous avons déterminé différentes raisons pour lesquelles un liquide ionique peut être considéré comme un point aberrant. Ces derniers sont listés dans les tableaux 4.2, 4.3 et 4.4 pour la conductivité, la température de fusion et la viscosité respectivement.

Quand plusieurs points aberrants provenaient d'une même référence, l'ensemble des données issues de la même référence a été considéré comme potentiellement aberrant. En effet, il y a peut-être un biais dans les mesures, même dans les cas pour lesquels la teneur en eau et en impuretés a été soigneusement vérifiée. Comme ce biais peut affecter les mesures, nous avons considéré qu'il était plus sûr de ne pas les utiliser pour la construction des modèles. Les liquides ioniques concernés sont les liquides ioniques 1, 2, 3, 5, 6, 7, 8, 9 du jeu d'entraînement de la conductivité (voir tableau 4.2) ainsi que les liquides ioniques 2, 3, 5, 6, 8 du jeu d'entraînement de la température de fusion (voir tableau 4.3).

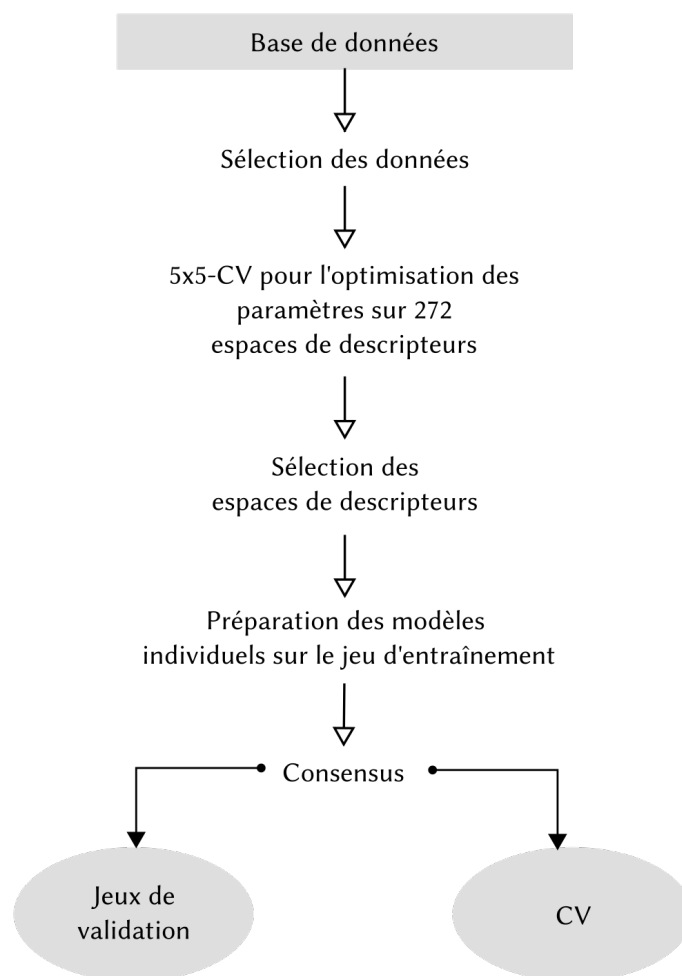


FIGURE 4.2 – Procédure suivie lors de la modélisation des liquides ioniques avec la SVR. À partir de notre base de données locale, nous avons soigneusement sélectionné des données pour constituer nos différents jeux d'entraînement (un pour la conductivité, un pour la température de fusion et un pour la viscosité). Nous avons exploité ces jeux d'entraînement afin de construire différents modèles SVR, basés sur différents ensembles de descripteurs ISIDA, à l'aide d'une 5×5 -CV. Les meilleurs modèles ont ensuite été sélectionnés, puis reconstruits sur le jeu d'entraînement pour créer un modèle consensus par propriété. Ce modèle consensus a ensuite été validé sur deux types de jeux de validation. L'un d'entre eux contient des données issues de la littérature récente, l'autre concerne les données mesurées par nos collègues.

Pour certains liquides ioniques, la valeur indiquée n'a pas été retrouvée dans la référence. C'est le cas, par exemple, du catalogue Sigma-Aldrich [173]. Étant donné qu'il n'était pas possible de confirmer ces valeurs, elles ont été rejetées. Il s'agit ici des liquides ioniques 1, 3, 8 du jeu d'entraînement de la conductivité (voir tableau 4.2).

Nous avons rencontré un problème similaire avec certaines valeurs qui n'ont pas été mesurées par les équipes les publiant. C'est notamment le cas des données trouvées dans des revues de la littérature. Comme nous ne pouvions pas valider la pureté des échantillons utilisés pour mesurer ces valeurs, ces composés ont été retirés du jeu d'entraînement. Ceci a été observé pour les liquides ioniques 2 et 5 du jeu d'entraînement de la conductivité (voir tableau 4.2), les liquides ioniques 5, 6, 8 du jeu d'entraînement de la température de fusion (voir tableau 4.3) et le liquide ionique 6 du jeu d'entraînement de la viscosité (voir tableau 4.4).

Pour les liquides ioniques 4 et 9 du jeu d'entraînement de la conductivité (voir tableau 4.2) ainsi que les liquides ioniques 2, 4 et 7 du jeu d'entraînement de la viscosité (voir tableau 4.4), d'autres valeurs trouvées dans la littérature n'étaient pas en accord avec la valeur utilisée dans le jeu d'entraînement. Ces autres valeurs sont généralement plus proches des valeurs prédites par la SVR. Ces valeurs ont été corrigées.

Pour le liquide ionique 1 du jeu d'entraînement de la température de fusion (voir tableau 4.3), nous n'avons trouvé aucune autre mesure le concernant dans la littérature. La seule source disponible était celle qui était entrée dans la base. Dans ce cas, même si la pureté semblait correcte, nous avons préféré retirer ce point.

Enfin, les points pour lesquels la teneur en eau (%w) n'était pas vérifiée et l'échantillon testé non séché ont été retirés. Les composés concernés sont les liquides ioniques 2, 3, 4, 7 du jeu d'entraînement de la température de fusion (voir tableau 4.3) et le liquide ionique 3 du jeu d'entraînement de la viscosité (voir tableau 4.4).

Finalement, cette analyse nous a permis d'identifier deux structures mal saisies dans la base de données : les liquides ioniques 1 et 5 du jeu d'entraînement de la viscosité (voir tableau 4.4). Ces liquides ioniques ont été réintégré dans le jeu d'entraînement après correction.

Suite à cette analyse des points aberrants, nous avons décidé d'imposer de nouveaux critères pour la construction de nos jeux d'entraînement. Nous avons veillé à ce que ces jeux contiennent des données relatives à des échantillons pour lesquels la teneur en eau ne dépasse pas 500 ppm. Si la teneur en eau n'était pas précisée dans l'article, l'échantillon étudié devait avoir été soigneusement séché.

4.2.3 Résultats obtenus sur les jeux de validation

Nous avons répété la procédure de modélisation sur des jeux d'entraînements nettoyés. Les RMSE et les MAE obtenues pour les différents jeux de données pour les trois propriétés modélisées sont indiquées dans le tableau 4.5. Ces résultats tiennent compte de l'application d'un domaine d'applicabilité de type contrôle de fragment.

Chapitre 4. Modélisation des liquides ioniques

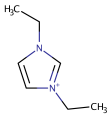
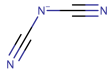
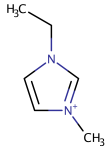
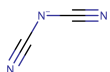
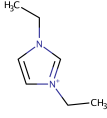
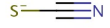
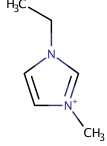
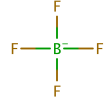
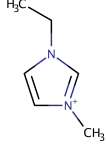
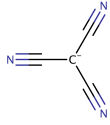
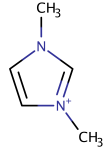

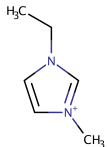

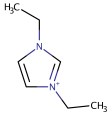
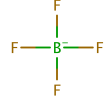
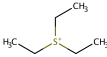

N° IL	Cation	Anion	Valeur Exp. (mS/cm)	Valeur Pred. (mS/cm)	Réf.	Problème détecté
1			27	5,3	[173]	4 points aberrants proviennent de cette référence, et cette valeur n'a pas été retrouvée dans la référence
2			28	16,9	[174]	2 points aberrants viennent de cette référence, et cette valeur n'a pas été mesurée par l'équipe
3			21	13,7	[173]	4 points aberrants proviennent de cette référence, et cette valeur n'a pas été retrouvée dans la référence
4			16,3	10	[175]	Autre valeur retrouvée dans la littérature : 10 mS/cm [176]
5			22	16,3	[174]	2 points aberrants viennent de cette référence. 2 références citées pour cet IL dans l'article en question ; une de ces références n'a pas été trouvée, l'autre donne une conductivité à 20 °C
6			15,5	10,3	[177]	2 points aberrants viennent de cette référence
7			14,8	9,4	[177]	2 points aberrants viennent de cette référence
8			12	7,7	[173]	4 points aberrants proviennent de cette référence, et cette valeur n'a pas été retrouvée dans la référence
9			8,2	5	[173]	4 points aberrants proviennent de cette référence, et la valeur trouvée dans la référence est de 5,5 mS/cm

Tableau 4.2 – Points aberrants détectés pour le jeu d'entraînement de la conductivité. Ces liquides ioniques sont considéré comme des points aberrants pour diverses raisons : divergences avec des valeurs trouvées dans la littérature, %w non précisé, source non fiable.

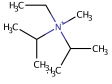

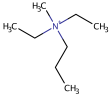
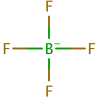
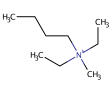

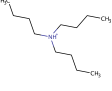

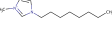



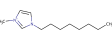

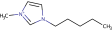

N° IL	Cation	Anion	Valeur Exp. (°C)	Valeur Pred. (°C)	Réf.	Problème détecté
1			328,4	209,2	[178]	Il n'était pas possible de comparer cette valeur à d'autres.
2			186	71,6	[177]	2 points aberrants viennent de cette référence, %w compris entre 400 ppm et 600 ppm
3			165	58,6	[177]	2 points aberrants viennent de cette référence, %w compris entre 400 ppm et 600 ppm
4			52	22,7	[179]	%w non contrôlé, échantillon non séché
5			0	74,2	[169]	3 points aberrants viennent de cette référence qui est une revue, %w non précisé, source non retrouvée
6			-79	7,3	[169]	3 points aberrants viennent de cette référence qui est une revue, %w non précisé, source non retrouvée
7			115	48,1	[180]	%w non contrôlé, échantillon non séché
8			-88	24,6	[169]	3 points aberrants viennent de cette référence qui est une revue, %w non précisé, source non retrouvée

Tableau 4.3 – Points aberrants détectés pour le jeu d'entraînement de la température de fusion. Les anomalies détectées pour ces liquides ioniques sont les suivantes : divergences avec des valeurs trouvées dans la littérature, %w non précisé, source non fiable.

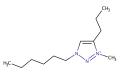
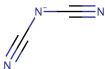



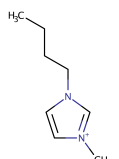
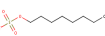
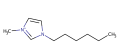



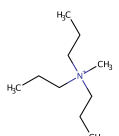

N° IL	Cation Structure	Anion Structure	Valeur Exp. (°C)	Valeur Pred. (°C)	Réf.	Problème détecté
1			977	517,3	[181]	Structure erronée
2			56	482,1	[30]	Autres références : 453 cP à 25,10 °C [182, 183]
3		Cl^-	716	225.8	[184]	%w : 1130 ppm
4			888,6	435,6	[185]	Autre référence [186] : 874,5 cP à 20 °C, 447,9 cP à 30 °C
5			496,5	140,8	[187]	Structure erronée
6			800	479.4	[183]	La référence est une revue, %w non précisé
7			593,4	305,8	[30]	Les autres valeurs trouvées dans la littérature sont entre 520 cP [188] et 540 cP [189] à 25 °C ; mesure faite par nos collègues à 25,10 °C : 520,3 cP

Tableau 4.4 – Points aberrants détectés pour le jeu d'entraînement de la viscosité. Les raisons pour lesquelles ces IL ont été identifiés comme étant des points aberrants sont les suivantes : mauvaise structure entrée dans la base, divergences avec des valeurs trouvées dans la littérature, %w élevé ou non précisé.

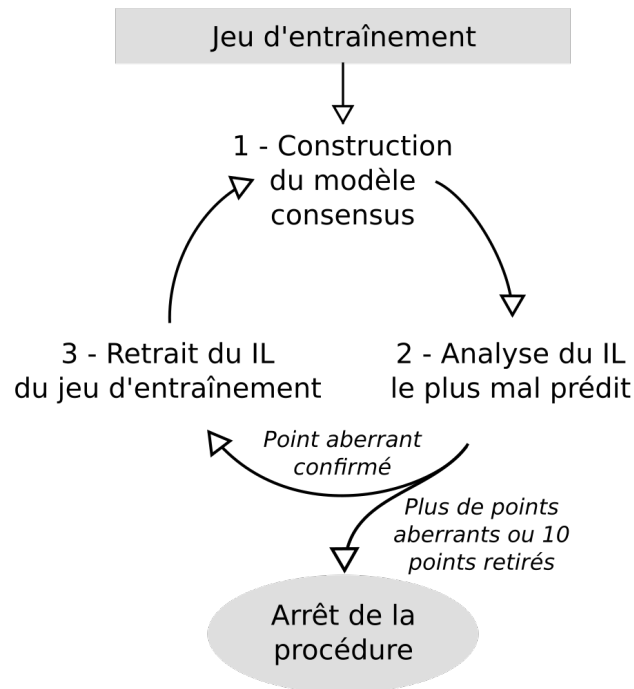


FIGURE 4.3 – Procédure de retrait des points aberrants pour les jeux d’entraînement.

Si les résultats obtenus en validation croisée sont en accord avec ces erreurs expérimentales, il n’en est pas toujours de même avec les RMSE obtenues sur les deux jeux de validation sur lesquels nous avons travaillé. Ainsi, le jeu de validation Solvionic a une RMSE supérieure à ce qui était attendu pour la conductivité et la température de fusion, tandis que le jeu de validation Littérature a une RMSE plus grande pour la viscosité. Les MAE quant à elles, correspondent aux erreurs attendues au vu des erreurs expérimentales. Étant donné que la MAE est moins sensible à la présence de points aberrants dans les données, les écarts observés entre la RMSE et la MAE sont donc des indices quand à la présence d’anomalies dans les données.

	Conductivité (mS/cm)	Température de fusion (°C)	Viscosité (cP)
RMSE CV	1,17	38,69	76,40
RMSE Littérature DA	1,15	23,87	90,83
RMSE Littérature DA nettoyé	1,15	23,87	69,12
RMSE Solvionic DA	2,70	47,24	54,78
RMSE Solvionic DA nettoyé	1,57	37,22	54,78
MAE CV	0,88	29,13	42,91
MAE Littérature DA	1,03	23,81	70,40
MAE Littérature DA nettoyé	1,03	23,81	56,71
MAE Solvionic DA	2,02	45,57	47,68
MAE Solvionic DA nettoyé	1,35	26,35	47,68

Tableau 4.5 – RMSE et MAE obtenues avec les modèles consensus de la SVM pour la conductivité, la température de fusion et la viscosité. Un domaine d’applicabilité (DA) de type contrôle de fragments a été appliqué. Les résultats notés « DA nettoyé » correspondent aux résultats obtenus après le retrait des points aberrants dans les jeux de validation.

Nous avons analysé les liquides ioniques ayant les plus grandes erreurs de prédiction dans les différents jeux de validation. Nous présentons dans les figures 4.4 pour la conductivité, 4.5 pour la température de fusion et 4.6 pour la viscosité, les graphes exp/pred ainsi que les détails des points aberrants identifiés.

Nous observons que le seul point mal prédit dans les jeux de données Littérature concerne le jeu de la viscosité. Il est lié à un échantillon dont la teneur en eau est supérieure à 500 ppm, seuil initialement choisi lors de la préparation des jeux d'entraînement (« $w\% = 0,0875$ » [189], ce qui correspond à 875 ppm, liquide n°1 dans le tableau 4.6).

Pour les liquides ioniques mal prédits des jeux de données Solvionic, nous avons relevé plusieurs cas de figure. Nous avons d'abord déterminé qu'un point se situait hors du domaine d'applicabilité du modèle en terme d'intervalles de valeurs : il s'agit du point n°1 dans le tableau 4.4. Il possède une conductivité de 13,85 mS/cm alors que la valeur de conductivité la plus haute relevée dans le jeu d'entraînement est de 8,6 mS/cm. Pour ce cas, le modèle se retrouve incapable d'extrapoler. Il arrive cependant à le prédire dans les valeurs de haute conductivité puisqu'il lui attribue une conductivité de 7,33 mS/cm.

Le second point mal prédit du jeu Solvionic pour la conductivité met en évidence une falaise d'activité potentielle. En effet, avec le remplacement d'un méthyl par un éthyl, la conductivité du liquide ionique est doublée. Si on compare le Tf₂N de 1-éthyl-1-butylpyrrolidinium au Tf₂N de 1-méthyl-1-butylpyrrolidinium (voir le tableau 4.6), on observe une grande différence de conductivité pour une modification mineure de la structure du cation (de 2,14 mS/cm pour le Tf₂N de 1-méthyl-1-butylpyrrolidinium à 8,58 mS/cm pour le Tf₂N de 1-éthyl-1-butylpyrrolidinium). Étant donné que nous observons la même tendance lorsque l'on remplace la chaîne butyl par une chaîne de type allyl (de 3,70 mS/cm pour le Tf₂N de 1-méthyl-1-allylpyrrolidinium à 7,4 mS/cm pour le Tf₂N de 1-éthyl-1-allylpyrrolidinium), nous supposons qu'il y a ici une falaise d'activité, et qu'il n'y a pas eu d'erreur de mesure.

Le troisième liquide ionique du tableau 4.4 met en évidence un défaut dans le modèle. En effet, quand nous étudions les différents Tf₂N de n-méthylimidazolium, avec la chaîne allant d'un groupement méthyl à un groupement pentyl (voir figure 4.7), nous observons une allure différente entre ce qui est prédit et ce qui est mesuré expérimentalement. Les deux séries ont une évolution linéaire, mais les prédictions obtenues pour le méthyl- et l'éthyl-méthylimidazolium sont plus basses que les valeurs expérimentales. Nous faisons l'hypothèse que le modèle, pour être le plus général possible, fait un compromis entre tous les composés du jeu d'entraînement. Dans ce cas, comme l'éthyl-méthylimidazolium peut être associé à des valeurs de conductivité basse (par exemple 0,82 mS/cm pour l'éthyl-méthylimidazolium 2-cyanométhyl-1,1,3,3-tétracyanoallyl), le modèle a sous-estimé la conductivité de ce liquide ionique.

Un liquide ionique a été identifié comme étant hautement hygroscopique dans la littérature. Il s'agit du numéro 4 dans le tableau 4.4. Étant donné qu'il est constitué de l'ion BF₄⁻, ceci n'est pas surprenant, car les liquides contenant cet anion le sont souvent [190]

Pour les liquides ioniques 4 du tableau 4.4 et 1 du tableau 4.5, les valeurs prédites ne convergent pas vers les valeurs expérimentales. Dans ces cas, les valeurs prédites sont plus proches de valeurs trouvées dans la littérature que des mesures expérimentales mesurées par notre équipe. Nous pensons que le problème pourrait être la quantité d'eau de l'échantillon utilisé pour faire les mesures.

Finalement, les mauvais résultats obtenus sur les liquides ioniques 2 et 3 du tableau 4.5 peuvent s'expliquer par le fait que les cations de ces liquides ioniques n'apparaissent pas dans le jeu d'entraînement.

Sans ces points aberrants, les modèles ont des performances qui sont comparables à celles de la validation croisée.

$R_1 \setminus R_2$	Methyl	Ethyl
Butyl	2,14 mS/cm [192]	8,58 mS/cm [193]
Allyl	3,70 mS/cm [194]	7,4 mS/cm [194]

Tableau 4.6 – Comparaison du Tf_2N de 1-méthyl-1-butylpyrrolidinium, du Tf_2N de 1-éthyl-1-butylpyrrolidinium, du Tf_2N de 1-méthyl-1-allylpyrrolidinium et du Tf_2N de 1-éthyl-1-allylpyrrolidinium.

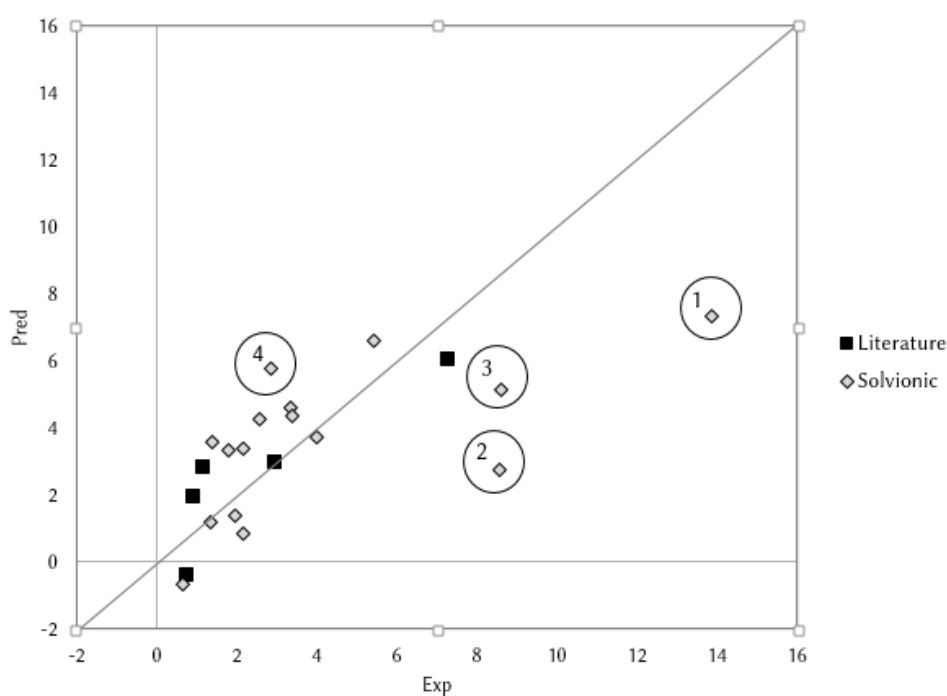
4.2.4 Mise en ligne des modèles

Les modèles SVR consensus développés pour cette étude sont disponibles et utilisables par tous sur notre serveur : infochim.u-strasbg.fr/webserv/VSEngine.html

Après la création d'un compte, l'utilisateur peut choisir de dessiner un liquide ionique à modéliser dans l'interface de dessin prévue à cet effet, ou bien d'importer un fichier sdf comportant un ou plusieurs liquides ioniques (voir figure 4.8). Cet outil traite les liquides ioniques comme des mélanges. Il faut donc cocher la case « mixture » dans l'interface. Après cette étape, l'utilisateur peut vérifier les structures, puis choisir de calculer la conductivité, la température de fusion et/ou la viscosité. À la fin de la procédure, pour chaque propriété, l'utilisateur peut choisir de voir les résultats dans une nouvelle page HTML ou de les télécharger dans un fichier csv. Pour chaque propriété est proposé une prédiction consensus sans domaine d'applicabilité, et si possible une prédiction avec l'application du domaine d'applicabilité de contrôle de fragments. L'écart type des deux prédictions est donné, de même que le nombre de modèles utilisés pour calculer la prédiction en tenant compte du domaine d'applicabilité.

4.3 Modélisation avec la TRR

Après avoir construit des modèles SVR consensus pour la modélisation des liquides ioniques, nous avons décidé de tester la TRR sur ces données. Le but ici était d'évaluer les performances de la TRR avec un cas réel.



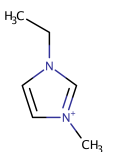
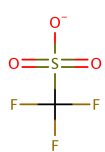
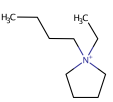
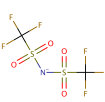
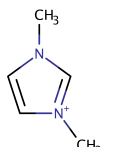
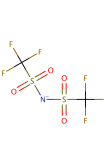
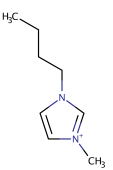
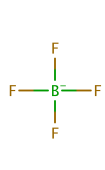
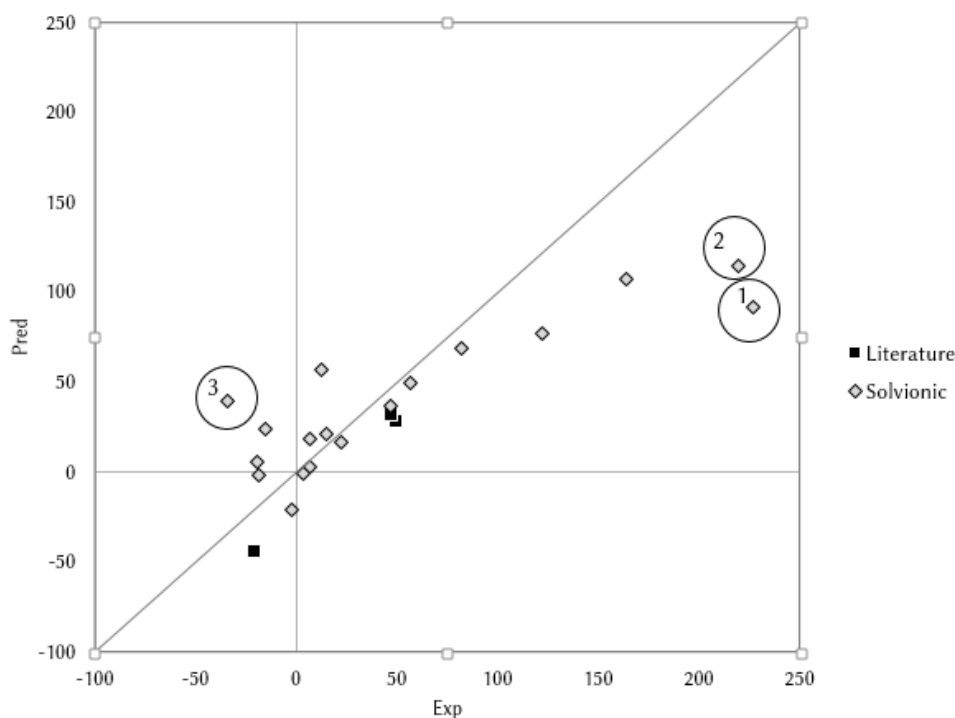
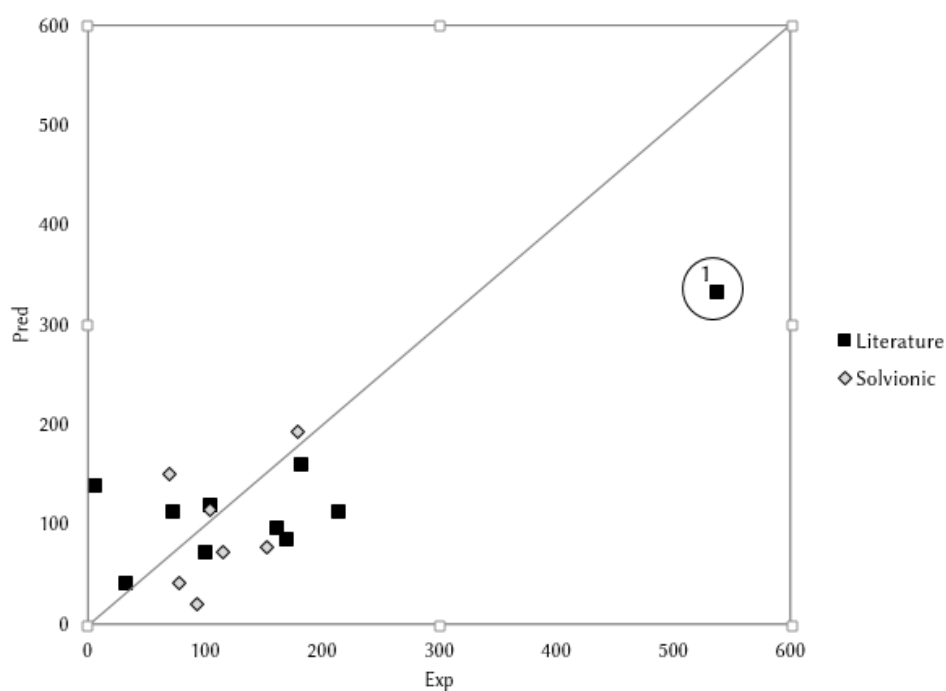
N° sur le graphe exp/pred	Dataset	Cation Structure	Anion Structure	Valeur Exp. (mS/cm)	Valeur Pred. (mS/cm)	Problème détecté
1	Solvionic			13,85	7,33	Hors de l'intervalle de valeurs du train
2	Solvionic			8,58	2,76	Falaise d'activité
3	Solvionic			8,61	5,13	Défaut du modèle
4	Solvionic			2,85	5,75	Littérature : 3,52 mS/cm [170] et 4,3 mS/cm [191] (w% < 0,03). Considéré comme étant hautement hygroscopique [190]

FIGURE 4.4 – Courbe exp/pred obtenue pour les jeux de validation de la conductivité (en haut), et tableau présentant les points aberrants détectés sur la courbe exp/pred (en bas). Les points aberrants proviennent tous du jeu de validation Solvionic. Les problèmes identifiés sont les suivants : valeur expérimentale située hors de l'intervalle de prédiction du modèle, falaise d'activité, défaut du modèle et composé hygroscopique.



N° sur le graphe exp/pred	Dataset	Cation Structure	Anion Structure	Valeur Exp. (°C)	Valeur Pred. (°C)	Problème détecté
1	Solvionic			227,00	98,04	Sur le site web ChemSpider les températures tendent vers 160 °C
2	Solvionic			220	114,5	Le cation est absent du jeu d'entraînement
3	Solvionic			-33,9	39,9	Le cation est absent du jeu d'entraînement

FIGURE 4.5 – Courbe exp/pred obtenue pour les jeux de validation de la température de fusion (en haut), et tableau présentant les points aberrants détectés sur la courbe exp/pred (en bas). Encore une fois, les points aberrants détectés proviennent tous du jeu de validation Solvionic. Pour deux des liquides ioniques identifiés, le problème vient d'un défaut dans le domaine d'applicabilité. Peu ou pas de cations présents dans le jeu d'entraînement possédait une structure leur ressemblant. Pour le liquide ionique restant, la valeurs prédite est plus proche de valeurs trouvées dans la littérature que des mesures expérimentales mesurées par notre équipe.





N° sur le graphe exp/pred	Dataset	Cation Structure	Anion Structure	Exp. Value (cP)	Pred. Va- lue (cP)	Problème détecté
1	Literature			538,9	324,85	%w > 500 ppm [189]

FIGURE 4.6 – Courbe exp/pred obtenue pour les jeux de validation de la viscosité (en haut), et tableau présentant les points aberrants détectés sur la courbe exp/pred (en bas). La première colonne correspond au point entouré sur la courbe exp/pred. Le point aberrant identifié provient du jeu de validation Littérature et correspond à un échantillon dont la teneur en eau est supérieure au seuil fixé pour la construction des jeux d’entraînement.

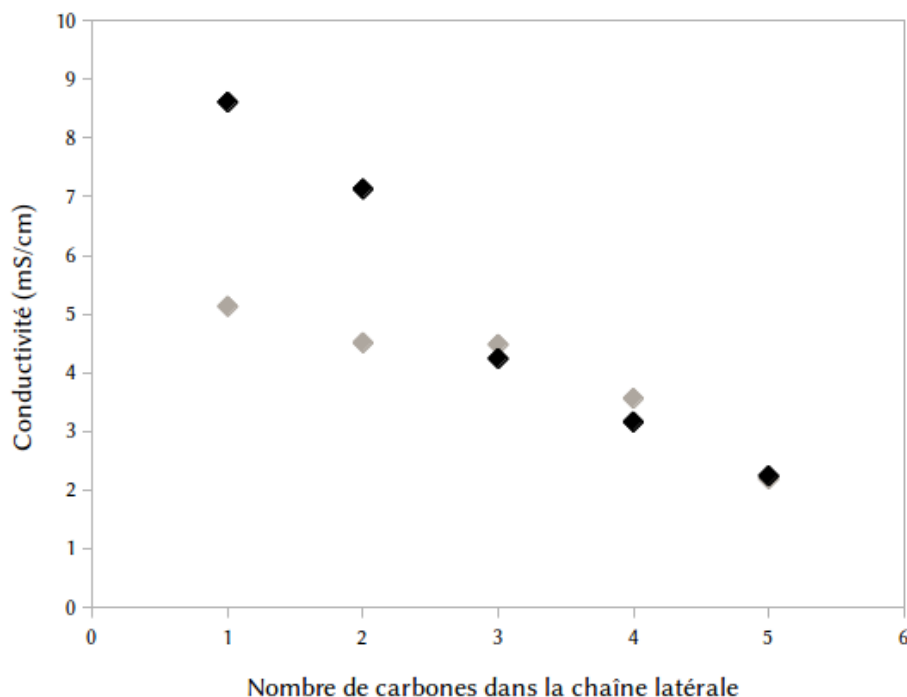


FIGURE 4.7 – Conductivité de différents liquides ioniques de type Tf_2N de n-méthylimidazolium en fonction du nombre de carbones de la chaîne latérale, avec la chaîne allant de $n=1$ à $n=5$. Les points noirs correspondent aux valeurs expérimentales, les points gris aux valeurs prédites.

4.3.1 Procédure

L'étude avec la RR s'est faite de la même façon que pour la SVR (voir le protocole en figure 4.2). Nous avons fait, pour chaque espace de descripteurs, une optimisation de g à l'aide d'une 5×5 -CV. Nous avons ensuite sélectionné de la même façon les modèles à utiliser pour le consensus (tri par RMSE, sélection du premier modèle et de ceux dont la RMSE est comparable selon un test de Student à 95 %). Nous avons finalement construit les modèles individuels et construit un modèle consensus correspondant à la moyenne arithmétique de l'ensemble des prédictions des différents modèles.

En ce qui concerne les modèles TRR, nous avons pris la liste des modèles RR sélectionnés avec leur g associé, puis nous avons optimisé gp à l'aide d'une validation croisée. Compte tenu des observations faites lors de l'étude méthodologique de la TRR, nous avons divisé le gp obtenu par 10 pour obtenir le gp heuristique avant de construire les modèles individuels, puis le modèle consensus.

Nous avons comparé les différentes méthodes de différentes façons. Nous avons fait une première comparaison en nous basant sur les modèles consensus. La seconde comparaison des modèles s'est faite en analysant les différents ensembles de modèles individuels utilisés pour construire les modèles consensus. Enfin, nous avons comparé la SVR, la RR et la TRR en

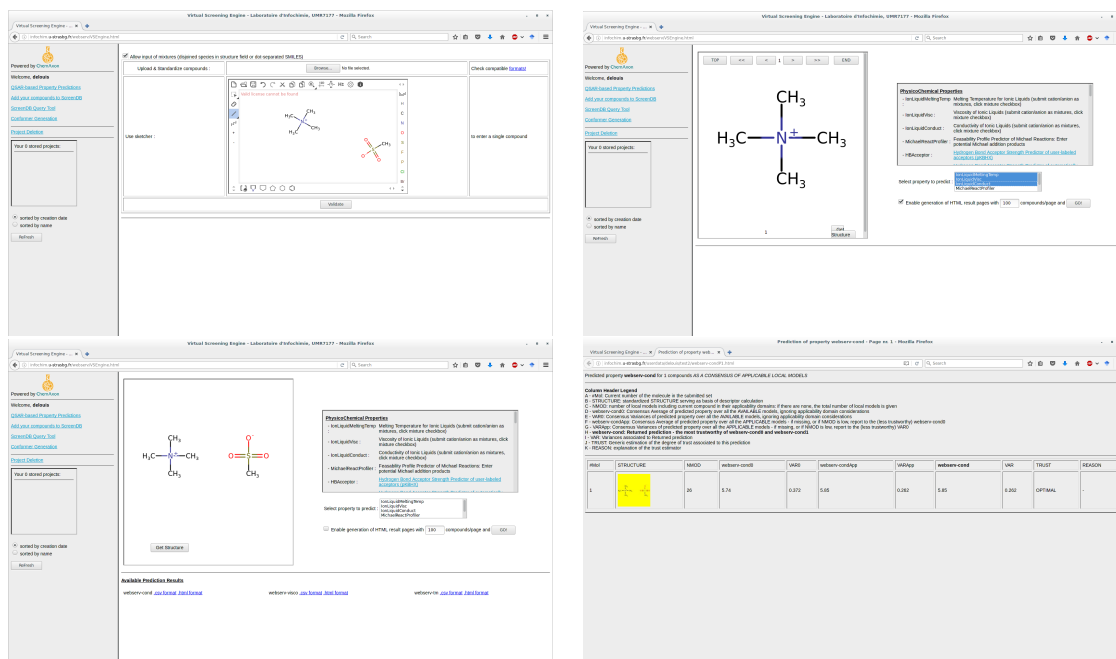


FIGURE 4.8 – Captures d’écran de l’application web. La première étape, illustrée en haut à gauche, consiste à préciser les structures, soit en utilisant le sketcher, soit en incluant un fichier sdf. La seconde étape, illustrée en haut à droite, est de sélectionner les modèles que l’on veut appliquer. L’image en bas à gauche illustre la page indiquant que les calculs sont en cours, et l’image en bas à droite montre la page HTML des résultats obtenus

utilisant, pour une propriété donnée, un modèle individuel avec un espace de descripteurs fixé pour les trois méthodes.

4.3.2 Modèles consensus

Nous avons commencé par regarder les résultats obtenus avec les modèles consensus. Nous avons regardé les RMSE de la SVR, de la RR et de la TRR, ainsi que l’effet transductif pour chaque propriété pour la validation croisée, et les jeux de validation Littérature et Solvionic avec et sans application du domaine d’applicabilité. Les résultats sont regroupés dans le tableau 4.7.

Pour la conductivité, les résultats obtenus sont à peu près du même ordre de grandeur entre les 3 méthodes pour tous les jeux de données. La SVR a de meilleures performances en validation croisée et pour le jeu Littérature avec domaine d’applicabilité. La TRR surpasse la RR et la SVR pour les jeux Littérature et Solvionic avec domaine d’applicabilité. Enfin, la RR est la meilleure méthode pour prédire le jeu de données Solvionic. Nous observons également un effet transductif assez élevé pour le jeu de test Littérature (12,56 %), négatif pour le jeu Solvionic (-3,56 %) et proche de 0 % pour les trois autres cas.

Pour la température de fusion, les RMSE sont du même ordre de grandeur pour les trois modèles sur l’ensemble des jeux de données, exception faite du jeu de validation Littérature

pour lequel la RR et la TRR sont surpassées par la SVR de 20 °C. Les trois modèles sont donc globalement comparables. Les effets transductifs observés ici sont positifs pour la validation croisée et le jeu de validation Littérature avec application de domaine d'applicabilité, proches de zéro pour le jeu de validation Solvionic avec domaine d'applicabilité, et négatifs pour le jeu de validation Littérature et le jeu de validation Solvionic.

Enfin, en ce qui concerne la viscosité, les résultats obtenus sont globalement du même ordre de grandeur entre les trois méthodes pour l'ensemble des jeux de données. La SVR se place en tête pour les deux jeux de validation avec domaine d'applicabilité, la TRR est la méthode la plus performante pour la validation croisée et le jeu de validation Solvionic, tandis que la RR a la RMSE la plus basse pour le jeu de validation Littérature. En ce qui concerne les effets transductifs relevés, il est positif pour le jeu de validation Solvionic (4,01 %), nul pour la validation croisée et les jeux de validation avec domaine d'applicabilité. Enfin, il est négatif pour le jeu de données Littérature (-7,34 %).

Conductivité (mS/cm)

	CV	Littérature	Lit. avec DA	Solvionic	Sol. avec DA
SVR	1,17	2,47	1,15	2,71	2,70
RR	1,55	2,51	1,57	2,62	2,66
TRR	1,54	2,19	1,56	2,71	2,66
TE(%)	0,06	12,56	0,19	-3,56	0,08

Température de fusion (°C)

	CV	Littérature	Lit. avec DA	Solvionic	Sol. avec DA
SVR	38,69	64,03	23,87	63,08	47,24
RR	33,20	84,10	32,65	63,02	49,96
TRR	32,82	84,96	29,55	63,41	49,80
TE (%)	1,15	-1,03	9,49	-0,61	0,32

Viscosité (cP)

	CV	Littérature	Lit. avec DA	Solvionic	Sol. avec DA
SVR	76,40	129,51	90,83	128,53	54,78
RR	74,49	128,21	108,83	111,06	56,37
TRR	74,32	137,62	109,07	106,61	56,10
TE (%)	0,22	-7,34	-0,22	4,01	0,48

Tableau 4.7 – RMSE des modèles consensus SVR, RR et TRR relevées pour la conductivité, la température de fusion et la viscosité.

4.3.3 Comparaison de l'ensemble des modèles individuels

Pour comprendre ce qui pouvait impacter ces effets transductifs, nous avons fait une étude des modèles individuels. Le tableau 4.8 reporte l'effet transductif moyen mesuré sur l'ensemble des modèles individuels, tandis que le tableau 4.9 reporte les intervalles des valeurs d'effets transductifs ainsi que le pourcentage d'effets transductifs supérieurs strict à 0. Ici, nous nous concentrons sur la validation croisée et sur les jeux de test sans domaine d'applicabilité.

En ce qui concerne les effets transductifs moyens (tableau 4.8), nous constatons qu'ils sont tous positifs ou proches de 0 %. Aucun effet transductif négatif, aussi faible soit il, n'a été relevé. La plupart de ces effets transductifs restent néanmoins inférieur à 1,5 %, ce qui signifie qu'en moyenne les performances de la TRR et de la RR sont comparables. On notera toutefois le cas du test Littérature pour la température de fusion, pour lequel on relève un effet transductif égal à 15,49 %. Souvent, les performances ont été améliorées pour les cations contenant un cycle cyclopropénium. Ceci peut sembler surprenant car cette structure est, en fait, non présente dans le jeu d'entraînement. Il s'agit d'une autre illustration de l'effet positif de la transduction pour des molécules situées hors du domaine d'applicabilité du modèle non transductif.

Observons maintenant les intervalles des valeurs d'effets transductifs relevés sur les modèles individuels (tableau 4.9). On constate que la majorité des cas ont des intervalles de valeurs assez faible, avec des amplitudes entre les effets transductifs minimum et maximum généralement inférieur à 2 %. La proportion de modèles pour lesquels l'effet transductif est strictement positif dépasse les 60 % pour tous les modèles, et atteint même 100 % des modèles pour la validation croisée de la température de fusion. On notera de nouveau l'exception du jeu de validation Littérature de la température de fusion, pour lequel la différence entre les effets transductifs minimum et maximum est de 61,8 %. On relève un effet transductif maximal à 44,9 %, et le nombre de modèles pour lesquels l'effet transductif est strictement positif est de 87,5 %.

	Validation croisée	Littérature	Solvionic
Conductivité (mS/cm)	0,05	0,07	0,12
Température de fusion (°C)	1,37	15,49	0,32
Viscosité (cP)	0,26	0,11	0,06

Tableau 4.8 – Moyenne des TE pour la conductivité, la température de fusion et la viscosité. Les TE moyens sont tous positifs et assez faibles, exception faite du jeu de validation pour la température de fusion, pour lequel le TE est de 15,49 %. Pour ce jeu de données, les performances ont été améliorées pour les cations contenant un cycle cyclopropénium, cation absent du jeu d'entraînement. Dans ce contexte, les composés hors du domaine d'applicabilité ont fortement bénéficié de la transduction.

	CV	Littérature	Solvionic
Conductivité (mS/cm)	[-0,10 ; 0,27], 78,6 %>0	[-0,34 ; 1,57], 64,3 %>0	[-0,02 ; 0,38], 96,4 %>0
Température de fusion (°C)	[0,19 ; 2,85], 100 %>0	[-16,9 ; 44,9], 87,5 %>0	[-0,15 ; 2,91], 66,7 %>0
Viscosité (cP)	[0,05 ; 0,87], 100 %>0	[-0,14 ; 0,68], 73,3 %>0	[-0,33 ; 0,41], 75,6 %>0

Tableau 4.9 – Intervalle des valeurs de TE et le pourcentage de TE supérieurs à zéro pour la conductivité, la température de fusion et la viscosité. La majorité des cas ont des intervalles de valeurs assez faible, avec des amplitudes entre le TE minimum et le TE maximum généralement inférieur à 2 %. La proportion de modèles ayant un TE strictement positif dépasse les 60 % pour tous les modèles, et atteint même 100 % des modèles pour la validation croisée de la température de fusion. Le jeu de validation Littérature de la température de fusion, est la seule exception : amplitude entre le TE minimum et le TE maximum de 61,8 % et TE maximal à 44,9 %.

4.3.4 Comparaison de la SVR, de la RR et de la TRR pour un espace de descripteurs fixé

Nous avons ensuite comparé, pour un ensemble de descripteurs fixé, les modèles obtenus pour la SVR, la RR et la TRR. Pour être sélectionné, l'ensemble de descripteurs doit avoir été utilisé dans les modèles consensus des trois méthodes de modélisation, et avoir obtenu des performances satisfaisantes sur les trois méthodes. Les résultats sont consignés dans le tableau 4.10.

Pour la conductivité, la TRR et la RR ont des RMSE identiques, excepté pour le jeu de validation de la littérature où la RR est très légèrement supérieure, ce qui rend ces méthodes comparables en terme de qualité de prédiction. Si les performances de la SVR sont comparables pour la validation croisée, ce n'est pas le cas pour les jeux de validation : on passe, sur le jeu Littérature, de 1,97 mS/cm pour les méthodes RR et TRR à 2,79 mS/cm pour la SVR (2,47 mS/cm pour le modèle SVR consensus), tandis que pour le jeu de validation Solvionic on passe de 8,87 mS/cm à 3,15 mS/cm (2,70 mS/cm pour le modèle SVR consensus).

Pour la température de fusion, les performances des trois méthodes sont comparables. Ainsi, si numériquement parlant la TRR surpasse la RR et la SVR pour la validation croisée et le jeu de validation Littérature, en pratique les différences entre les modèles sont assez minimes pour que l'on puisse utiliser indifféremment l'une ou l'autre des méthodes.

Enfin, pour la viscosité, nous avons également des résultats d'ordre de grandeur comparable entre les trois méthodes. Cependant, on observe sur les jeux de validation une différence de 5 cP suffisamment nette pour qu'il reste plus intéressant d'utiliser, pour la Littérature, les modèles RR et TRR, et pour Solvionic le modèle SVR. De nouveau, si la TRR surpasse numériquement la RR, en pratique le gain de performance est trop faible pour différencier les performances de ces modèles.

4.4 En résumé

Nous avons modélisé la conductivité, la température de fusion et la viscosité d'un premier type de solvant : les liquides ioniques. Pour cela, nous avons trois jeux de données par pro-

Conductivité (mS/cm, fragmentation IA(2-2)_FC)

	CV	Littérature	Solvionic
SVR	1,42	2,79	3,15
RR	1,41	1,97	2,87
TRR	1,41	1,96	2,87

Température de fusion (°C, fragmentation IIAB(2-3)_R)

	CV	Littérature	Solvionic
SVR	37,52	79,83	51,94
RR	38,25	81,11	52,20
TRR	37,36	78,31	52,19

Viscosité (cP, fragmentation IAB(3-8))

	CV	Littérature	Solvionic
SVR	75,69	134,39	112,60
RR	73,61	129,88	117,41
TRR	73,55	129,79	117,33

Tableau 4.10 – RMSE des modèles SVR, RR et TRR relevées pour la conductivité, la température de fusion et la viscosité. Pour chaque propriété, les RMSE d’un seul modèle représentatif ont été indiquées (pour la conductivité il s’agit de séquences d’atomes de longueur 2 atomes avec ajout de charges formelles, pour la température de fusion d’atomes centrés étendus comprenant les atomes et les liaisons de longueur variant entre 2 et 3 atomes, et pour la viscosité de séquences d’atomes et de liaisons de longueur entre 3 et 8 atomes). Globalement les trois méthodes d’apprentissage donnent des résultats comparables.

priété : un jeu d'entraînement dont les données, issues de la littérature, ont été soigneusement sélectionnées en fonction de la pureté de l'échantillon (notamment en ce qui concerne la quantité d'eau), un jeu de validation Littérature, dont les données sont issues de la littérature récente, et un jeu de validation Solvionic sur lequel nous avons travaillé en collaboration avec l'IPHC et la société Solvionic.

Ces différentes propriétés ont été modélisées grâce à trois méthodes d'apprentissage automatique : la SVR, la RR et la TRR. Les modèles consensus SVR se sont avérés efficaces pour détecter des points aberrants. Les RMSE observées pour les modèles SVR consensus sont du même ordre de grandeur que les valeurs expérimentales. Ces modèles sont disponibles en ligne, sur le serveur du laboratoire : infochim.u-strasbg.fr/webserv/VSEngine.html. Les résultats obtenus avec la TRR sont assez comparables à ceux de la RR et de la SVR.

Chapitre 5

Modélisation des électrolytes

Nous nous sommes intéressée aux électrolytes de batteries Li-ion. Il s'agit d'un travail fait en collaboration avec en collaboration avec Laurent Joubert de l'université de Rouen, Alexandre Chagnes de Chimie ParisTech et Jean-Marie Tarascon du Collège de France au sein de l'ANR DEVEGA [9]. Le but de ce projet est la découverte de nouveaux électrolytes liquides pour la conception d'une nouvelle génération de batteries. L'application visée ici est la batterie pour voitures électriques. Comme pour la modélisation des liquides ioniques, nous avons fait notre modélisation en deux étapes : une première modélisation avec la SVR pour pouvoir rapidement transmettre des suggestions de molécules, et une seconde phase où l'on teste la TRR.

5.1 Données

Notre base de données nous a été transmise par nos collaborateurs de Chimie ParisTech. Elle contient une centaine de composés répartis en plusieurs familles chimiques : carbonates, esters, éthers, sulfones et sulfinyles. Ces composés sont destinés à entrer dans la composition de solvants électrolytiques pour une nouvelle génération de batteries au lithium. Toutefois, par facilité de langage, nous les désignerons simplement comme étant des électrolytes. Pour chacune de ces molécules, des données sur la conductivité, la constante diélectrique, le potentiel d'oxydation, la température d'ébullition, la température de fusion, la température finale de la cellule électrochimique et la viscosité ont été collectées. Toutefois, ces propriétés ne sont pas disponibles toutes ensemble pour une même molécule : le profil de propriété contient des valeurs manquantes. Ces données proviennent soit de la littérature, soit des mesures expérimentales faites par nos collègues. Dans le cadre de cette étude, nous allons nous focaliser sur ces 6 propriétés : la conductivité, la constante diélectrique, le potentiel d'oxydation, la température d'ébullition, la température de fusion et la viscosité. Le nombre de données disponibles pour chacune des propriétés est illustré dans la figure 5.1. Certaines propriétés (potentiel d'oxydation et conductivité) possèdent un nombre de valeurs supérieur au nombre d'électrolytes. Cela provient du fait qu'une même molécule peut avoir été testée dans différentes

conditions expérimentales. Au contraire, le nombre de données enregistrées sur certaines propriétés est très faible comparé aux données habituellement disponibles dans la littérature. Par exemple, la température de fusion rassemble 60 mesures alors que cette propriété est fréquemment étudiée en chémoinformatique sur des jeux de données rassemblant typiquement plus de 70 000 points. Ceci est dû au fait que dans ce projet, les données couvrent un espace chimique restreint, propre aux applications technologiques pour les batteries Li-ion.

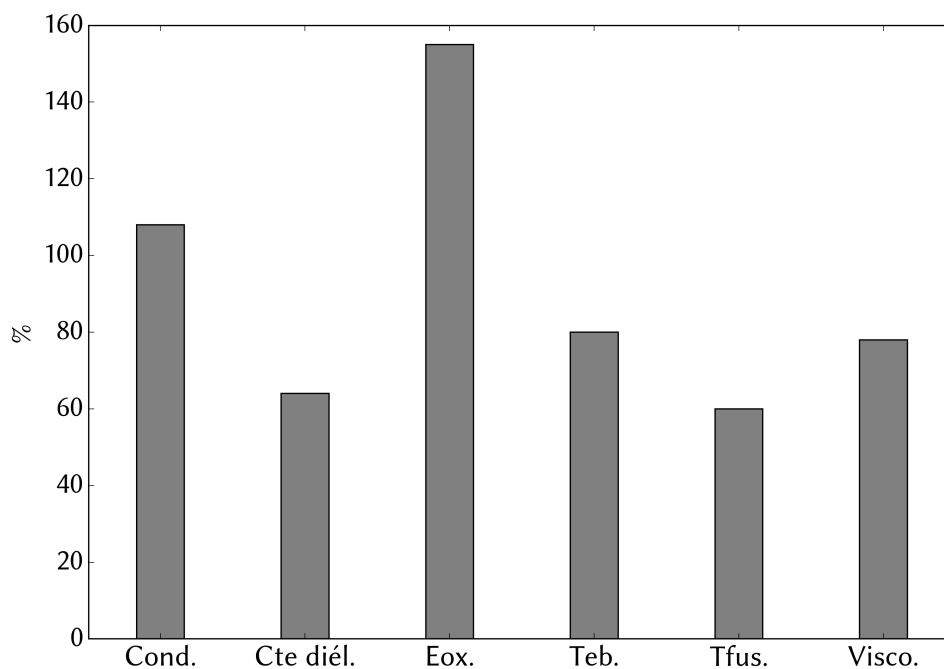


FIGURE 5.1 – Représentation graphique de la répartition du nombre de valeurs dans nos jeux de données. Nous indiquons ici le nombre de données utilisées pour chacune des propriétés étudiées (conductivité, constante diélectrique, potentiel d’oxydation, température d’ébullition, température de fusion et viscosité). Certaines propriétés (potentiel d’oxydation et conductivité) possèdent un nombre de valeurs supérieur au nombre d’électrolytes. Cela provient du fait qu’une même molécule peut avoir été testée dans différentes conditions expérimentales.

Les mesures de la conductivité répertoriées dans la base de données ont été faites dans diverses conditions expérimentales. L’électrode utilisée est toujours une électrode de platine. Les sels suivants sont utilisés : LiPF_6 , LiBF_4 , LiClO_4 et LiTFSI (TFSI correspondant à l’ion Tf_2N chez les liquides ioniques). La concentration en sels est fixée à 1 M. Le sel le plus courant dans les batteries Li-ion est le LiPF_6 . Toutefois, dans d’autres types de batteries telles que les batteries Li-S, d’autres sels sont aussi utilisés, en particulier le LiTFSI , c’est pourquoi il a été possible de comparer les mesures de conductivité en échangeant la nature du sel. La reproductibilité des mesures en échangeant ces sels est estimée à 0,4 mS/cm (voir figure 5.2). Cette étude est comparable aux variations observées quand une mesure publiée a été reproduite expérimentalement. Ainsi, les données de conductivité concernant les sels LiPF_6 et LiTFSI ont été fusionnées pour l’étude QSPR.

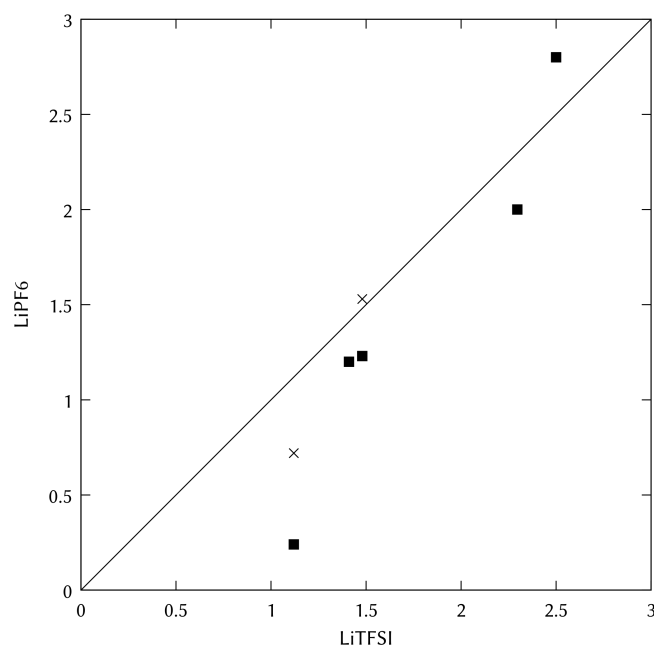


FIGURE 5.2 – Corrélation entre la conductivité de divers électrolytes en présence de LiTFSI et de LiPF₆. Chaque point correspond à un électrolyte différent. Les carrés correspondent à des mesures répertoriées dans la littérature. Les croix correspondent à des mesures expérimentales faites par nos collaborateurs de Chimie ParisTech.

Pour mesurer le potentiel d'oxydation (E_{ox}), plusieurs conditions expérimentales différentes sont possibles. L'électrode utilisée pour réaliser les mesures peut être en platine, carbone vitreux, or, LiCoO₂, Li_{1-x}Mn₂O₄ ou LiCr_{0,015}Mn_{1,985}O₄. Les sels utilisés sont LiPF₆, LiBF₄, LiClO₄ et LiTf₂N. Dans la plupart des cas, la concentration en sel est de 1 M et la tension est de 5 mV/s. Cette fois encore, il a été possible de comparer l'effet des conditions expérimentales sur la mesure du potentiel d'oxydation. Il a été observé que la nature de l'électrode joue un rôle important sur la mesure. Par exemple, pour un même sel (LiTFSI), lorsque l'on échange une électrode en platine par une électrode en carbone vitreux (*glassy carbon* en anglais), les potentiels d'oxydation ne corrèlent que médiocrement (avec un coefficient de détermination de 0,32). Toutefois, suivant le conseil et l'expertise de notre collègue Alexandre Chagnes de Chimie Paris Tech, nous avons fusionné les données LiTFSI et LiPF₆ mesurées sur platine.

Les données de la température d'ébullition et de la température de fusion sont acquises dans la littérature exclusivement. La mesure de la température de fusion a généralement été faite avec un banc Kofler.

La constante diélectrique et la viscosité sont des propriétés pouvant principalement être affectées par la température et la pression. Les données utilisées sont donc restreintes ici à 25 °C et 1 atmosphère. Pour ces deux propriétés, il est plus simple de travailler avec leur logarithme. Le logarithme de la viscosité permet, en théorie [195, 196], d'obtenir un modèle additif. La dépendance en température de la viscosité est bien représentée par un modèle exponentiel. Aussi le logarithme de la viscosité est, au sens thermodynamique, interprété

comme une quantité extensive, et donc, est potentiellement plus simple à modéliser par un modèle QSPR additionnant des contributions de différentes fonctions chimiques. Quant à la constante diélectrique, elle est traitée par analogie aux approches utilisées pour modéliser la constante diélectrique des matériaux composites. La plus connue des règles de mélange, la règle de Lichtenecker [197] propose que le logarithme de la constante diélectrique soit une combinaison linéaire des logarithmes des constantes diélectriques de ses constituants. Cela justifie de plutôt chercher une équation QSPR pour le logarithme de la constante diélectrique.

5.2 Modélisation avec la SVR

Pour les besoins de notre collaboration sur ce projet, nous avons commencé la modélisation des électrolytes en utilisant la SVR, comme ce fut le cas avec les liquides ioniques. Nous avons ainsi pu identifier quelques points aberrants, construire des modèles avec des performances raisonnables et faire des propositions de candidats pour nos collègues de Paris.

5.2.1 Procédure

Pour modéliser les électrolytes, nous avons utilisé des fragments ISIDA. Nous avons généré 175 espaces de descripteurs différents. Pour la SVR, le paramètre ϵ a été estimé à 0,01 pour la conductivité, la constante diélectrique, le potentiel d'oxydation et la viscosité. Cette valeur correspond au nombre de chiffres significatifs dans la base de données pour ces propriétés. Pour les températures d'ébullition et de fusion, ϵ a été fixé à 2. Ceci correspond à une estimation optimiste de la qualité des mesures estimées. La procédure suivie est illustrée en figure 5.3. Étant donné que nous n'avons, pour chaque propriété, qu'un seul jeu de données, nous avons choisi de faire un processus de double validation croisée. Nous avons fait une validation croisée interne pour optimiser C , puis une validation croisée externe pour tester le modèle. Ceci a été fait pour chaque espace de descripteurs différent. Nous avons donc construit 175 modèles différents. Nous avons retenu entre 9 et 19 modèles pour faire un modèle consensus final. La sélection des modèles s'est faite simplement en prenant les meilleurs modèles en terme de RMSE. Le résultat du consensus est obtenu en prenant la moyenne arithmétique des estimations de chaque modèle individuel.

5.2.2 Résultats

Les résultats obtenus sont compilés dans le tableau 5.1. Les modèles sont prédictifs, avec un R^2 compris entre 0,64 pour la conductivité et 0,94 pour le logarithme de la viscosité.

Les modèles de conductivité électrique et du potentiel d'oxydation sont les moins prédictifs : on note un R^2 de 0,64 pour la conductivité et de 0,68 pour le potentiel d'oxydation, ainsi qu'une RMSE de 2,13 mS/cm pour la première et de 0,48 V pour la seconde. Les performances de ces modèles sont limitées par la fusion de données acquises dans des conditions expérimentales

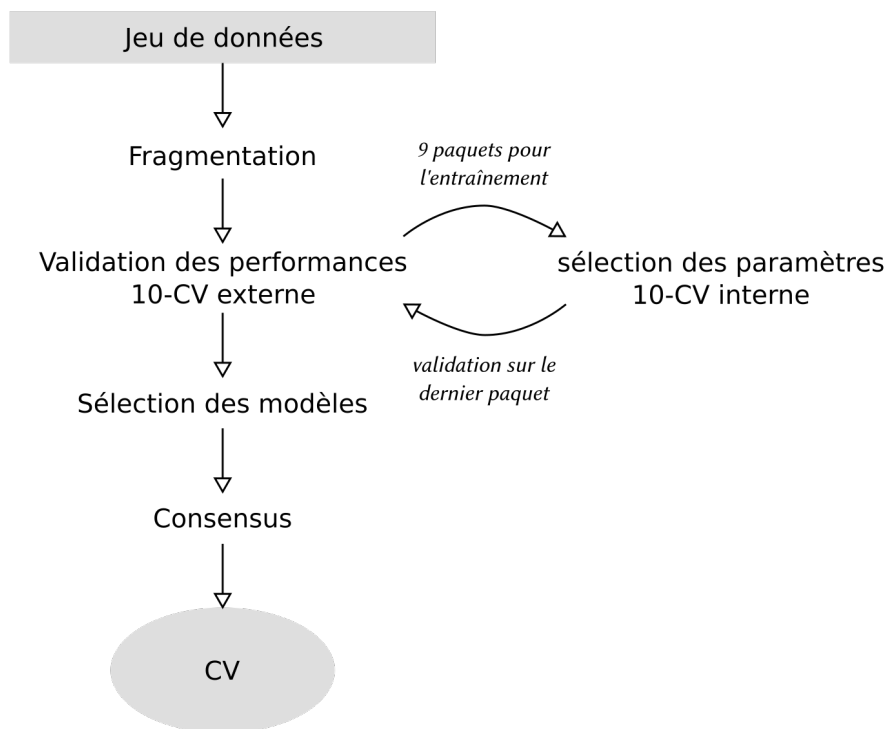


FIGURE 5.3 – Protocole suivi lors de la modélisation des électrolytes avec la SVR. Après la sélection des données, nous avons fait une validation croisée interne pour optimiser C . Nous avons ensuite fait une validation croisée externe pour évaluer les performances des modèles. Nous avons finalement choisi les meilleurs modèles SVR individuels afin de faire un modèle consensus.

différentes. Ces fusions de données sont justifiées par des observations empiriques mais sont aussi nécessaires : certaines données ne peuvent pas être acquises en présence de LiPF₆ et sont donc acquises en présence de LiTFSI. Elles participent ainsi à la définition précise de ces propriétés. Ceci augmente néanmoins le bruit présent dans nos données, et donc par conséquent la qualité des modèles est plus basse.

Les performances des modèles sur les logarithmes de la constante diélectrique et de la viscosité sont très bonnes, avec des R² de 0,87 et de 0,94 respectivement, et des RMSE de 0,17 log et 0,13 log(cP) respectivement. Toutefois, quand les estimations des modèles sont exprimées dans les unités natives de ces propriétés, ces performances (constante diélectrique : RMSE=22,57, R²=0,57 ; viscosité : RMSE=4,43 cP, R²=0,82) sont plus proches de celles qui peuvent être observées quand une équation QSPR est recherchée directement sur les valeurs non transformées de la constante diélectrique ou de la viscosité. L'intérêt de travailler avec le logarithme est donc limité.

Les données les plus compliquées à modéliser sont certainement les températures de fusion et d'ébullition. Les sources d'erreurs pour ces propriétés sont nombreuses, et certaines ont déjà été rencontrées lors de la modélisation des liquides ioniques. Les températures relevées dans la littérature concernent des substances dont la pureté n'est pas toujours établie. Par ailleurs, la température de fusion dépend de la phase cristalline du solide et peut être parfois confondue avec une température de transition entre deux phases cristallines ou avec une température de transition vitreuse. La température d'ébullition est aussi parfois difficile à détecter expérimentalement. S'ajoutent à ceci des phénomènes parasites, comme la surfusion, qui eux aussi contribuent à augmenter l'incertitude observée sur les valeurs expérimentales. La demi-largeur de l'intervalle de confiance autour des valeurs estimées est de l'ordre de 90 °C. Ceci ne se traduit pas vraiment dans le coefficient de détermination ou celui de corrélation, car ces erreurs sont comparées au domaine dans lequel ces températures sont mesurées et qui couvre de -150 °C à 330 °C.

Propriété	RMSE	R ²	Nb de modèles SVR dans le consensus
Conductivité (mS/cm)	2,13	0,64	19
Constante diélectrique (log(ε))	0,17	0,87	14
Potentiel d'oxydation (V)	0,45	0,68	13
Température d'ébullition (°C)	37,12	0,77	9
Température de fusion (°C)	32,56	0,73	11
Viscosité (log(cP))	0,13	0,94	19

Tableau 5.1 – RMSE et R² obtenus sur les modèles SVR consensus construits pour la conductivité, le logarithme de la constante diélectrique, le potentiel d'oxydation, la température d'ébullition et la température de fusion. Les modèles de conductivité électrique et du potentiel d'oxydation sont les moins prédictifs. Les performances des modèles sur les logarithmes de la constante diélectrique et de la viscosité sont très bonnes. Les données les plus compliquées à modéliser sont les températures de fusion et d'ébullition.

5.2.3 Points aberrants

Nous avons identifié pour chaque propriété les composés les plus mal prédits. Nous avons identifié entre 0 et 4 molécules potentiellement aberrantes selon la propriété. La liste de ces molécules est donnée dans les tableaux 5.2 et 5.3. Nous pouvons remarquer que ces points ne sont généralement dans le domaine d'applicabilité que de quelques modèles.

Ces données sont en cours d'investigation au niveau expérimental pour être confirmées ou infirmées, nous ne pouvons donc pas à l'heure actuelle confirmer que ces points sont des points aberrants. Étant donné que le caractère de ces points aberrants n'a pas été démontré, ils n'ont pas été retirés des jeux de données.

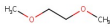

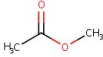

5.2.4 Predictor

ISIDA/Predictor est un outil développé au laboratoire par Guillaume Beck lors de son stage au sein du laboratoire (voir figure 5.4). Il peut être installé sur Linux, Mac et Windows. Il est chargé de calculer, pour une ou plusieurs nouvelles molécules, les descripteurs moléculaires et d'appliquer le modèle QSPR pour estimer les valeurs des propriétés de ces molécules. L'utilisateur reçoit les estimations de chaque modèle ainsi que la valeur consensus. Un domaine d'applicabilité de type contrôle de fragments est utilisé sur les modèles individuels. De plus, si moins de 10 % des modèles consensus participent au calcul de la prédiction, le composé est considéré comme étant hors du domaine d'applicabilité du modèle consensus.

Nous avons intégré dans le Predictor les modèles SVM utilisant les descripteurs ISIDA des 6 propriétés (conductivité, log de la constante diélectrique, potentiel d'oxydation, température d'ébullition, température de fusion et log de la viscosité). L'utilisateur choisit un fichier SDF à analyser et un modèle. Il reçoit les résultats sous forme de table mentionnant l'appartenance ou non de la molécule aux différents domaines d'applicabilité. Ces résultats sont également sauvegardés dans un fichier csv.

5.2.5 Criblage et sélection de candidats

Nous avons fabriqué une chimiothèque virtuelle de 15 045 esters et sulfones énumérés, à l'aide du logiciel Marvin Sketch [198], sur des châssis moléculaires proposés par l'équipe chargée de la synthèse. Nous avons filtré certaines structures chimiques posant des problèmes évidents tels que les peroxydes ou les acides carboxyliques. Ensuite, nous avons utilisé le logiciel Predictor afin de prédire les propriétés de ces molécules virtuelles. Nous avons filtré les résultats sur trois propriétés afin de ne garder que les structures intéressantes pour le développement des batteries. Ainsi, nous avons sélectionné les molécules possédant un potentiel d'oxydation supérieur à 3 V, une température d'ébullition supérieure à 50 °C et une température de fusion inférieure à 20 °C. Seules 3 832 molécules ont passé ce filtre.

Conductivité (mS/cm)				
Molécule	Structure	Valeur expérimentale	Valeur prédite	Nb de modèle (19)
1		11,09	9,22	1
2		2,80	7,81	1
3		14,76	8,01 – 9,83	15
4		9,11	3,37 – 6,70	2

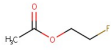
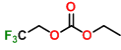
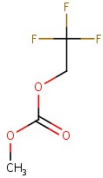
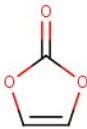
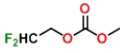
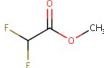
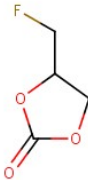

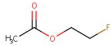
Constante diélectrique (log(ε))				
Molécule	Structure	Valeur expérimentale	Valeur prédite	Nb de modèle (14)
1		0,90	1,05 – 1,06	2
2		0,85	0,95	2
3		0,98	0,93 – 0,94	4
4		2,10	1,95	4

Tableau 5.2 – Points les plus mal prédits pour la conductivité et la constante diélectrique. Pour ces points, la différence entre la valeur expérimentale et la valeur prédite dépassait 3 RMSE pour plusieurs modèles individuels. Ces composés sont en cours d’investigation par nos collègues.

Potentiel d'oxydation (V)				
Molécule	Structure	Valeur expérimentale	Valeur prédite	Nb de modèle (13)
1		6,40	5,87	1
2		4,50	5,40 – 5,91	8
3		6,50	5,94	1
4		5,50	4,99 – 5	3

Température d'ébullition (°C)				
Molécule	Structure	Valeur expérimentale	Valeur prédite	Nb de modèle (11)
1		119	171,22	1

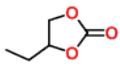
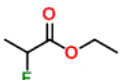
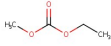
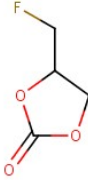
Viscosité (log(cP))				
Molécule	Structure	Valeur expérimentale	Valeur prédite	Nb de modèle (19)
1		0,49	0,43 – 0,44	4
2		-0,09	0,17	1
3		-0,19	-0,13 – -0,15	5
4		0,88	0,42 – 0,45	3

Tableau 5.3 – Points les plus mal prédits pour le potentiel d'oxydation, la température d'ébullition et la viscosité. Ces composés sont en cours d'investigation par nos collègues. À noter que pour la température de fusion, aucun composé n'a obtenu une erreur de prédiction suffisante pour être considéré comme potentiellement aberrant.

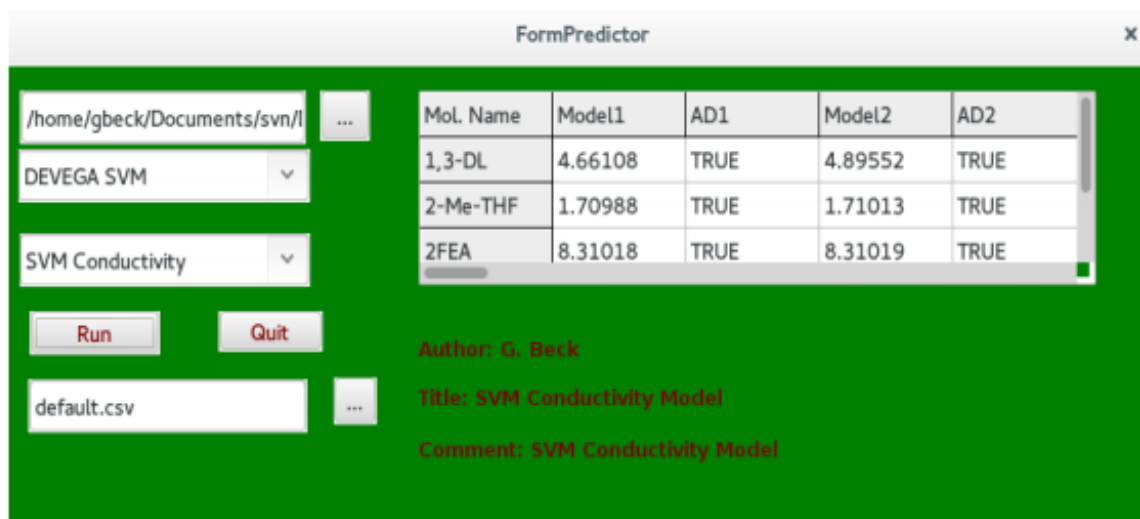


FIGURE 5.4 – Capture d'écran du logiciel ISIDA/Predictor. L'utilisateur choisit un fichier SDF à analyser et un modèle. Il reçoit les résultats sous forme de table mentionnant l'appartenance ou non de la molécule aux différents domaines d'applicabilité. Ces résultats sont également sauvegardés dans un fichier csv.

Nous avons ensuite utilisé une méthode de type produit de rang [199, 200] afin de voir quelles molécules sont les plus prometteuses. Pour chaque propriété, nous avons donné des rangs pour ordonner les prédictions de la meilleure (rang n°1) à la moins bonne (rang n°3 832). Nous avons ensuite fait, pour chaque molécule, le produit de ses rangs pour chaque propriété. Les produits ayant des valeurs les plus faibles vont donc être les molécules étant classées en tête de liste le plus souvent dans les 6 propriétés. Cette approche nécessite toutefois de disposer d'une estimation pour chacune des 6 propriétés, ce qui est rarement le cas. En effet, pour certaines propriétés, la molécule est hors du domaine d'applicabilité et le modèle ne fournit pas d'estimation fiable. Dans ce cas, la valeur moyenne de la propriété est utilisée par défaut pour une molécule dont on ne dispose pas d'estimation plus crédible. Ces moyennes sont de 5,21 V pour le potentiel d'oxydation, 3,59 mS/cm pour la conductivité, 190,8 °C pour la température d'ébullition, 19,15 °C pour la température de fusion, 0,22 log(cP) pour le logarithme de la viscosité et 1,11 log pour le logarithme de la constante diélectrique.

Nous avons finalement envoyé à nos partenaires de Chimie ParisTech les structures chimiques les plus intéressantes selon nos calculs. Ils ont synthétisé et caractérisé les composés listés dans le tableau 5.4. Ils se sont concentrés sur des esters et des dérivés. Toutes ces molécules ont été contrôlées par RMN ^1H et ^{13}C ainsi que par chromatographie en phase gazeuse couplée à de la spectrométrie de masse (*Gas chromatography-mass spectrometry* en anglais, GC-MS). Elles ont été ensuite caractérisées par DSC afin d'obtenir leurs points d'ébullition. L'appareil de DSC utilisé étant limité à -70 °C, les points de fusion n'ont en revanche pas pu être déterminés. Ces molécules ont ensuite été mélangées à une concentration de 1 M avec le LiPF_6 afin d'obtenir des électrolytes caractérisables. Pour 4 d'entre eux (molécules 9 à 12), une réaction violente a été observée par nos collègues lors de la dissolution avec le sel de lithium. Cette réactivité avec

les molécules, possédant une fonction gem-diméthyl en alpha d'un acétyl ou d'un carbonate, vis-à-vis du LiPF_6 n'était pas rapportée dans la littérature. Des analyses complémentaires sont actuellement en cours pour déterminer les produits issus de la réaction afin de comprendre cette dernière. Enfin, les conductivités et les potentiels d'oxydation des électrolytes 1 à 8 ont été déterminés.

Les données expérimentales sont renseignées dans le tableau 5.4. Nous constatons que nos modèles fonctionnent bien. En effet, si on tient compte de l'erreur des modèles (intervalle de confiance : ± 3 RMSE), les erreurs de prédictions sont presque toujours inférieures à ces intervalles. Le candidat numéro 7, dont le potentiel d'oxydation prédit à 5,08 V, a été mesuré à 3,7 V est la seule exception. Il s'agit de l'erreur la plus significative dans les estimations, et la seule dont l'erreur est supérieure à 3 RMSE. Nous pouvons relever des erreurs supérieures à 2 RMSE pour les candidats 2 (potentiel d'oxydation), 3 (température d'ébullition), 7 (conductivité et potentiel d'oxydation) et 12 (température d'ébullition). Nous observons également de nombreux cas pour lesquels l'erreur de prédiction est inférieure à 0,5 RMSE. Ces résultats sont représentés sur la figure 5.5. Les RMSE obtenues sont de 2,56 mS/cm pour la conductivité, de 0,63 V pour le potentiel d'oxydation (0,46 V lorsque l'on retire le candidat 7) et de 49,23 °C pour la température d'ébullition. Ces RMSE sont comparables à celles obtenues lors de la construction des modèles. Le Predictor peut donc fournir des informations intéressantes aux expérimentateurs souhaitant concevoir de nouveaux électrolytes.

Parmi les candidats étudiés ici, le composé numéro 8 possède une conductivité de 8,69 mS/cm, un potentiel d'oxydation de 5,5 V et une température d'ébullition de 100 °C. Ces propriétés font de ce composé le meilleur candidat.

5.3 Modélisation avec la TRR

Après la construction de modèles SVR et la proposition de candidats pour nos collègues de Paris, nous avons testé la méthode TRR sur ces données.

5.3.1 Procédure

Selon les propriétés, nous avons suivi deux procédures différentes.

Pour la constante diélectrique et la viscosité, nous avons trouvé des nouvelles données à ajouter à notre base de données. Nous avons collecté des données supplémentaires, essentiellement dans [3, 201]. Nous avons donc travaillé avec 224 composés pour la constante diélectrique et 565 molécules pour la viscosité. Pour les modéliser, nous avons suivi la procédure de l'étude de l'impact de la taille relative du jeu d'entraînement décrite dans le chapitre 3 (figure 3.4, page 49). Pour rappel, ce protocole consistait à partager le jeu de données complet en deux parts égales, à mettre une des moitiés de côté pour l'utiliser comme jeu de données étiquetées, et à tirer aléatoirement un certain pourcentage x de données de l'autre moitié (x variant entre

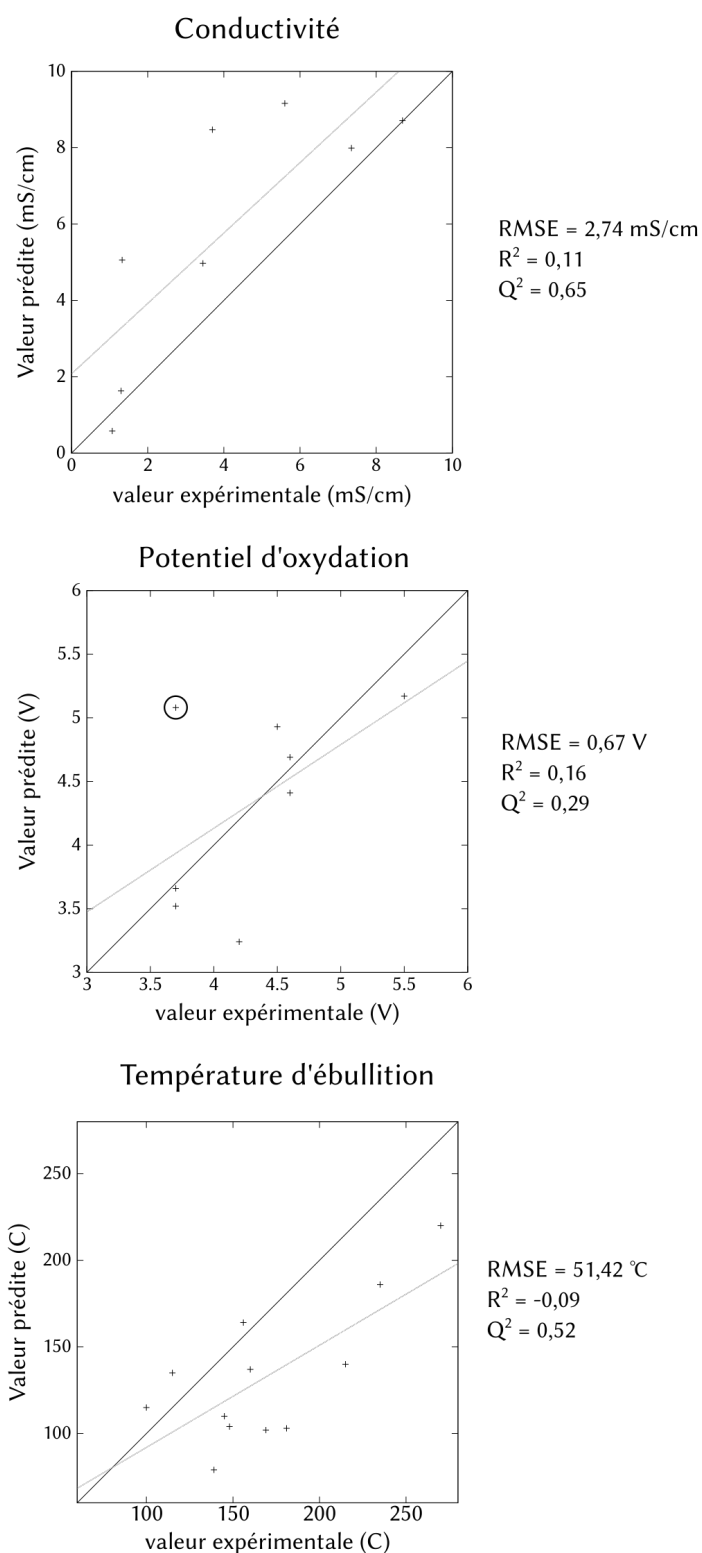


FIGURE 5.5 – Courbes exp/pred de la conductivité, la température d'ébullition et du potentiel d'oxydation pour les données criblées. Le point entouré sur le graphe du potentiel d'oxydation est le candidat numéro 7, dont le potentiel d'oxydation prédit à 5,08 V, a été mesuré à 3,7 V. Lorsque l'on retire ce point, nous obtenons les statistiques suivantes : RMSE = 0,46 ; $R^2 = 0,44$ et $Q^2 = 0,68$.

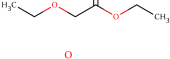
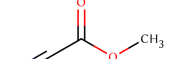
N°	Structure	Conductivité prédite (mS/cm)	Conductivité mesurée (mS/cm)	Potentiel prédit (V)	Potentiel mesuré (V)	Température d'ébullition prédite (°C)	Température d'ébullition mesurée (°C)
1		9,16	5,6	3,66	3,7	110	145
2		1,63	1,3	3,24	4,2	137	160
3		4,97	3,45	4,41	4,6	140	215
4		0,58	1,07	4,93	4,5	186	235
5		5,06	1,33	4,69	4,6	220	270
6		7,99	7,35	3,52	3,7	164	156
7		8,47	3,7	<i>5,08</i>	3,7	135	115
8		8,71	8,69	5,17	5,5	115	100
9		/	/	/	/	79	139
10		/	/	/	/	104	148
11		/	/	/	/	102	169
12		/	/	/	/	103	181

Tableau 5.4 – Molécules candidates testées expérimentalement. En gras, ce sont les cas pour lesquels la valeur prédite est proche de la valeur expérimentale (différence inférieure à 0,5 RMSE), et en italique ce sont les cas pour lesquels la différence entre les deux est supérieure à 3 RMSE. Pour rappel, RMSE(conductivité) = 2,13 mS/cm, RMSE(potentiel d'oxydation) = 0,45 V et RMSE(température d'ébullition) = 37,12 °C.

5 % et 100 %) pour former notre jeu d'entraînement. Nous avons ensuite optimisé les paramètres g , puis gp et construit deux modèles : un modèle RR avec le paramètre g , et un modèle TRR avec le paramètres g et le paramètre gp heuristique conformément à ce qui a été préconisé lors de l'étude théorique (figure 3.4, page 49).

Pour les autres propriétés, nous n'avons pas enrichi les jeux de données. Il faut noter que, dans le cas du potentiel d'oxydation, la base de données dont nous disposons est, à notre connaissance, exhaustive. Aussi, nous avons construit des modèles RR et TRR en répétant la procédure suivie pour la construction des modèles SVR de ces propriétés.

Nous présentons ici, pour chaque propriété modélisée, les résultats obtenus pour l'ensemble de descripteurs donnant les meilleurs résultats.

5.3.2 Résultats

Pour la constante diélectrique et la viscosité, nous avons obtenus des résultats cohérents avec ce qui avait déjà été observé lors de l'étude méthodologique avec les jeux de données A2AR, LogS, et pKa (voir les figures 5.6 et 5.7). En effet, on constate que les erreurs obtenues par la méthode TRR sont généralement plus faibles que celles obtenues par la méthode RR. Comme observé précédemment, dans certains cas, la transduction est absente, ce qui conduit à une dégradation des modèles TRR comparés aux modèles RR. Toutefois, la méthode mise en place permet de tirer profit de l'effet transductif tout en limitant les conséquences négatives éventuelles.

Il est également intéressant de noter que, pour la constante diélectrique, les jeux d'entraînement correspondant à une taille relative de 5 % par rapport au jeu de données non étiquetées étaient trop petits pour construire des modèles fiables.

Pour la conductivité et le potentiel d'oxydation, nous présentons les performances associées au meilleur modèle dans le tableau 5.5. Les R^2 pour les modèles TRR sont supérieurs à 0,90 pour les deux propriétés.

Enfin, les effets transductifs observés sont nuls (pour le potentiel d'oxydation) ou positifs et inférieurs à 1 % (pour la conductivité). La construction d'un modèle RR est donc suffisant ici pour obtenir des performances acceptables, et les bénéfices de la transduction sont minimes.

Propriété	Fragmentation	R^2 TRR	RMSE RR	RMSE TRR	TE
Conductivité (mS/cm)	IIAB(2-4)	0,94	0,92	0,91	1,09
Potentiel d'oxydation (V)	IIAB(2-4)_R	0,91	0,24	0,24	0,00

Tableau 5.5 – Statistiques obtenues avec la RR et la TRR pour la conductivité et le potentiel d'oxydation. Les fragmentations retenues sont des atomes centrés (de type étendu dans le cas du potentiel d'oxydation) comprenant des atomes et des liaisons de taille variant entre 2 et 4 atomes. Les modèles sont performants et le TE obtenu est positif pour la conductivité et nul pour le potentiel d'oxydation.

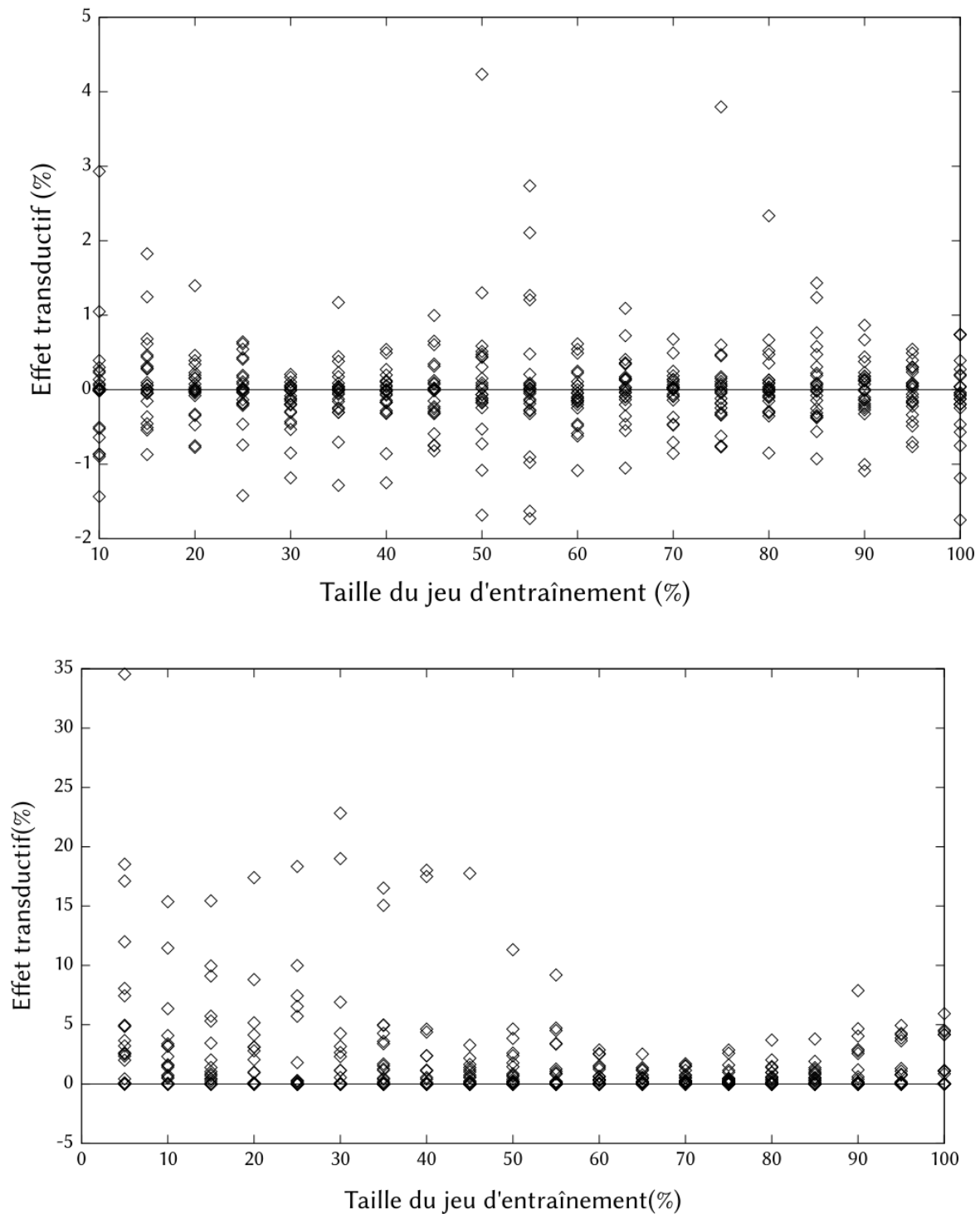


FIGURE 5.6 – TE en fonction de la taille relative du jeu d'entraînement pour la constante diélectrique (en haut) et la viscosité (en bas). Pour la viscosité, nous obtenons un graphe similaire à ceux obtenus lors de l'étude méthodologique : les TE sont généralement relativement faibles, plus amples pour les plus petites tailles. Pour la constante diélectrique, le graphe montre un équilibre entre les cas positifs et négatifs.

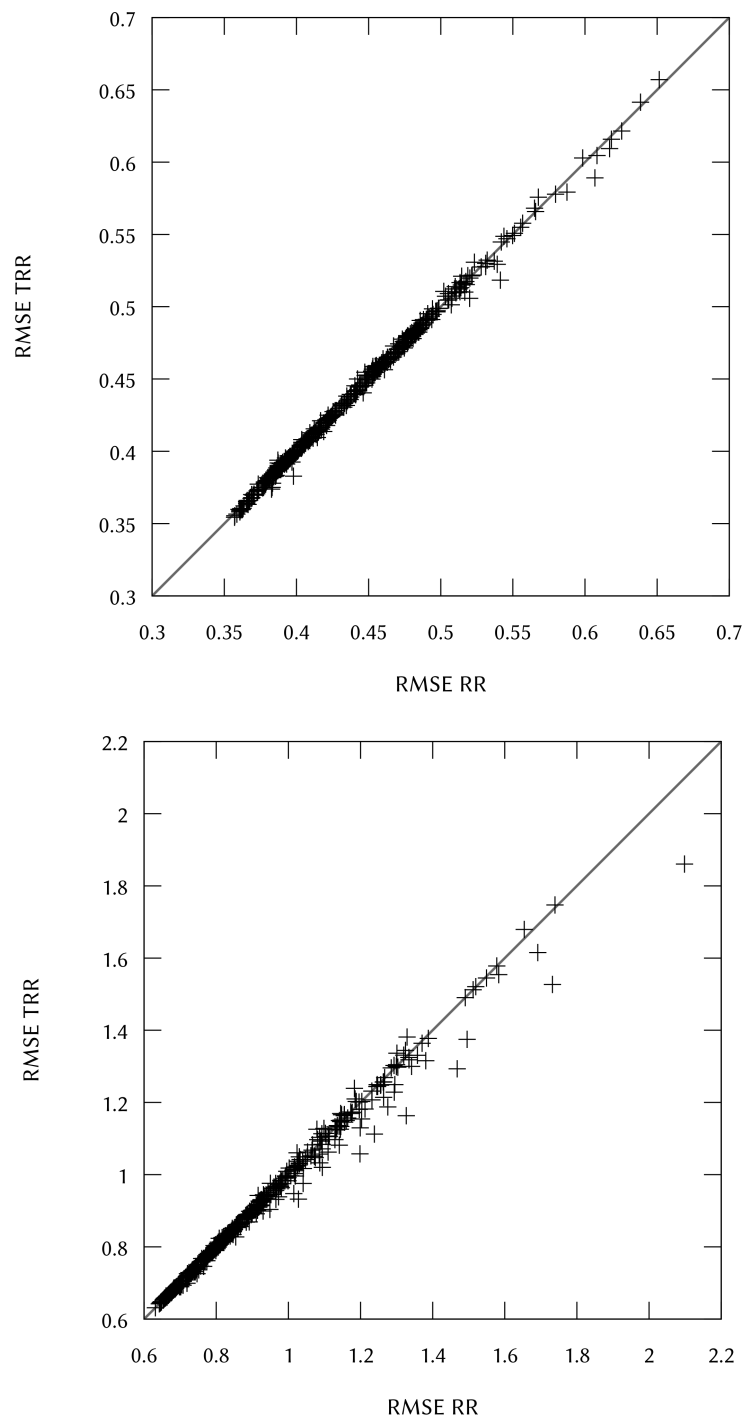


FIGURE 5.7 – Graphe RMSE TRR en fonction de la RMSE RR pour la constante diélectrique (en haut) et la viscosité (en bas). Nos obtenons des résultats similaires à ceux obtenus lors de l'étude méthodologique. Plus les RMSE sont élevées et plus la dispersion est élevée.

5.4 En résumé

Les modèles SVR consensus développés pour 6 propriétés ont été mis à disposition de nos collègues, et des candidats ont été proposés. Les résultats expérimentaux obtenus sur une série de composés de type esters et dérivés correspondent aux erreurs attendues avec ces modèles, et ont permis d'identifier un composé potentiellement intéressant pour son utilisation en tant qu'électrolyte.

En ce qui concerne les modèles TRR, les résultats obtenus lors de l'étude méthodologique (chapitre 3) se confirment. Les modèles individuels ont des performances en validation croisée satisfaisantes, même si les effets transductifs restent faibles.

Conclusion

Le but de cette thèse était de modéliser deux types de solvants à haute valeur technologique : les liquides ioniques et les électrolytes pour batteries Li-ion.

Le travail effectué sur les liquides ioniques résulte d'une collaboration avec l'équipe d'Isabelle Billard à l'IPHC de Strasbourg, ainsi qu'avec la société Solvionic, spécialisée dans la production de liquides ioniques. Nos efforts se sont concentrés sur trois propriétés : la conductivité, la température de fusion et la viscosité. Les modèles SVR construits pour ce projet ont été testés sur des liquides ioniques fournis par la société Solvionic pour lesquels les mesures n'étaient pas disponibles. Les mesures expérimentales effectuées sur ces liquides ioniques ont été réalisées par l'équipe d'Isabelle Billard en collaboration avec notre équipe.

Pour modéliser ces substances composées d'un cation et d'un anion, nous avons simplement concaténé les descripteurs ISIDA générés pour les cations aux descripteurs ISIDA générés pour les anions afin de pouvoir décrire l'ensemble du liquide ionique. Cette stratégie nous permet de travailler avec une grande diversité de liquides ioniques (nous n'avons pas besoin de travailler à anion constant) et de pouvoir construire des modèles applicables à des liquides ioniques constitués d'anions absents du jeu d'entraînement.

La qualité des données utilisées est déterminante pour la construction de modèles prédictifs performants. En effet, en utilisant des données concernant des échantillons qui ne contiennent pas plus de 500 ppm d'eau, nous avons pu, par la suite, identifier des liquides ioniques mal séchés dans le jeu de données Solvionic. De plus, les modèles construits ont une précision de prédiction comparable aux erreurs expérimentales des différentes propriétés modélisées.

La modélisation des électrolytes s'est faite dans le cadre de l'ANR DEVEGA [9]. Sur ce projet, nous avons collaboré avec les équipes de Laurent Joubert de l'université de Rouen, Alexandre Chagnes de Chimie ParisTech et Jean-Marie Tarascon du Collège de France. Nous avons modélisé 6 propriétés différentes : la conductivité, la constante diélectrique, le potentiel d'oxydation, la température d'ébullition, la température de fusion et la viscosité. Le but était de proposer de nouveaux électrolytes potentiels pour la conception de batteries Li-ion pour véhicules électriques. Nos collaborateurs de Chimie ParisTech ont ensuite synthétisé et mesuré la conductivité, le potentiel d'oxydation et la température d'ébullition des molécules que nous avons proposées.

Là encore, nous avons réussi à construire des modèles malgré le peu de données à notre disposition. Les modèles SVR construits pour la conductivité, la constante diélectrique, le potentiel d'oxydation et la viscosité ont des erreurs de prédictions raisonnables. Les modèles construits pour la température d'ébullition et la température de fusion présentent des erreurs assez grandes, mais finalement comparables aux erreurs typiquement affichées dans d'autres études QSPR (allant généralement jusqu'à 30 °C pour la température de fusion et 40 °C pour la température d'ébullition [202,203]). Les données utilisées, et donc les modèles QSPR développés, sont restreints à l'espace chimique qui concerne le développement des électrolytes.

Nous avons proposé à nos collaborateurs de nouveaux composés à synthétiser et tester en utilisant nos modèles pour estimer le profil physico-chimique d'une chimiothèque combinatoire de plus de 15 000 structures. Les molécules ont été priorisées sur la base du produit de rang des estimations de leurs propriétés. L'approche de produit de rang utilisée pour sélectionner des candidats nous a permis de proposer des composés à synthétiser et tester expérimentalement à nos collaborateurs. Les mesures de la conductivité, la température d'ébullition et le potentiel d'oxydation corrélaient bien avec les prédictions faites par nos modèles, et un des candidats possède des caractéristiques intéressantes pour son emploi possible en tant qu'électrolyte.

En raison de la faible quantité de données disponibles, nous avons envisagé d'utiliser l'approche transductive par le biais de la régression ridge transductive (TRR). La transduction consiste, au moment de l'apprentissage QSPR, à utiliser les molécules dont on cherche à estimer les propriétés pour améliorer la qualité des estimations concernant spécifiquement ces molécules. La TRR est une méthode nouvelle pour le domaine de la chémoinformatique, nous avons donc mené une étude méthodologique avant de l'appliquer aux solvants modélisés. Cette étude méthodologique a été menée sur trois jeux de données bien connus : la solubilité aqueuse, la constante d'acidité et la constante d'association avec la protéine A2AR.

Nous avons ainsi pu déterminer qu'une optimisation séquentielle du paramètre g puis du paramètre gp permet de construire un modèle TRR améliorant le modèle RR correspondant, mais pas systématiquement. Ceci nous a conduit à établir une définition de la transduction qui consiste en une amélioration d'un modèle par un procédé transductif. Cette définition permet d'identifier les situations où le procédé transductif n'améliore pas le modèle comme étant non transductives. Concernant la TRR, le paramètre transductif gp est donc nécessairement plus petit que le paramètre g puisqu'il s'agit d'une correction transductive du modèle RR. Nous avons pu observer que les effets transductifs les plus intéressants apparaissaient quand la taille du jeu d'entraînement faisait entre 5 et 20 % de la taille du jeu de validation.

Nous avons constaté que la TRR permettait d'obtenir de meilleures prédictions que son homologue RR dans une vaste majorité de cas. Toutefois, en l'absence de transduction, la correction transductive peut dégrader les performances des modèles QSPR. Nous avons mis au point une méthode permettant de tirer avantage de la transduction et d'en contrôler les effets délétères. Toutefois, pour éviter de détruire accidentellement les performances de nos modèles, il faut accepter que l'effet transductif soit faible (il est généralement inférieur à 5 %) ce qui

limite finalement l'intérêt de cette approche. Ces résultats ont été confirmés systématiquement dans des situations réelles issues de nos modélisations des propriétés des liquides ioniques et des électrolytes.

Le travail effectué lors de cette thèse ouvre les perspectives suivantes :

- En ce qui concerne la modélisation des liquides ioniques, il aurait été intéressant de modéliser également la miscibilité du liquide ionique avec l'eau, en vue d'applications de ces substances aux procédés de séparation liquide-liquide [6]. Nous n'avons, à ce jour, pas pu rassembler suffisamment d'informations à ce sujet. Cependant, il serait intéressant de surveiller la littérature pour construire progressivement un jeu de données suffisant pour pouvoir modéliser cette propriété. Nous pouvons également envisager de modéliser l'indice de réfraction utilisé dans la création de liquides ioniques avec une optique particulière (par exemple la microscopie sur des échantillons troubles) [204].
- En ce qui concerne le projet DEVEGA, nous attendons des analyses concernant les points aberrants détectés. Pour la viscosité et la constante diélectrique, nous disposons maintenant de données à plusieurs températures. Nous avons commencé la modélisation de ces propriétés en fonction de la température. Nous pourrions également modéliser l'affinité de l'électrolyte pour les Li^+ [86] et construire des modèles pour des nouvelles familles de composés (par exemple les liquides ioniques) et pour d'autres modèles de batteries.
- Que ce soit pour la modélisation des liquides ioniques où celle des électrolytes, nous pourrions utiliser des modèles de classification de l'état physique (solide ou liquide). De plus, pour utiliser correctement les modèles associés aux propriétés de transport, il faudrait distinguer si une substance liquide est, *a priori*, newtonienne ou non newtonienne. Ceci demanderait aussi de construire un modèle de classification dédié.
- Nos travaux pourraient aussi bénéficier d'autres approches d'apprentissage automatique. Les méthodes de tâches multiples (*multi-task learning* en anglais) [205] pourraient améliorer les modèles des propriétés peu fournies en données (par exemple le potentiel d'oxydation) grâce au co-apprentissage de propriétés riches en données (la viscosité). L'accès à un environnement expérimental autoriserait aussi à utiliser des techniques d'apprentissage actif (*active learning* en anglais) à l'image de l'équipe de Gisbert Schneider [206]. Des techniques de visualisation de l'espace chimique telles que la GTM (cartographie topographique générative, *Generative Topographic Mapping* en anglais) [207], une méthode couramment employée dans notre laboratoire, offrirait de nouvelles perspectives de conception rationnelle de solvants.
- Enfin, il serait intéressant, par la suite, d'étudier d'autres méthodes transductives ou semi-supervisées pour vérifier si les limitations observées avec la TRR se généralisent. Par exemple, nous pourrions essayer de coupler la transduction avec la GTM. En effet, la GTM construit un feuillet sur des données qui n'ont pas besoin d'être étiquetées,

il serait donc intéressant de voir l'impact de l'ajout de données non étiquetées dans les performances de la GTM. Des résultats en ce sens ont déjà été produits dans notre équipe [208], mais il manque une étude systématique à l'image de celle proposée dans cette thèse.

Bibliographie

- [1] Tom Welton. Solvents and sustainable chemistry. In *Proceedings of the Royal Society A*, volume 471, page 20150502. The Royal Society, 2015.
- [2] Johann Gasteiger and Thomas Engel. *Chemoinformatics : a textbook*. John Wiley & Sons, 2006.
- [3] William M Haynes. *CRC handbook of chemistry and physics*. CRC press, 2014.
- [4] Annegret Stark and Kenneth R. Seddon. *Ionic Liquids*. John Wiley and Sons, Inc., 2007.
- [5] Dimitrios Tsaoulidis, Valentina Dore, Panagiota Angeli, Natalia V Plechkova, and Kenneth R Seddon. Extraction of dioxouranium (vi) in small channels using ionic liquids. *Chemical Engineering Research and Design*, 91(4) :681–687, 2013.
- [6] Isabelle Billard, Ali Ouadi, and Clotilde Gaillard. Liquid–liquid extraction of actinides, lanthanides, and fission products by use of ionic liquids : from discovery to understanding. *Analytical and Bioanalytical Chemistry*, 400(6) :1555–1566, 2011.
- [7] Natalia V. Plechkova and Kenneth R. Seddon. Applications of ionic liquids in the chemical industry. *Chemical Society Reviews*, 37(1) :123–150, 2008.
- [8] Kang Xu. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chemical Reviews*, 104(10) :4303–4418, 2004. PMID : 15669157.
- [9] Alexandre Chagnes. Gestion des variabilités spatio-temporelles des énergies (ds0204), projet devega, design d'électrolytes à 5 v via l'approche génomique des électrolytes, 2014.
- [10] Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. pages 148–155. Morgan Kaufmann Publishers Inc., 1998.
- [11] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds. ; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3) :542–542, 2009.
- [12] Rudra Narayan Das and Kunal Roy. Advances in qspr/qstr models of ionic liquids for the design of greener solvents of the future. *Molecular Diversity*, 17(1) :151–196, 2013.
- [13] Robert Hayes, Gregory G Warr, and Rob Atkin. Structure and nanostructure in ionic liquids. *Chemical reviews*, 115(13) :6357–6426, 2015.

- [14] Chieu D Tran, Silvia H De Paoli Lacerda, and Daniel Oliveira. Absorption of water by room-temperature ionic liquids : effect of anions on concentration and state of water. *Applied spectroscopy*, 57(2) :152–157, 2003.
- [15] Maciej Galiński, Andrzej Lewandowski, and Izabela Stępniaak. Ionic liquids as electrolytes. *Electrochimica Acta*, 51(26) :5567–5580, 2006.
- [16] Kenneth R. Seddon, Annegret Stark, and María-José Torres. Influence of chloride, water, and organic solvents on the physical properties of ionic liquids. *Pure and Applied Chemistry*, 72(12) :2275–2287, 2000.
- [17] Hélène Olivier-Bourbigou, L. Magna, and D. Morvan. Ionic liquids and catalysis : Recent progress from knowledge to applications. *Applied Catalysis A : General*, 373(1) :1–56, 2010.
- [18] Daniel Joseph Tempel, Philip Bruce Henderson, and Jeffrey Richard Brzozowski. Reactive liquid based gas storage and delivery systems, February 6 2007. US Patent 7,172,646.
- [19] Hiroyuki Ohno and Yukinobu Fukaya. Task specific ionic liquids for cellulose technology. *Chemistry Letters*, 38(1) :2–7, 2009.
- [20] Ermanno F. Borra, Omar Seddiki, Roger Angel, Daniel Eisenstein, Paul Hickson, Kenneth R. Seddon, and Simon P. Worden. Deposition of metal films on an ionic liquid as a basis for a lunar telescope. *Nature*, 447(7147) :979–981, 2007.
- [21] Przemysław Majewski, Agnieszka Pernak, Marian Grzymisławski, Katarzyna Iwanik, and Juliusz Pernak. Ionic liquids in embalming and tissue preservation : Can traditional formalin-fixation be replaced safely ? *Acta Histochemica*, 105(2) :135–142, 2003.
- [22] Muhammad Moniruzzaman, Yoshiro Tahara, Miki Tamura, Noriho Kamiya, and Masahiro Goto. Ionic liquid-assisted transdermal delivery of sparingly soluble drugs. *Chemical Communications*, 46(9) :1452–1454, 2010.
- [23] Michael Zakrewsky, Katherine S Lovejoy, Theresa L Kern, Tarryn E Miller, Vivian Le, Amber Nagy, Andrew M Goumas, Rashi S Iyer, Rico E Del Sesto, Andrew T Koppisch, et al. Ionic liquids as a class of materials for transdermal delivery and pathogen neutralization. *Proceedings of the National Academy of Sciences*, 111(37) :13313–13318, 2014.
- [24] Li Chen, Genevieve E. Mullen, Myriam Le Roch, Cody G. Cassity, Nicolas Gouault, Henry Y. Fadamiro, Robert E. Barletta, Richard A. O'Brien, Richard E. Sykora, Alexandra C. Stenson, Kevin N. West, Howard E. Horne, Jeffrey M. Hendrich, Kang Rui Xiang, and James H. Davis. On the formation of a protic ionic liquid in nature. *Angewandte Chemie*, 126(44) :11956–11959, 2014.
- [25] Alan R Katritzky, Andre Lomaka, Ruslan Petrukhin, Ritu Jain, Mati Karelson, Ann E Visser, and Robin D Rogers. Qspr correlation of the melting point for pyridinium bromides, potential ionic liquids. *Journal of Chemical Information and Computer Sciences*, 42(1) :71–74, 2002.

-
- [26] Katsumi Tochigi and Hiroshi Yamamoto. Estimation of ionic conductivity and viscosity of ionic liquids using a qspr model. *The Journal of Physical Chemistry C*, 111(43) :15989–15994, 2007.
- [27] Andreas Klamt. The cosmo and cosmo-rs solvation models. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 1(5) :699–709, 2011.
- [28] AR Katritzky, VS Lobanov, and M Karelson. Codessa : reference manual. *University of Florida, Gainesville, FL*, 1994.
- [29] R Todeschini, V Consonni, A Mauri, and M Pavan. Dragon-software for the calculation of molecular descriptors. *Web version*, 3, 2004.
- [30] I. Billard, G. Marcou, A. Ouadi, and A. Varnek. In silico design of new ionic liquids based on quantitative structure-property relationship models of ionic liquid viscosity. *The Journal of Physical Chemistry B*, 115(1) :93–98, 2011.
- [31] Hiroyuki Matsuda, Hiroshi Yamamoto, Kiyofumi Kurihara, and Katsumi Tochigi. Computer-aided reverse design for ionic liquids by qspr using descriptors of group contribution type for ionic conductivities and viscosities. *Fluid Phase Equilibria*, 261(1-2) :434–443, 2007.
- [32] Claudia L Aguirre, Luis A Cisternas, and José O Valderrama. Melting-point estimation of ionic liquids by a group contribution method. *International Journal of Thermophysics*, 33(1) :34–46, 2012.
- [33] Riccardo Bini, Cinzia Chiappe, Celia Duce, Alessio Micheli, Roberto Solaro, Antonina Starita, and Maria Rosaria Tiné. Ionic liquids : prediction of their melting points by a recursive neural network model. *Green Chemistry*, 10(3) :306–309, 2008.
- [34] Gonçalo Carrera and Joao Aires-de Sousa. Estimation of melting points of pyridinium bromide ionic liquids with decision trees and neural networks. *Green Chemistry*, 7(1) :20–27, 2005.
- [35] Gonçalo V. S. M. Carrera, Luís C. Branco, Joao Aires-de Sousa, and Carlos A. M. Afonso. Exploration of quantitative structure–property relationships (qspr) for the design of new guanidinium ionic liquids. *Tetrahedron*, 64(9) :2216–2224, 2008.
- [36] David M Eike, Joan F Brennecke, and Edward J Maginn. Predicting melting points of quaternary ammonium ionic liquids. *Green Chemistry*, 5(3) :323–328, 2003.
- [37] Nasrin Farahani, Farhad Gharagheizi, Seyyed Alireza Mirkhani, and Kaniki Tumba. Ionic liquids : Prediction of melting point by molecular-based model. *Thermochimica Acta*, 549 :17–34, 2012.
- [38] Mohammad H Fatemi and Parisa Izadian. In silico prediction of melting points of ionic liquids by using multilayer perceptron neural networks. *Journal of Theoretical and Computational Chemistry*, 11(01) :127–141, 2012.
-

- [39] Farhad Gharagheizi, Poorandokht Ilani-Kashkouli, and Amir H Mohammadi. Computation of normal melting temperature of ionic liquids using a group contribution method. *Fluid Phase Equilibria*, 329 :1–7, 2012.
- [40] Subin Hada, Robert H. Herring, Sarah E. Davis, and Mario R. Eden. Multivariate characterization, modeling, and design of ionic liquid molecules. *Computers & Chemical Engineering*, 81 :310–322, 2015.
- [41] Yan Huo, Shuqian Xia, Yan Zhang, and Peisheng Ma. Group contribution method for predicting melting points of imidazolium and benzimidazolium ionic liquids. *Industrial and Engineering Chemistry Research*, 48(4) :2212–2217, 2009.
- [42] Alan R Katritzky, Ritu Jain, Andre Lomaka, Ruslan Petrukhin, Mati Karelson, Ann E Visser, and Robin D Rogers. Correlation of the melting points of potential ionic liquids (imidazolium bromides and benzimidazolium bromides) using the codessa program. *Journal of Chemical Information and Computer Sciences*, 42(2) :225–231, 2002.
- [43] Natalia Kireeva, Sergey L Kuznetsov, and Aslan Yu Tsivadze. Toward navigating chemical space of ionic liquids : prediction of melting points using generative topographic maps. *Industrial and Engineering Chemistry Research*, 51(44) :14337–14343, 2012.
- [44] Juan A. Lazzús. A group contribution method to predict the melting point of ionic liquids. *Fluid Phase Equilibria*, 313 :1–6, 2012.
- [45] Ignacio Lopez-Martin, Enrico Burello, Paul N. Davey, Kenneth R. Seddon, and Gadi Rothenberg. Anion and cation effects on imidazolium salt melting points : a descriptor modelling study. *ChemPhysChem*, 8(5) :690–695, 2007.
- [46] Ulrich Preiss, Safak Bulut, and Ingo Krossing. In silico prediction of the melting points of ionic liquids from thermodynamic considerations : a case study on 67 salts with a melting point range of 337 c. *The Journal of Physical Chemistry B*, 114(34) :11133–11140, 2010.
- [47] Ulrich P Preiss, Witali Beichel, Anna MT Erle, Yauheni U Paulechka, and Ingo Krossing. Is universal, simple melting point prediction possible ? *ChemPhysChem*, 12(16) :2959–2972, 2011.
- [48] Yueying Ren, Jin Qin, Huanxiang Liu, Xiaojun Yao, and Mancang Liu. Qspr study on the melting points of a diverse set of potential ionic liquids by projection pursuit regression. *QSAR and Combinatorial Science*, 28(11-12) :1237–1244, 2009.
- [49] Ning Sun, Xuezhong He, Kun Dong, Xiangping Zhang, Xingmei Lu, Hongyan He, and Suojiang Zhang. Prediction of the melting points for two kinds of room temperature ionic liquids. *Fluid phase equilibria*, 246(1) :137–142, 2006.
- [50] José S. Torrecilla, Francisco Rodríguez, Jose L. Bravo, Gadi Rothenberg, Kenneth R. Seddon, and Ignacio Lopez-Martin. Optimising an artificial neural network for predicting the melting point of ionic liquids. *Physical Chemistry Chemical Physics*, 10(38) :5826–5831, 2008.

-
- [51] Steven Trohalaki and Ruth Pachter. Prediction of melting points for ionic liquids. *QSAR and Combinatorial Science*, 24(4) :485–490, 2005.
- [52] Steven Trohalaki, Ruth Pachter, Greg W. Drake, and Tommy Hawkins. Quantitative structure-property relationships for melting points and densities of ionic liquids. *Energy & Fuels*, 19(1) :279–284, 2005.
- [53] José O. Valderrama and Roberto E. Rojas. Data selection and estimation of the normal melting temperature of ionic liquids using a method based on homologous cations. *Comptes Rendus Chimie*, 15(8) :693–699, 2012.
- [54] Alexandre Varnek, Natalia Kireeva, Igor V. Tetko, Igor I. Baskin, and Vitaly P. Solov'ev. Exhaustive qspr studies of a large diverse set of ionic liquids : How accurately can we predict melting points ? *Journal of Chemical Information and Modeling*, 47(3) :1111–1122, 2007.
- [55] Fangyou Yan, Shuqian Xia, Qiang Wang, Zhen Yang, and Peisheng Ma. Predicting the melting points of ionic liquids by the quantitative structure property relationship method using a topological index. *The Journal of Chemical Thermodynamics*, 62 :196–200, 2013.
- [56] Rafael Alcalde, Gregorio Garcia, Mert Atilhan, and Santiago Aparicio. Systematic study on the viscosity of ionic liquids : Measurement and prediction. *Industrial and Engineering Chemistry Research*, 54(43) :10918–10924, 2015.
- [57] Maciej Barycki, Anita Sosnowska, Agnieszka Gajewicz, Maciej Bobrowski, Dorota Wi-leńska, Piotr Skurski, Artur Giełdoń, Cezary Czaplewski, Stefanie Uhl, and Edith Laux. Temperature-dependent structure-property modeling of viscosity for ionic liquids. *Fluid Phase Equilibria*, 427 :9–17, 2016.
- [58] C. I. Daniel, J. Albo, E. Santos, C. A. M. Portugal, J. G. Crespo, and A. Irabien. A group contribution method for the influence of the temperature in the viscosity of magnetic ionic liquids. *Fluid Phase Equilibria*, 360 :29–35, 2013.
- [59] Juan de Riva, Victor R. Ferro, Lourdes del Olmo, Elia Ruiz, Rafael Lopez, and Jose Palomar. Statistical refinement and fitting of experimental viscosity-to-temperature data in ionic liquids. *Industrial and Engineering Chemistry Research*, 53(25) :10475–10484, 2014.
- [60] Pablo Díaz-Rodríguez, John C. Cancilla, Gemma Matute, and José S. Torrecilla. Determination of physicochemical properties of pyridinium-based ionic liquid binary mixtures with a common component through neural networks. *Industrial and Engineering Chemistry Research*, 53(2) :1015–1019, 2014.
- [61] Pablo Díaz-Rodríguez, John C. Cancilla, Gemma Matute, and José S. Torrecilla. Viscosity estimation of binary mixtures of ionic liquids through a multi-layer perceptron model. *Journal of Industrial and Engineering Chemistry*, 21(0) :1350–1353, 2015.
- [62] Chao Han, Guangren Yu, Lu Wen, Dachuan Zhao, Charles Asumana, and Xiaochun Chen. Data and qspr study for viscosity of imidazolium-based ionic liquids. *Fluid Phase Equilibria*, 300(1) :95–104, 2011.
-

- [63] Seyyed Alireza Mirkhani and Farhad Gharagheizi. Predictive quantitative structure–property relationship model for the estimation of ionic liquid viscosity. *Industrial and Engineering Chemistry Research*, 51(5) :2470–2477, 2012.
- [64] Kamil Paduszyński and Urszula Domańska. Viscosity of ionic liquids : An extensive database and a new group contribution model based on a feed-forward artificial neural network. *Journal of Chemical Information and Modeling*, 2014.
- [65] Guangren Yu, Dachuan Zhao, Lu Wen, Shendu Yang, and Xiaochun Chen. Viscosity of ionic liquids : Database, observation, and quantitative structure-property relationship analysis. *AIChE Journal*, 58(9) :2885–2899, 2012.
- [66] Yongsheng Zhao, Ying Huang, Xiangping Zhang, and Suojiang Zhang. A quantitative prediction of the viscosity of ionic liquids using σ -profile molecular descriptors. *Physical Chemistry Chemical Physics*, 17(5) :3761–3767, 2015.
- [67] Nan Zhao, Johan Jacquemin, Ryan Oozeerally, and Volkan Degirmenci. New method for the estimation of viscosity of pure and mixtures of ionic liquids based on the unifac–visco model. *Journal of Chemical and Engineering Data*, 61(6) :2160–2169.
- [68] Yu Cao, Jia Yu, Hang Song, Xianlong Wang, and Shun Yao. Prediction of the electric conductivity of ionic liquids by two chemometrics methods. *Journal of the Serbian Chemical Society*, 78(5) :653–667, 2013.
- [69] Pablo Díaz-Rodríguez, John C. Cancilla, Gemma Matute, and José S. Torrecilla. Conductivity of ionic liquids : a neural network approach. *Industrial and Engineering Chemistry Research*, 54(1) :55–58, 2014.
- [70] Farhad Gharagheizi, Mehdi Sattari, Poorandokht Ilani-Kashkouli, Amir H. Mohammadi, Deresh Ramjugernath, and Dominique Richon. A “non-linear” quantitative structure-property relationship for the prediction of electrical conductivity of ionic liquids. *Chemical Engineering Science*, 101 :478–485, 2013.
- [71] Farhad Gharagheizi, Poorandokht Ilani-Kashkouli, Mehdi Sattari, Amir H Mohammadi, Deresh Ramjugernath, and Dominique Richon. Development of a lssvm-gc model for estimating the electrical conductivity of ionic liquids. *Chemical Engineering Research and Design*, 92(1) :66–79, 2014.
- [72] L Cammarata, SG Kazarian, PA Salter, and T Welton. Molecular states of water in room temperature ionic liquids. *Physical Chemistry Chemical Physics*, 3(23) :5192–5200, 2001.
- [73] Jason A Widegren, Arno Laesecke, and Joseph W Magee. The effect of dissolved water on the viscosities of hydrophobic room-temperature ionic liquids. *Chemical Communications*, (12) :1610–1612, 2005.
- [74] Jason A Widegren, Eric M Saurer, Kenneth N Marsh, and Joseph W Magee. Electrolytic conductivity of four imidazolium-based room-temperature ionic liquids and the effect of a water impurity. *The journal of Chemical Thermodynamics*, 37(6) :569–575, 2005.

-
- [75] Takatsugu Endo, Tatsuya Kato, Ken-ichi Tozaki, and Keiko Nishikawa. Phase behaviors of room temperature ionic liquid linked with cation conformational changes : 1-butyl-3-methylimidazolium hexafluorophosphate. *The Journal of Physical Chemistry B*, 114(1) :407–411, 2010.
- [76] C Chiappe, P Margari, A Mezzetta, CS Pomelli, S Koutsoumpos, M Papamichael, P Gianios, and K Moutzouris. Temperature effects on the viscosity and the wavelength-dependent refractive index of imidazolium-based ionic liquids with a phosphorus-containing anion. *Physical Chemistry Chemical Physics*, 19(12) :8201–8209, 2017.
- [77] David Linden and Thomas B Reddy. Handbook of batteries. 3rd, 2002.
- [78] Yu Wang and Wei-Hong Zhong. Development of electrolytes towards achieving safe and high-performance energy-storage devices : A review. *ChemElectroChem*, 2(1) :22–36, 2015.
- [79] Elise Nanini-Maury. *Formulation d'électrolytes haut potentiel pour la caractérisation d'électrodes positives innovantes : batteries lithium-ion pour le véhicule électrique*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2014.
- [80] Jennifer Jones. *Etude des interfaces électrodes/électrolyte et des phénomènes de solubilité dans l'accumulateur lithium-ion*. PhD thesis, Tours, 2010.
- [81] T. Nagaura, M. Yokokawa, and T. Hashimoto. Rechargeable organic electrolyte cell, May 9 1989. US Patent 4,828,834.
- [82] Lidan Xing, Jenel Vatamanu, Oleg Borodin, Grant D Smith, and Dmitry Bedrov. Electrode/electrolyte interface in sulfolane-based electrolytes for li ion batteries : a molecular dynamics simulation study. *The Journal of Physical Chemistry C*, 116(45) :23871–23881, 2012.
- [83] P Ganesh, PRC Kent, and De-en Jiang. Solid-electrolyte interphase formation and electrolyte reduction at li-ion battery graphite anodes : Insights from first-principles molecular dynamics. *The Journal of Physical Chemistry C*, 116(46) :24476–24481, 2012.
- [84] Kevin Leung and Joanne L Budzien. Ab initio molecular dynamics simulations of the initial stages of solid–electrolyte interphase formation on lithium ion battery graphitic anodes. *Physical Chemistry Chemical Physics*, 12(25) :6583–6586, 2010.
- [85] Kevin Leung. Electronic structure modeling of electrochemical reactions at electrode/electrolyte interfaces in lithium ion batteries. *The Journal of Physical Chemistry C*, 117(4) :1539–1547, 2012.
- [86] Min Hee Park, Yoon Sup Lee, Hochun Lee, and Young-Kyu Han. Low li+ binding affinity : An important characteristic for additives to form solid electrolyte interphases in li-ion batteries. *Journal of Power Sources*, 196(11) :5109–5114, 2011.
- [87] Lidan Xing, Chaoyang Wang, Weishan Li, Mengqing Xu, Xuliang Meng, and Shaofei Zhao. Theoretical insight into oxidative decomposition of propylene carbonate in the lithium ion battery. *The Journal of Physical Chemistry B*, 113(15) :5181–5187, 2009.
-

- [88] Lidan Xing, Weishan Li, Chaoyang Wang, Fenglong Gu, Mengqing Xu, Chunlin Tan, and Jin Yi. Theoretical investigations on oxidative stability of solvents and oxidative decomposition mechanism of ethylene carbonate for lithium ion battery use. *The Journal of Physical Chemistry B*, 113(52) :16596–16602, 2009.
- [89] Dmitry Bedrov, Grant D Smith, and Adri CT Van Duin. Reactions of singly-reduced ethylene carbonate in lithium battery electrolytes : a molecular dynamics simulation study using the reaxff. *The Journal of Physical Chemistry A*, 116(11) :2978–2985, 2012.
- [90] Young-Kyu Han, Jaehoon Jung, Sunghoon Yu, and Hochun Lee. Understanding the characteristics of high-voltage additives in li-ion batteries : Solvent effects. *Journal of Power Sources*, 187(2) :581–585, 2009.
- [91] Shyue Ping Ong, Oliviero Andreussi, Yabi Wu, Nicola Marzari, and Gerbrand Ceder. Electrochemical windows of room-temperature ionic liquids from molecular dynamics and density functional theory calculations. *Chemistry of Materials*, 23(11) :2979–2986, 2011.
- [92] Mathew D Halls and Ken Tasaki. High-throughput quantum chemistry and virtual screening for lithium ion battery electrolyte additives. *Journal of Power Sources*, 195(5) :1472–1478, 2010.
- [93] F Eckert and A Klamt. Cosmothem, version c3. 0, release 13.01. *COSMOlogic GmbH & Co. KG, Leverkusen, Germany*, 2013.
- [94] Pipeline Pilot. Version 8.5. accelrys. *Inc. : San Diego, CA*, 92121, 2011.
- [95] Christoph Schütter, Tamara Husch, Martin Korth, and Andrea Balducci. Toward new solvents for edlcs : from computational screening to electrochemical validation. *The Journal of Physical Chemistry C*, 119(24) :13413–13424, 2015.
- [96] Tamara Husch, Nusret Duygu Yilmazer, Andrea Balducci, and Martin Korth. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents : computing infrastructure and collective properties. *Physical Chemistry Chemical Physics*, 17(5) :3394–3401, 2015.
- [97] Christoph Schütter, Tamara Husch, Venkatasubramanian Viswanathan, Stefano Passerini, Andrea Balducci, and Martin Korth. Rational design of new electrolyte materials for electrochemical double layer capacitors. *Journal of Power Sources*, 326 :541–548, 2016.
- [98] Martin Korth. Computational studies of solid electrolyte interphase formation. *Chemical Modelling*, 11 :57, 2014.
- [99] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for chemoinformatics, volume 41 (2 volume set)*, volume 41. John Wiley & Sons, 2009.
- [100] A. Varnek, D. Fourches, F. Hoonakker, and V. P. Solov'ev. Substructural fragments : an universal language to encode reactions, molecular and supramolecular structures. *Journal of Computer-Aided Molecular Design*, 19(9) :693–703, 2005.

-
- [101] Alexandre Varnek, Denis Fourches, Dragos Horvath, Olga Klimchuk, Cedric Gaudin, Philippe Vayer, Vitaly Solov'ev, Frank Hoonakker, Igor V Tetko, and Gilles Marcou. Isida-platform for virtual screening based on fragment and pharmacophoric descriptors. *Current Computer-Aided Drug Design*, 4(3) :191–198, 2008.
- [102] Fanny Bonachéra, Benjamin Parent, Frédérique Barbosa, Nicolas Froloff, and Dragos Horvath. Fuzzy tricentric pharmacophore fingerprints. 1. topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *Journal of Chemical Information and Modeling*, 46(6) :2457–2477, 2006.
- [103] Fanny Bonachéra and Dragos Horvath. Fuzzy tricentric pharmacophore fingerprints. 2. application of topological fuzzy pharmacophore triplets in quantitative structure- activity relationships. *Journal of Chemical Information and Modeling*, 48(2) :409–425, 2008.
- [104] Fiorella Ruggiu, Gilles Marcou, Alexandre Varnek, and Dragos Horvath. Isida property-labelled fragment descriptors. *Molecular Informatics*, 29(12) :855–868, 2010.
- [105] Fiorella Ruggiu, Vitaly Solov'ev, Gilles Marcou, Dragos Horvath, Jérôme Graton, Jean-Yves Le Questel, and Alexandre Varnek. Individual hydrogen-bond strength qspr modeling with isida local descriptors : a step towards polyfunctional molecules. *Molecular Informatics*, 33(6-7) :477–487, 2014.
- [106] Mircea Braban, Iuliana Pop, Xavier Willard, and Dragos Horvath. Reactivity prediction models applied to the selection of novel candidate building blocks for high-throughput organic synthesis of combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 39(6) :1119–1127, 1999.
- [107] Aurélie de Luca. *Espaces chimiques optimaux pour la recherche par similarité, la classification et la modélisation de réactions chimiques représentées par des graphes condensés de réactions*. PhD thesis, Université de Strasbourg, september 2015.
- [108] ChemAxon. Jchem 15.8.10, 2015. URL : <http://www.chemaxon.com>.
- [109] Chemical Computing Group ULC. Molecular operating environment (moe), 2014. 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.
- [110] David R. Lide. *CRC handbook of chemistry and physics*. CRC press, 1994.
- [111] P Labute. Moe molar refractivity model. unpublished. *Source code in MOE/lib/svl/quasar.svl/q_mref.svl*, 1998.
- [112] Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5) :868–873, 1999.
- [113] P Labute. Moe logp (octanol/water) model. unpublished. *Source code in MOE/lib/svl/quasar.svl/q_logp.svl*, 1998.
-

- [114] LH Hall and LB Kier. The molecular connectivity chi indices and kappa shape indices in structure-property modelling. *Reviews of Computational Chemistry*. v2, pages 367–422, 1991.
- [115] LB Kier and LH Hall. Nature of structure-activity-relationships and their relation to molecular connectivity. *European Journal of Medicinal Chemistry*, 12(4) :307–312, 1977.
- [116] A.T. Balaban. Five new topological indices for the branching of tree-like graphs. *Theoretica Chimica Acta*, 53 :355–375, 1979.
- [117] Alexandru T Balaban. Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89(5) :399–404, 1982.
- [118] Michel Petitjean. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *Journal of Chemical Information and Computer Sciences*, 32(4) :331–337, 1992.
- [119] Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1) :17–20, 1947.
- [120] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22) :3219–3228, 1980.
- [121] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems*, pages 155–161, 1997.
- [122] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.
- [123] Chih-Chung Chang and Chih-Jen Lin. Libsvm : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3) :27 :1–27 :27, 2011.
- [124] Ian H Witten and Eibe Frank. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
- [125] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3) :4, 2006.
- [126] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-supervised learning*. 2010.
- [127] Guo-Zheng Li, Jack Y. Yang, Wen-Cong Lu, Dan Li, and Mary Qu Yang. Improving prediction accuracy of drug activities by utilising unlabelled instances with feature selection. *International journal of computational biology and drug design*, 1(1) :1–13, 2008.
- [128] Zhi-Hua Zhou and Ming Li. Semi-supervised regression with co-training. In *International Joint Conference on Artificial Intelligence*, volume 5, pages 908–913, 2005.
- [129] S Stanley Young, Vijay K Gombar, Michael R Emptage, Neal F Cariello, and Christophe Lambert. Mixture deconvolution and analysis of ames mutagenicity data. *Chemometrics and Intelligent Laboratory Systems*, 60(1) :5–11, 2002.

-
- [130] Xia Ning, Huzefa Rangwala, and George Karypis. Multi-assay-based structure-activity relationship models : improving structure-activity relationship models by incorporating activity information from related targets. *Journal of Chemical Information and Modeling*, 49(11) :2444–2456, 2009.
- [131] JB Brown, Yasushi Okuno, Gilles Marcou, Alexandre Varnek, and Dragos Horvath. Computational chemogenomics : Is it more than inductive transfer ? *Journal of Computer-Aided Molecular Design*, 28(6) :597–618, 2014.
- [132] Jurica Levatić, Saso Džeroski, Fran Supek, and Tomislav Smuc. Semi-supervised learning for quantitative structure-activity modeling. *Informatica*, 37(2) :173–179, 2013.
- [133] Kurt Driessens, Peter Reutemann, Bernhard Pfahringer, and Claire Leschi. Using weighted nearest neighbor to benefit from unlabeled data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 60–69. Springer, 2006.
- [134] Mark Culp and George Michailidis. A co-training algorithm for multi-view data with applications in data fusion. *Journal of Chemometrics*, 23(6) :294–303, 2009.
- [135] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2004.
- [136] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.
- [137] SL Holbeck. Update on nci in vitro drug screen utilities. *European Journal of Cancer*, 40(6) :785–793, 2004.
- [138] Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius Ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of chemical information and modeling*, 49(9) :2077–2081, 2009.
- [139] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1) :31–71, 1997.
- [140] J. Levatić, F. Supek, and S. Džeroski. Improving qsar models by exploiting unlabeled data from public ddatabase od bioactive drug-like molecules. In *7th Jožef Stefan International Postgraduate School Students' Conference*, pages 114–123, Ljubljana, Slovenia, 2015.
- [141] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL : a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1) :D1100–D1107, 2012.
- [142] Evgeny Kondratovich, Igor I. Baskin, and Alexandre Varnek. Transductive support vector machines : Promising approach to model small and unbalanced datasets. *Molecular Informatics*, 32(3) :261–266, 2013.
-

- [143] Corinna Cortes and Mehryar Mohri. On transductive regression. *Advances in Neural Information Processing Systems*, 19 :305–313, 2007.
- [144] Andrei Nikolaevich Tikhonov, Vasilii Iakovlevich Arsenin, and Fritz John. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.
- [145] Arthur E Hoerl and Robert W Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.
- [146] Paola Gramatica. Principles of qsar models validation : internal and external. *Molecular Informatics*, 26(5) :694–701, 2007.
- [147] Horvath Dragos, Marcou Gilles, and Varnek Alexandre. Predicting the predictability : a unified approach to the applicability domain problem of qsar models. *Journal of Chemical Information and Modeling*, 49(7) :1762–1776, 2009.
- [148] Arthur Stanley Eddington. *Stellar Movements and the Structure of the Universe*. Macmillan and Company, limited, 1914.
- [149] Karl Pearson. Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Proceedings of the Royal Society of London*, 59(353-358) :69–71, 1895.
- [150] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572, 1901.
- [151] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15 :246–263, 1886.
- [152] Sheldon M Ross. *Initiation aux probabilités*. PPUR presses polytechniques, 2007.
- [153] Fiorella Ruggiu. *Property-enriched fragment descriptors for adaptive QSAR*. PhD thesis, Strasbourg, 2014.
- [154] Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry : miniperspective. *Journal of Medicinal Chemistry*, 55(7) :2932–2942, 2012.
- [155] Free Pascal Team. *Free Pascal : A 32, 64 and 16 bit professional Pascal compiler*. Fairfax, VA, 1993-2016. version 3.0. URL : <http://www.freepascal.org>.
- [156] Lazarus Team. *Lazarus : The professional Free Pascal RAD IDE*. Fairfax, VA, 1993-2016. version 1.6. URL : <http://www.lazarus-ide.org>.
- [157] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. Golden section search in one dimension. *Numerical Recipes in C : the Art of Scientific Computing*, 2, 1992.
- [158] Jarmo Huuskonen. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *Journal of Chemical Information and Computer Sciences*, 40(3) :773–777, 2000.

-
- [159] Nathan R McElroy and Peter C Jurs. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *Journal of Chemical Information and Computer Sciences*, 41(5) :1237–1247, 2001.
- [160] Yingqing Ran, Neera Jain, and Samuel H Yalkowsky. Prediction of aqueous solubility of organic compounds by the general solubility equation (gse). *Journal of Chemical Information and Computer Sciences*, 41(5) :1208–1217, 2001.
- [161] Denise Yaffe, Yoram Cohen, Gabriela Espinosa, Alex Arenas, and Francesc Giralt. A fuzzy artmap based on quantitative structure- property relationships (qsprs) for predicting aqueous solubility of organic compounds. *Journal of Chemical Information and Computer Sciences*, 41(5) :1177–1207, 2001.
- [162] Igor Baskin, Gilles Marcou, and Alexandre Varnek. 2nd strasbourg summer school on chemoinformatics. Strasbourg, 2010.
- [163] Syracuse Research. Physprom, 2013 ed. (ed. : S. r. corporation), 2017. Corporation, Inc., 2017.
- [164] Mourad Elhabiri, Pavel Sidorov, Elena Cesar-Rodo, Gilles Marcou, Don Antoine Lanfranchi, Elisabeth Davioud-Charvet, Dragos Horvath, and Alexandre Varnek. Electrochemical properties of substituted 2-methyl-1, 4-naphthoquinones : Redox behavior predictions. *Chemistry-A European Journal*, 21(8) :3415–3424, 2015.
- [165] Gilles Marcou and Igor V. Tetko. 3rd strasbourg summer school on chemoinformatics. Strasbourg, 2012.
- [166] Anthony J Harmar, Rebecca A Hills, Edward M Rosser, Martin Jones, O Peter Buneman, Donald R Dunbar, Stuart D Greenhill, Valerie A Hale, Joanna L Sharman, Tom I Bonner, et al. Iuphar-db : the iuphar database of g protein-coupled receptors and ion channels. *Nucleic acids research*, 37(suppl 1) :D680–D685, 2008.
- [167] Yanli Wang, Evan Bolton, Svetlana Dracheva, Karen Karapetyan, Benjamin A. Shoemaker, Tugba O. Suzek, Jiyao Wang, Jewen Xiao, Jian Zhang, and Stephen H. Bryant. An overview of the pubchem bioassay resource. *Nucleic Acids Research*, 38(suppl 1) :D255–D266, 2010.
- [168] Jonathan G Huddleston, Ann E Visser, W Matthew Reichert, Heather D Willauer, Grant A Broker, and Robin D Rogers. Characterization and comparison of hydrophilic and hydrophobic room temperature ionic liquids incorporating the imidazolium cation. *Green Chemistry*, 3(4) :156–164, 2001.
- [169] Dandan Han and Kyung Ho Row. Recent applications of ionic liquids in separation technology. *Molecules*, 15(4) :2405–2426, 2010.
- [170] Oliver Zech, Alexander Stoppa, Richard Buchner, and Werner Kunz. The conductivity of imidazolium-based ionic liquids from (248 to 468) kb variation of the anion. *Journal of Chemical and Engineering Data*, 55(5) :1774–1778, 2010.
-

- [171] BE Mbondo Tsamba, S Sarraute, Mounir Traikia, and P Husson. Transport properties and ionic association in pure imidazolium-based ionic liquids as a function of temperature. *Journal of Chemical and Engineering Data*, 59(6) :1747–1754, 2014.
- [172] Fiorella Ruggiu, Patrick Gizzi, Jean-Luc Galzi, Marcel Hibert, Jacques Haiech, Igor Baskin, Dragos Horvath, Gilles Marcou, and Alexandre Varnek. Quantitative structure-property relationship modeling : a valuable support in high-throughput screening quality control. *Analytical chemistry*, 86(5) :2510–2520, 2014.
- [173] Sigma-aldrich catalogue. URL : <http://www.sigmaaldrich.com/france.html>.
- [174] Yukihiro Yoshida, Masatoshi Kondo, and Gunzi Saito. Ionic liquids formed with polycyano 1, 1, 3, 3-tetracyanoallyl anions : Substituent effects of anions on liquid properties. *The Journal of Physical Chemistry B*, 113(26) :8960–8966, 2009.
- [175] J. Vila, P. Gines, E. Rilo, O. Cabeza, and L. M. Varela. Great increase of the electrical conductivity of ionic liquids in aqueous solutions. *Fluid Phase Equilibria*, 247(1) :32–39, 2006.
- [176] Minato Egashira, Shigeto Okada, and Jun-ichi Yamaki. The effect of the coexistence of anion species in imidazolium cation-based molten salt systems. *Solid State Ionics*, 148(3) :457–461, 2002.
- [177] Zhi-Bin Zhou, Hajime Matsumoto, and Kuniaki Tatsumi. Low-melting, low-viscous, hydrophobic ionic liquids : 1-alkyl (alkyl ether)-3-methylimidazolium perfluoroalkyltrifluoroborate. *Chemistry-A European Journal*, 10(24) :6581–6591, 2004.
- [178] Kang Xu, Michael S. Ding, and T. Richard Jow. Quaternary onium salts as non-aqueous electrolytes for electrochemical capacitors. *Journal of The Electrochemical Society*, 148(3) :A267–A274, 2001.
- [179] Om D. Gupta, Brendan Twamley, and M. Shreeve Jean'ne. Low melting and slightly viscous ionic liquids via protonation of trialkylamines by perfluoroalkyl β -diketones. *Tetrahedron letters*, 45(8) :1733–1736, 2004.
- [180] Giovanni B. Appetecchi, Maria Montanino, Maria Carewska, Margherita Moreno, Fabrizio Alessandrini, and Stefano Passerini. Chemical–physical properties of bis (perfluoroalkyl-sulfonyl) imide-based ionic liquids. *Electrochimica Acta*, 56(3) :1300–1307, 2011.
- [181] Yukihiro Yoshida, Osamu Baba, Carlos Larriba, and Gunzi Saito. Imidazolium-based ionic liquids formed with dicyanamide anion : influence of cationic structure on ionic conductivity. *The Journal of Physical Chemistry B*, 111(42) :12204–12210, 2007.
- [182] Hajime Matsumoto, Hiroyuki Kageyama, and Yoshinori Miyazaki. Room temperature molten salts based on tetraalkylammonium cations and bis (trifluoromethylsulfonyl) imide. *Chemistry Letters*, (2) :182–183, 2001.

-
- [183] Alain Berthod, M. J. Ruiz-Angel, and Samuel Carda-Broch. Ionic liquids in separation techniques. *Journal of Chromatography A*, 1184(1) :6–18, 2008.
- [184] Luís C. Branco, João N. Rosa, Joaquim J. Moura Ramos, and Carlos A. M. Afonso. Preparation and characterization of new room temperature ionic liquids. *Chemistry—A European Journal*, 8(16) :3671–3677, 2002.
- [185] Glen McHale, Chris Hardacre, Rile Ge, Nicola Doy, Ray W. K. Allen, Jordan M. MacInnes, Mark R. Bown, and Michael I. Newton. Density-viscosity product of small-volume ionic liquid samples using quartz crystal impedance analysis. *Analytical Chemistry*, 80(15) :5806–5811, 2008.
- [186] Peter Wasserscheid, Roy van Hal, and Andreas Bösmann. 1-n-butyl-3-methylimidazolium ([bmim]) octylsulfate—an even ‘greener’ ionic liquid. *Green Chemistry*, 4(4) :400–404, 2002.
- [187] Kenneth R. Harris, Mitsuhiro Kanakubo, and Lawrence A. Woolf. Temperature and pressure dependence of the viscosity of the ionic liquids 1-hexyl-3-methylimidazolium hexafluorophosphate and 1-butyl-3-methylimidazolium bis (trifluoromethylsulfonyl) imide. *Journal of Chemical and Engineering Data*, 52(3) :1080–1085, 2007.
- [188] D. Behar, P. Neta, and Carl Schultheisz. Reaction kinetics in ionic liquids as studied by pulse radiolysis : redox reactions in the solvents methyltributylammonium bis (trifluoromethylsulfonyl)imide and n-butylpyridinium tetrafluoroborate. *The Journal of Physical Chemistry A*, 106(13) :3139–3147, 2002.
- [189] Arijit Bhattacharjee, Andreia Luís, João H. Santos, José A. Lopes-da Silva, Mara G. Freire, Pedro J. Carvalho, and João A. P. Coutinho. Thermophysical properties of sulfonium- and ammonium-based ionic liquids. *Fluid Phase Equilibria*, 381 :36–45, 2014.
- [190] John S. Wilkes and Michael J. Zaworotko. Air and water stable 1-ethyl-3-methylimidazolium based ionic liquids. *Journal of the Chemical Society, Chemical Communications*, (13) :965–967, 1992.
- [191] Gwan-Hong Min, Tae-eun Yim, Hyun-Yeong Lee, Dal-Ho Huh, Eun-joo Lee, Jun-young Mun, Seung M. Oh, and Young-Gyu Kim. Synthesis and properties of ionic liquids : imidazolium tetrafluoroborates with unsaturated side chains. *Bulletin of the Korean Chemical Society*, 27(6) :847–852, 2006.
- [192] Safak Bulut, Philipp Eiden, Witali Beichel, John M. Slattery, Tom F. Beyersdorff, Thomas J. S. Schubert, and Ingo Krossing. Temperature dependence of the viscosity and conductivity of mildly functionalized and non-functionalized [tf2n]-ionic liquids. *ChemPhysChem*, 12(12) :2296–2310, 2011.
- [193] Yukihiro Yoshida and Gunzi Saito. Ionic liquids based on diethylmethyl (2-methoxyethyl) ammonium cations and bis (perfluoroalkanesulfonyl) amide anions : influence of anion structure on liquid properties. *Physical Chemistry Chemical Physics*, 13(45) :20302–20310, 2011.
-

- [194] A. Orita, K. Kamijima, and M. Yoshida. Allyl-functionalized ionic liquids as electrolytes for electric double-layer capacitors. *Journal of Power Sources*, 195(21) :7471–7479, 2010.
- [195] Svante Arrhenius. The viscosity of solutions. *Biochemical Journal*, 11(2) :112–133, 1917.
- [196] Osborne Reynolds. On the theory of lubrication and its application to mr. beauchamp tower’s experiments, including an experimental determination of the viscosity of olive oil. *Proceedings of the Royal Society of London*, 40(242-245) :191–203, 1886.
- [197] K Lichtenecker. Die herleitung des logarithmischen mischungsgesetzes aus allgemeinen prinzipien der stationaren stromung. *phys. Z*, 32 :255–260, 1931.
- [198] ChemAxon. Marvin 17.5.0, 2017. URL : <http://www.chemaxon.com>.
- [199] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products : a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3) :83–92, 2004.
- [200] Rob Eisinga, Rainer Breitling, and Tom Heskes. The exact probability distribution of the rank product statistics for replicated experiments. *FEBS letters*, 587(6) :677–682, 2013.
- [201] Dabir S Viswanath, Tushar K Ghosh, Dasika HL Prasad, Nidamarty VK Dutt, and Kalipatnapu Y Rani. Experimental data. In *Viscosity of Liquids*, pages 443–643. Springer, 2007.
- [202] John C. Dearden. Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. *Environmental Toxicology and Chemistry*, 22(8) :1696–1709, 2003.
- [203] Ioana Oprisiu, Gilles Marcou, Dragos Horvath, Damien Bernard Brunel, Fabien Rivollet, and Alexandre Varnek. Publicly available models to predict normal boiling point of organic compounds. *Thermochimica Acta*, 553 :60–67, 2013.
- [204] Hideki Ishida, Yukari Gobara, Mayumi Kobayashi, and Toshinobu Suzuki. Use of ionic liquid for scanning electron microscopy of protists. *International Journal of New Technology and Research*, 2 :43–46, 2016.
- [205] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.
- [206] Daniel Reker and Gisbert Schneider. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today*, 20(4) :458–465, 2015.
- [207] Christopher M Bishop, Markus Svensén, and Christopher KI Williams. Gtm : The generative topographic mapping. *Neural computation*, 10(1) :215–234, 1998.
- [208] Pavel Sidorov, Helena Gaspar, Gilles Marcou, Alexandre Varnek, and Dragos Horvath. Mappability of drug-like space : towards a polypharmacologically competent map of drug-relevant compounds. *Journal of Computer-Aided Molecular Design*, 29(12) :1087–1108, 2015.

**Modélisation QSPR de solvants
d'intérêt technologique :
les liquides ioniques et
les électrolytes pour batteries Li-ion**

Résumé

Cette thèse a pour but de modéliser les liquides ioniques et les électrolytes pour batteries Li-ion. Nous avons développé des modèles SVR afin de prédire 9 propriétés d'intérêt pour ces solvants. Les modèles construits pour les liquides ioniques ont permis la détection de divers problèmes, et sont accessibles sur le site web du laboratoire : infochim.u-strasbg.fr/webserv/VSEngine.html. Les modèles construits pour les électrolytes ont permis la modélisation de candidats testés expérimentalement par nos collaborateurs. Le nombre de données étant limité pour ces solvants, nous avons également testé l'approche transductive par le biais de la TRR (Transductive Ridge Regression). Nous avons mis en place un protocole d'optimisation des paramètres de la méthode et appliqué la TRR aux solvants étudiés. Les résultats obtenus par la TRR sont légèrement meilleurs que ceux de la Régression Ridge, mais restent modestes si on veut éviter une détérioration accidentelle du modèle.

Mots-clés : liquides ioniques, électrolytes, QSPR, transduction, SVR, RR, TRR

Abstract

This thesis is dedicated to the modelling of ionic liquids and electrolytes of Li-ion batteries. We developed several SVR models in order to predict 9 interesting properties of these solvents. The models built for the ionic liquids allowed us to detect several problems, and are freely available on the laboratory's website: infochim.u-strasbg.fr/webserv/VSEngine.html. The models built for the electrolytes were used to model some candidates tested experimentally by our colleagues. As the amount of data is quite small for these solvents, we also tested the transductive approach with the help of the TRR (Transductive Ridge Regression). We have developed an optimization procedure for the method's parameters, and applied the TRR to the studied solvents. The results obtained with the TRR are slightly better than of the Ridge Regression but stay modest if we want to avoid any accidental damage of the model.

Keywords: ionic liquids, electrolytes, QSPR, transduction, SVR, RR, TRR