Relations structure-activité pour le métabolisme et la toxicité

Christophe Muller

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

UMR 7177

THÈSE

présentée par

Christophe MULLER

soutenue le : **24 janvier 2013**

pour obtenir le grade de

Docteur de l'université de Strasbourg

Discipline / Spécialité : Chimie / Chémoinformatique

Relations Structure-Activité pour le métabolisme et la toxicité.

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg

RAPPORTEURS :

M. BUREAU Ronan	Professeur, Université de Caen
Mme. CAMPROUX Anne-Claude	Professeur, Université Paris Diderot

---

MEMBRES DU JURY :

M. ROGNAN Didier	Docteur, Université de Strasbourg
Mme. RICHERT Lysiane	Professeur, Université de Franche-Comté
M. VAYER Philippe	Docteur, Technologies Servier

Remerciements

Je tiens tout d'abord à remercier le Professeur Alexandre Varnek pour m'avoir accueilli au sein de son laboratoire et pour m'avoir prodigué de nombreux conseils scientifiques et humains. Je le remercie également de m'avoir fortement conseillé de faire du monitorat, je ne pensais sincèrement pas que je prendrai autant de plaisir à enseigner. Et finalement je le remercie de m'avoir permis de participer à de nombreux congrès, particulièrement ceux dans les pays de l'est, scientifiquement parlant bien entendu.

Je veux également remercier tout spécialement le Docteur Gilles Marcou qui a TOUJOURS été là pour discuter de mes projets et m'orienter, même lorsque son emploi du temps était plus que serré. Merci aussi pour ta bonne humeur quoiqu'il arrive.

Ensuite je voudrai adresser ma sincère reconnaissance au Professeur Ronan Bureau, au Professeur Anne-Claude Camproux, au Docteur Didier Rognan, au Professeur Lysiane Richert, ainsi qu'au Docteur Philippe Vayer pour avoir accepté de juger mon travail et de faire partie de mon jury de thèse. Je remercie particulièrement ces deux derniers pour l'aide qu'ils m'ont apportés dans certains projets.

Je tiens également à remercier tous les collaborateurs et collègues membres du laboratoire: Docteur Igor Baskin, Docteur Vitaly Solov'yev, Docteur Natalia Kireeva, Docteur Vladminir Chupakhin, Docteur Pavel Polishchuk, Docteur Olga Klimchuk (merci pour ta super bonne humeur), Docteur Fanny Bonachera (merci pour les conversations « pause café »), et Docteur Dragos Horvath (merci pour tes blagues et pour le fait de t'énerver constamment contre tes ordis, ca me fait bien sourire).

J'adresse des remerciements particuliers à mes collègues doctorants ou ex collègues doctorants: Docteur Ioana Ioprisiu (merci pour ta bonne humeur et toutes les conversations de tout et n'importe quoi qu'on a pu avoir), Laurent Hoffer (mon collègue de bureau et colocataire attitré lors des conférences hors de Strasbourg. Merci pour toutes les conversations « originales » qu'on a pu avoir), Aurélie de Luca (qui n'a vraiment pas de chance question santé, merci pour les conversations potins qu'on a eu), Fiorella Ruggiu (qui a connu les bonheurs du monitorat en même temps que moi), Tetiana Khristova (merci pour ta gentillesse), et Héléna Gaspar (nouvelle collègue dans le bureau, merci de me supporter sans broncher).

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

Je tiens également à remercier les membres des autres laboratoires avec qui j'ai pu travailler ou avec qui j'ai passé de bons moments lors des manifestations telles que les écoles d'été en Chémoinformatique: Coraline (KaLy-Cell) et Tul (KaLy-Cell, une vraie encyclopédie du médicament) ; et les membres du Laboratoire d'Innovation Thérapeutique (merci pour les bons moments et les rigolades qu'on a pu avoir).

Je finirai ces remerciements avec ma famille (surtout mes parents et ma sœur) qui a toujours cru en moi, et mes amis (vivi, zarbi, tché, apprenti, pooms, …) qui m'ont aidé à bien décompresser le weekend, et parfois même trop bien.

Table des matières

Table des figures

Tables des figures

Tables des figures

Introduction

Il n'est pas possible de mesurer toutes les propriétés ADME/Tox pour tous les composés dès les premières étapes de recherche de médicaments. Cela engendrerai un trop grand coût financier et temporel. De nouvelles méthodes de prédictions informatiques doivent donc être développées afin d'aider l'expérimentaliste à filtrer les composés sur lesquels les tests biologiques doivent être fait. Seul l'aspect métabolisme et toxicité seront étudiés dans cette thèse.

Le fait que le métabolisme associé à un composé est souvent représenté sous la forme d'une succession de réactions qui n'est pas une forme directement utilisable pour les méthodes traditionnelles de Chémoinformatique rend la prédiction du métabolisme compliquée. De ce fait, les modèles de prédiction des métabolites, voir de chemins métaboliques n'utilisant pas directement l'information réactionnelle sont très utilisés dans les compagnies pharmaceutiques ; les principaux axes de recherches étant la sélectivité substrat/enzyme et la régiosélectivité, c'est-à-dire la localisation des sites d'une molécule capable de subir des biotransformations vis-à-vis d'une enzyme donnée.

Il existe de nombreux logiciels spécialisés, plus ou moins élaborés, dans la prédiction de la régiosélectivité, mais le plus connu est MetaSite. Le principal souci de cette méthode est qu'il n'existe pas de seuil défini pour distinguer les sites réellement oxydés des sites non oxydés. Nous proposons ici une méthode originale pour transformer les réactions métaboliques en objets facilement utilisables, et ainsi obtenir des modèles où l'information réactionnelle serait directement contenue dans les descripteurs.

Les mécanismes par lesquels une molécule peut induire une toxicité sont nombreux et variés. Dans certains cas, les mécanismes qui font qu'une cellule meurt ne sont pas toujours connus. Cela rend la modélisation de la toxicité très compliquée. Dans cette thèse 2 types de toxicité sont étudiés : la mutagénicité et l'hépatotoxicité. La mutagénicité est aisément mise en évidence par le test d'Ames mais il est impossible d'envisager une campagne de criblage pour un trop grand nombre de molécules étant donné le temps nécessaire à la mise en place du test ainsi que le coût financier que cela représente. Il sera donc proposé dans cette thèse de construire un modèle prédictif de la mutagénicité pour limiter les coûts.

Contexte

L'hépatotoxicité est l'une des toxicités les plus compliquées à prédire. Contrairement à d'autres toxicités comme la toxicité gastro-intestinale, cardiovasculaire ou hématologique, les tests sur les animaux ne permettent d'anticiper que dans 50% des cas une hépatotoxicité humaine. De plus les modèles QSAR construit jusqu'à maintenant souffrent véritablement de précision dans leur prédictions de l'hépatotoxicité. Il existe donc un réel besoin de développer de nouveaux tests biologiques et/ou des modèles informatiques capables d'estimer avec plus de précision ce type de toxicité. Nous proposons ici de construire des modèles hybrides utilisant des descripteurs moléculaires mais aussi biologiques.

Cette thèse comporte 3 parties : (i) la première partie propose une description des stratégies et techniques employées nécessaires à l'élaboration et l'utilisation de modèles QSAR/QSPR, ainsi qu'une brève présentation des bases de données exploitées ; (ii) la deuxième partie se penche sur l'utilisation des Graphes Condensés de Réactions pour la représentation des réactions métaboliques, l'évaluation de cette approche, et finalement la prédiction des sites d'oxydation pour les substrats d'enzymes du cytochrome P450 ; (iii) la troisième partie s'attachera à l'élaboration de modèles prédictifs de la mutagénicité ainsi que de l'hépatotoxicité.

PREMIERE PARTIE : Développement, validation et applications des modèles QSAR.

Dans ce premier chapitre, une étude bibliographique des différentes méthodologies QSAR/QSPR est proposée. Le traitement des données, molécules et réactions, est tout d'abord mis en avant. Puis les différentes étapes de développement, validation et application des modèles sont introduites. Et enfin, une brève présentation des bases de données exploitées est proposée.

1. Méthodologie QSAR/QSPR.

Un modèle QSPR/QSAR est un modèle qui tente de relier, de manière qualitative ou quantitative, la structure des molécules à une propriété ou activité donnée. L'élaboration de tels modèles suit 3 étapes :

- Développement du modèle. Des descripteurs sont choisis afin de traduire de manière numérique la structure des molécules et une méthode d'apprentissage est sélectionnée afin de construire un modèle.
- Validation du modèle. Plusieurs procédures visant à estimer les performances du modèle et sa robustesse sont mises en place.
- Application du modèle. Le domaine dans lequel le modèle est applicable est défini afin d'éviter des extrapolations hasardeuses.

La construction de tels modèles étant tributaire de la qualité des données, il faut avant tout vérifier les données et choisir une représentation unique pour celles-ci.

1.1. Nettoyage des données

Les performances des modèles QSAR/QSPR dépendent fortement de la qualité des structures du jeu de données et des activités ciblées. De petites erreurs dans la structure du composé peuvent entraîner une perte significative du pouvoir prédictif d'un modèle QSAR [1]. Avant toute modélisation il est donc indispensable de vérifier et de corriger si nécessaire les données.

Il y'a plusieurs éléments à vérifier dans les étapes de nettoyage d'un jeu de données [2]. Il faut tout d'abord vérifier que les structures sont correctes d'un point de vue chimique (règle de valence, …). Des structures erronées entraînent la génération de mauvais descripteurs et donc de mauvais modèles.

Les structures doivent être standardisées (représenter toutes les structures sous forme aromatique, …) afin que chaque molécule soit représentée de la même manière. Deux représentations différentes de la même structure peuvent entraîner la génération de descripteurs différents, et perturbent l'identification des duplicats.

Il faut ensuite veiller à supprimer les composés inorganiques qui ont un comportement différents des composés organiques qu'on cherche en général à modéliser. On écarte les mélanges, sauf si l'on connaît le principe actif, dans quel cas on garde uniquement la molécule expliquant l'activité. On fait de même pour les sels.

Les duplicats doivent aussi être traités. Leur présence à la fois dans le jeu d'entraînement et le jeu de test affecte directement la valeur des critères statistiques d'évaluation des performances du modèle. Il faut tout particulièrement prêter attention aux structures identiques qui ne possèdent pas la même activité. Les duplicats sont éliminés du jeu et seul un exemple du composé possédant l'activité correcte est conservée.

Il faut si possible vérifier l'activité des structures finalement retenues, ce qui n'est pas toujours aisé pour de grands jeu de données.

Seulement une fois toutes ces étapes remplies il est possible de réaliser des modèles QSAR dans de bonnes conditions.

1.2. Graphe Condensé de Réaction

Une réaction chimique [3] est une transformation d'une ou plusieurs molécules. Au cours de celle-ci on aura des formations de liaisons ainsi que des ruptures de liaisons. Toutefois, ces réarrangements des molécules sont dépendants des conditions réactionnelles. Mais quelles que soient les conditions, la réaction chimique se fait sans variation de masse. Les travaux de Lavoisier ont permis de mettre en évidence cette constatation : « Rien ne se perd, rien ne se crée, tout se transforme ».

Généralement les réactions sont présentées par une flèche caractérisant le sens de la réaction. A gauche de cette flèche on trouve les réactifs, et à droite de celle-ci les produits. Cependant sous cette forme il est difficile de coder numériquement (descripteurs) les transformations qui ont lieux. C'est pourquoi les réactions sont représentées sous la forme d'un Graphe Condensé de Réaction.

Un GCR [4] est une pseudomolécule correspondant à la superposition des graphes moléculaires des produits et des réactifs représenté par des liaisons conventionnelles ainsi que par des liaisons dynamiques correspondant aux transformations chimiques (Figure 1-1).

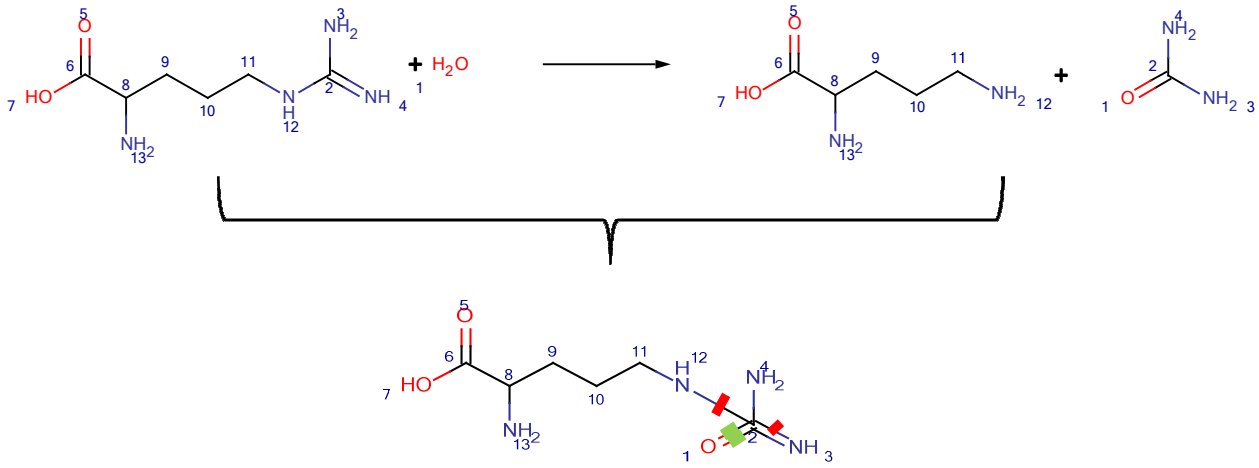


Figure 1-1. Réaction d'hydrolyse (KEGG R00551) et son graphe condensé de réaction.
Les liaisons « dynamiques » correspondantes aux transformations chimiques sont données en couleur : C=O (double liaison formée), NH-C (simple liaison cassée) et C=NH (double transformée en simple). Les numéros des atomes sont établis par la procédure du mapping.

Sous cette forme les méthodes traditionnelles de chémoinformatique sont alors applicables.

Cependant pour générer un GCR il faut pouvoir associer à chaque atome des réactifs son image dans les produits. Ce procédé est réalisable par des logiciels automatiques de « atom-to-atom » mapping. Cette procédure consiste à attribuer un même numéro au même atome du côté des réactifs et des produits. La réaction présentée Figure 1-1 a été mappée.

L'algorithme permettant de générer les GCRs est présenté dans la partie *logiciels développés* (8.1 GCR Designer).

1.3. Descripteurs moléculaires

Les descripteurs moléculaires ont pour but de décrire de manière numérique la structure d'une molécule. Un lien peut éventuellement être trouvé entre les descripteurs calculés et la propriété souhaitée pour la molécule. A ce jour plus de

6000 descripteurs ont été répertoriés [5]. On distingue plusieurs types de descripteurs :

- Descripteurs 1D : ces descripteurs sont directement calculés à partir de la formule brute d'une molécule. Ils correspondent à de simples comptages des propriétés de la molécule comme le nombre d'atomes, la masse moléculaire, etc.

- Descripteurs 2D : ils sont calculés à partir de la structure 2D de la molécule ou encore de la matrice de connectivité. On distingue 3 sous-types de descripteurs à deux dimensions : les indices topologiques (indices de Wiener [6], de Randić [7], etc.) et constitutionnels (nombre de liaisons simple, double, nombre de cycles, etc.), les propriétés physico-chimiques (BCUT [8], pharmacophore 2-, 3-, 4-points, etc.), et enfin les descripteurs basés sur les fragments (énumération de séquences d'atomes et de liaisons, etc.).

- Descripteurs 3D : la structure 3D de la molécule est requise pour calculer ce type de descripteurs. On distingue à nouveau l'utilisation d'indices et de propriétés physico-chimiques adaptés à la 3D et calculés grâce aux distances inter-atomiques ; mais aussi de nouveaux descripteurs de surface, volume ou encore quantique.

- Descripteurs 4D : ils correspondent à la mesure des propriétés 3D (potentiel électrostatique, d'hydrophobicité et de liaison hydrogène) d'une molécule en tout point de l'espace. Les approches les plus connus pour ces calculs sont CoMFA [9] et GRID [10].

1.3.1. Descripteurs ISIDA SMF

On distingue 2 types de descripteurs ISIDA [11]: les descripteurs SMF (Substructural Molecular Fragment)[12] et les descripteurs IPLF (ISIDA Property-Labelled Fragment)[13].

Afin de décrire les graphes moléculaires et d'établir des modèles de classification, deux classes de descripteurs fragmentaux ISIDA SMF développées par le laboratoire

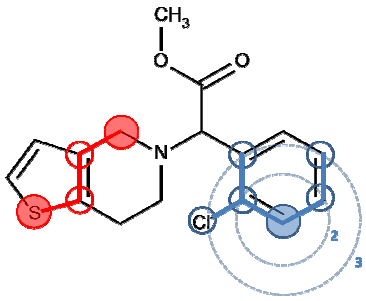
d'Infochimie seront utilisées: les « séquences » d'atomes et les « atomes unis ». Il s'agit de motifs qui sont détectés dans les structures 2D des composés analysés. Chaque motif est utilisé comme un descripteur, dont la valeur est le nombre de fois qu'il a été détecté dans la structure :

Les descripteurs de type I correspondent aux séquences d'atomes (IA), aux séquences de liaisons (IB), ou aux séquences d'atomes et de liaisons (IAB). Pour chaque type de séquence, le nombre minimal Nmin et maximal Nmax des atomes inclus est défini. La séquence correspond au chemin le plus court à parcourir pour relier 2 atomes d'une structure moléculaire ; et dans le cas de deux chemins de mêmes longueurs, les 2 chemins seront choisis.

Les descripteurs de type II correspondent aux atomes unis. Un atome uni représente un atome central et son environnement direct (1ère sphère de coordination) pouvant inclure les atomes et les liaisons (IIAB), les atomes (IIA) uniquement, ou les liaisons (IIB) seulement. L'état d'hybridation et l'environnement des atomes unis (IIHy) peuvent également être pris en compte et sont alors codés par des types atomiques donnés ; par exemple CD=Csp^2, CB=C aromatique, CO=carbonyle, etc. Dans le cas des atomes unis augmentés (type III), les sphères de coordination de taille Nmin à Nmax sont ajoutées simultanément. Pour le type IV, les sphères de coordination de taille Nmin à Nmax sont ajoutées indépendamment. La Figure 1-2 montre un exemple de fragmentations.

Pour les descripteurs de type I et II, il est également possible de considérer les paires d'atomes (IAP/IIAP). Ces descripteurs sont générés en notant explicitement le label des atomes situés aux extrémités du fragment considéré ainsi que la distance topologique qui sépare ces atomes.

Dans le cas des GCRs la même procédure de génération des fragments est appliquée. La seule différence repose sur le fait qu'on observe l'apparition de liaisons dynamiques dans l'énumération des fragments. De plus, seule la génération de descripteurs contenant au moins une liaison dynamique ou uniquement des liaisons dynamiques peut être entreprise. Un exemple de fragmentation d'un GCR est proposé Figure 1-3.



	SEQUENCES (I)	ATOMES UNIS (II)	ATOMES UNIS AUGMENTES (III)	ATOMES UNIS AUGMENTES (IV)
Atomes et liaisons	S-C=C-C ; S-C=C ; C=C-C ; S-C ; C=C ; C-C	C(=C)(-C)	C(=C)(-C)(=C-C) (=C-Cl)(-C=C)	C(=C)(-C); C(=C-C)(=C-Cl)(-C=C)
Atomes	SCCC ; SCC ; CCC ; SC ; CC	C(C)(C)	C(C)(C)(CC)(CCl)(CC)	C(C)(C); C(CC)(CCl)(CC)
Liaisons	-=- ; -= ; - ; =	C(=)(-)	C(=)(-)(=-)(=-)(-=)	C(=)(-); C(=-)(=-)(-=)

Figure 1-2. Enumération des descripteurs SMF pour des fragments de longueur de 2 à 3 atomes.
Les séquences sont calculées pour le chemin rouge, et les atomes unis pour les chemins bleus.

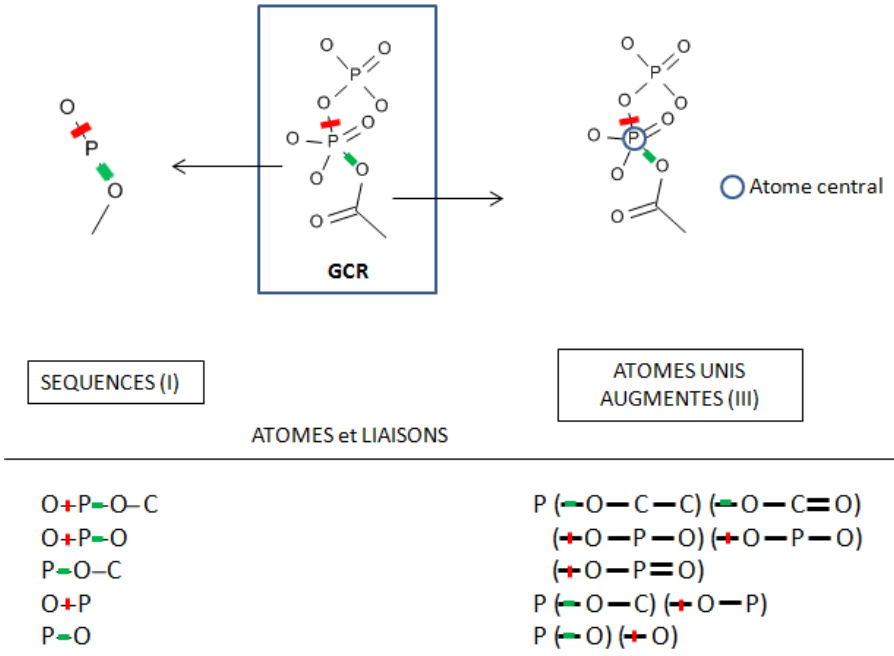


Figure 1-3. Enumération des descripteurs SMF dans le cas des GCRs.
Seules les séquences contenant au minimum une liaison dynamique (rouge si cassée, verte si créée) sont considérées. Cet exemple montre les séquences d'atomes et liaisons de longueur 2 à 4 qui sont générées ainsi que les atomes unis augmentés de longueur 2 à 3 atomes. Les fragments contenant uniquement les atomes ou les liaisons peuvent être déduits des fragments ci-dessus en omettant respectivement les symboles des atomes ou des liaisons.

1.3.2. Descripteurs ISIDA IPLF

Ces descripteurs se basent sur une représentation modifiée de la structure initiale de la molécule. Au lieu de considérer le label des atomes dans l'énumération des fragments comme pour les descripteurs SMF, on considère les propriétés de ces atomes. Ces propriétés peuvent être des propriétés pharmacophoriques ou encore des propriétés basées sur le potentiel électrostatique. Uniquement la coloration (attribution d'une propriété à un atome) pharmacophorique sera traitée dans cette thèse. Ainsi tous les atomes correspondant à des accepteurs de liaisons hydrogènes auront un même label « A ». Un exemple de coloration pharmacophorique est présenté pour 2 micro-espèces du Bupropion sur la Figure 1-4:

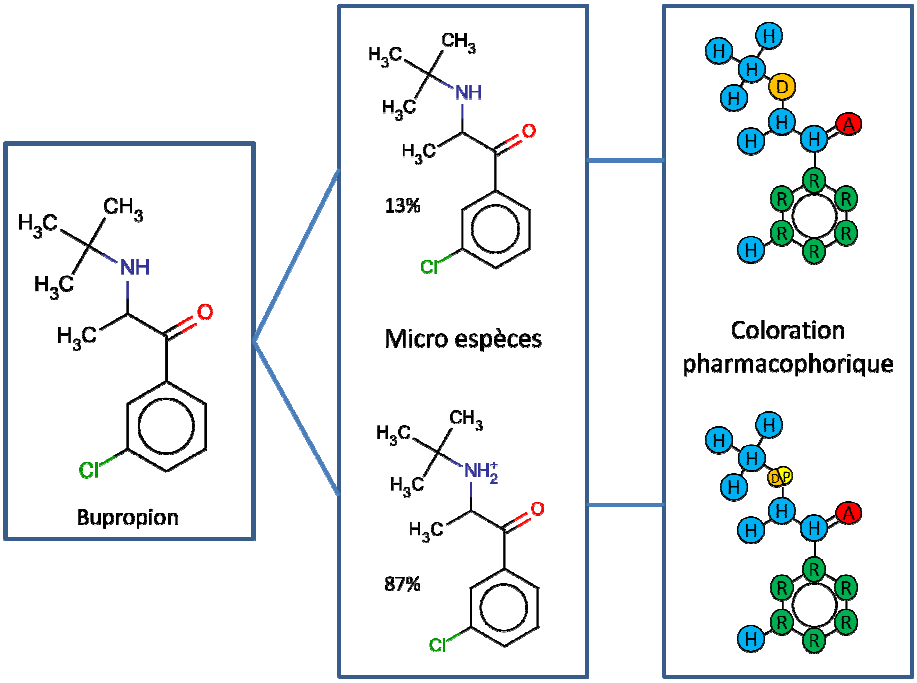


Figure 1-4. Micro-espèces du Bupropion pour un pH de 7,4 et coloration pharmacophorique.

La coloration pharmacophorique est faite à l'aide du logiciel « pharmacophore mapper » [14] de ChemAxon. La nomenclature est établie selon les règles suivantes :

- R : atomes aromatiques
- P : centre de groupements fonctionnels chargés positivement
- N : centre de groupements fonctionnels chargés négativement

- D : donneur de liaisons hydrogène
- A : accepteur de liaisons hydrogène
- H : les atomes hydrophobes non concernés par les règles précédentes
- F : les atomes ne suivant aucune des règles citées

Un atome peut bien entendu posséder plusieurs propriétés pharmacophoriques à la fois.

Le calcul des descripteurs peut se faire soit uniquement pour la micro-espèce majoritaire, soit pour chaque micro-espèce présente à un pH égal à 7,4 en pondérant les descripteurs par le taux de présence de celles-ci. Dans le second cas, seules les micro-espèces ayant un niveau de population significatif (supérieur à 1%) sont conservées.

Dans la Figure 1-4, la séquence « A=H-R », par exemple, sera comptabilisé 13 fois dans la micro-espèce du haut et 87 fois dans la micro-espèce du bas afin de rendre compte des différents taux de présence de chacune d'elle. Cela donne une occurrence totale de 100 pour ce fragment dans la molécule. Pour les atomes possédant plusieurs propriétés, les séquences contenant ces atomes seront comptabilisées pour chaque propriété indépendamment. Pour la micro-espèce du haut dans la Figure 1-4, la séquence « A/D-H-H », par exemple, sera comptabilisé 13 fois comme « A-H-H » et 13 fois comme « D-H-H ».

1.3.3. Descripteurs MOE

Le logiciel MOE (Molecular Operating Environment) [15] permet, entre autres, de générer des descripteurs 2D pour un jeu de molécules. Ces descripteurs sont calculés à partir de la table de connectivité de la molécule et permettent, tout comme les descripteurs fragmentaux, de s'affranchir de la conformation des molécules.

Les 183 descripteurs 2D disponibles ont été générés. On distingue les descripteurs physicochimiques (poids moléculaire, densité, somme des polarisabilités atomiques), les comptages d'atomes et de liaisons, les descripteurs se basant sur une approximation de la surface accessible de van der Waals des atomes et de leur contributions atomiques à certaines propriétés, les indices de Kier&Hall et de Kappa, les descripteurs associés aux matrices de distance et de proximité (rayon

de la molécule, diamètre, …), les descripteurs pharmacophoriques, et enfin les charges partielles.

La liste entière des descripteurs MOE 2D générés est disponible sur http://www.chemcomp.com/journal/descr.htm.

1.4. Machines d'apprentissages

L'objectif d'une modélisation QSPR est de trouver une fonction qui relie la structure d'une molécule, codée par des descripteurs moléculaires, à une propriété. Les machines d'apprentissage ont pour but de calculer cette fonction. Selon la méthode la fonction ne sera pas la même. Dans ces travaux des méthodes d'apprentissage supervisé ont été employées : SVM (Machine à Vecteurs Supports), J48 (arbre de décision), RF (Random Forest/forêt aléatoire), NB (Bayésien Naïf), VP (Perceptron Votant), SQS (Stochastic QSAR Sampler), JRip (apprentissage de règles), ASNN (Associative neural network), AMORE (A MORE flexible neural network) ; et des méthodes d'apprentissage non supervisé : K-Means (clustering) et carte de Kohonen.

1.4.1. SVM (Machine à Vecteurs Supports)

La SVM [16, 17] est une méthode de classification supervisée introduite par Vapnik. Le but de cette méthode est de trouver un hyperplan qui sépare au mieux les objets de 2 classes, c'est-à-dire un plan pour lequel tous les objets appartenant à la 1ère classe se situent d'un côté, et tous les objets appartenant à la 2nd classe de l'autre côté. Pour ce faire l'algorithme cherche l'hyperplan tel que la distance entre les plus proches voisins des 2 classes (vecteurs supports) et cet hyperplan soit maximale (voir Figure 1-5).

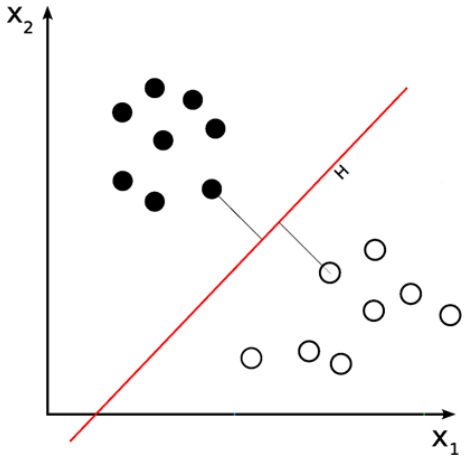


Figure 1-5. Classification binaire dans le cas d'une SVM.

Dans le cas des problèmes non linéairement séparables un tel plan ne peut pas être trouvé dans l'espace des descripteurs. Il convient alors d'utiliser une fonction noyau Φ qui a pour rôle de projeter les objets des 2 classes dans un espace de plus grande dimensionnalité. Cette projection correspond en fait à une transformation non linéaire des descripteurs. Ce procédé permet de rendre des problèmes non linéairement séparables dans l'espace des descripteurs en problème linéairement séparable dans l'espace de redescription (voir Figure 1-6).

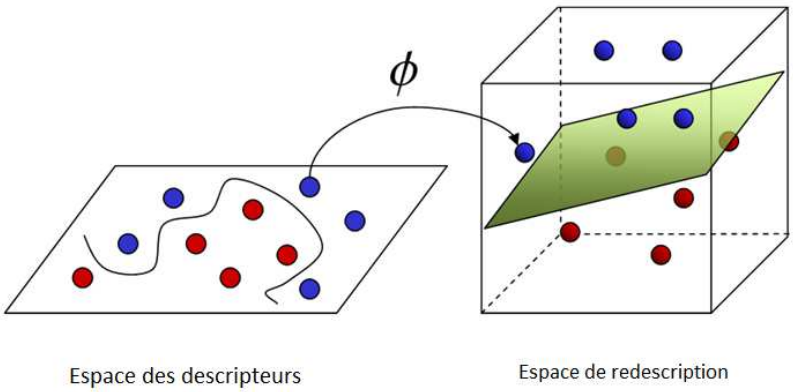


Figure 1-6. Projection non linéaire des objets de l'espace des descripteurs dans l'espace de redescription.

Le noyau de Tanimoto a été tout particulièrement été utilisé au cours de cette thèse. Il permet d'obtenir des performances similaires par rapport aux autres noyaux mais possède un nombre réduit de paramètres à optimiser ce qui rend son utilisation plus aisée. L'utilisation de ce noyau permet de calculer la distance entre les objets

dans l'espace de redescription comme la similarité de Tanimoto de ces objets dans l'espace des descripteurs.

Un facteur supplémentaire déterminant la tolérance de la SVM vis-à-vis des exemples mal classés peut être varié. Ce facteur nommé coût est proportionnel à la distance entre l'exemple mal classé et l'hyperplan. Le risque à trop augmenter le coût est de faire du sur-apprentissage. Mais ce risque est contrôlé par l'expressivité de la fonction noyau, c'est-à-dire la capacité à trouver un espace de redescription trop spécialisé pour le jeu d'entraînement.

La SVM donne la possibilité d'avoir des prédictions sous forme de probabilités [18]:

La sortie normale d'une SVM correspond à la distance d'un exemple avec l'hyperplan. Tout d'abord un partitionnement de l'espace de redescription est fait. Une fonction sigmoïde de la distance au plan est parfois utilisée pour calculer une probabilité. Si la probabilité est supérieure à 0,5, l'exemple appartient à la classe positive ; si elle est inférieure à 0,5 l'exemple appartient à la classe négative.

1.4.2. Arbre de décision

L'arbre de décision [19] est un modèle de classification supervisé qui divise successivement et le plus efficacement possible un jeu de données en sous-ensembles ne contenant (pratiquement) plus que des exemples d'une seule classe. La division des données en sous-ensembles est réalisée à l'aide de tests définis à l'aide des descripteurs.

La structure du modèle créé fait penser à celle d'un arbre : les feuilles de l'arbre représentent les étiquettes des classes, les nœuds correspondent aux attributs et les branches correspondent aux valeurs que prennent ces attributs.

L'algorithme général de construction d'un arbre est le suivant :

- Initialisation de l'arbre, on affecte le meilleur attribut à la racine.
- Pour chaque valeur de l'attribut à la racine on crée un nouveau nœud fils.
- On classe les exemples du jeu d'apprentissage dans les nœuds fils.
- Si tous les exemples d'un nœud fils sont homogènes on affecte leur classe au nœud, sinon on recommence à partir de ce nœud.

Le critère de sélection du meilleur attribut servant à séparer les exemples en sous-ensembles dépend de la méthode choisie, dans C4.5 par exemple le critère de sélection de l'attribut est le gain d'information lié à l'entropie.

Les arbres de décisions sont aussi soumis à des heuristiques variées visant à les simplifier. En effet le risque de sur-apprentissage est d'autant plus important que l'arbre est complexe. Cette tâche est appelée l'élagage.

Les arbres de décision sont souvent utilisés car ils ont l'avantage d'être rapide à construire, ils supportent le bruit et le modèle obtenu est en général facilement interprétable. Un exemple d'arbre est donné Figure 1-7.

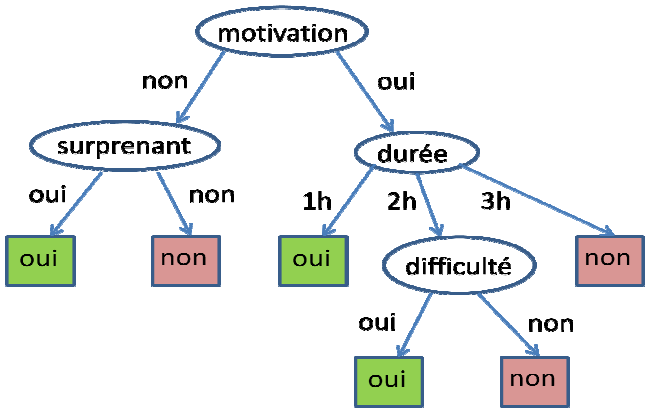


Figure 1-7. Exemple d'arbre de décision pour la question « Cette présentation est-elle intéressante ? »

Sous Weka, l'algorithme C4.5 est appelé J48. Dans la représentation des arbres de cette méthode chaque feuille sera associée à 3 valeurs : le 1er représente le label de la classe, le 2^{e} représente le nombre total d'instances dans la feuille (exemples correctement prédits et incorrectement prédits), et le 3^{e} nombre représente le nombre d'instances incorrectement prédites.

Par exemple, dans l'arbre précédent, pour le chemin motivation|non et surprenant|oui, supposons que la réponse donnée par le logiciel Weka serait « oui (10/2) ». Cela signifie que parmi les 10 personnes ayant trouvé la présentation surprenante mais manquant de motivation, la majorité a pensé que la présentation était intéressante, et 2 personnes l'ont trouvé inintéressante.

1.4.3. Forêt aléatoire (RF)

Une forêt aléatoire [20] consiste en un nombre défini (habituellement entre 100 et 200) d'arbres de décision utilisés pour calculer un vote à la majorité. L'objet sera prédit comme appartenant à la classe qui a reçu le plus de voix.

La réponse de chaque arbre dépend du sous-ensemble de descripteurs choisis aléatoirement. Le même nombre arbitraire de descripteurs est utilisé pour tous les arbres de la forêt. Chaque arbre est construit sans élagage et est donc largement biaisé en raison du sur-apprentissage.

Cette technique est généralement employée lorsque le nombre d'attributs à étudier devient très grand.

Il a d'ailleurs été démontré que cette technique tend vers une solution optimale pour un jeu de données. Quand le nombre d'arbres devient grand, le modèle obtenu est le meilleur qui puisse être trouvé, compte tenu du jeu de données soumis.

1.4.4. Bayésien Naïf (NB)

Le bayésien naïf [21-23] est une technique simple utilisant tous les descripteurs pour prendre une décision en leur donnant une contribution égale et en les considérant indépendant les un des autres. L'approche paraît simpliste étant donné que les attributs ne sont généralement pas indépendants, ni même d'importance égale dans les problèmes de la vie courante. Cependant l'approche fonctionne bien en pratique.

Le classeur bayésien se base sur la formule de Bayes :

$$\text{Eq. 1-1: } P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- P(h) est la probabilité à priori de l'hypothèse h, c'est-à-dire la probabilité à priori d'appartenir à une classe.
- P(D) est la probabilité à priori des données D, c'est-à-dire des descripteurs.
- P(h|D) est la probabilité à posteriori de h étant donné D.
- P(D|h) est la probabilité à posteriori de D étant donné h.

Etant donné 2 hypothèses h_1 (l'objet appartient à la classe 1) et h_2 (l'objet appartient à la classe 2), on cherchera généralement à trouver l'hypothèse la plus probable étant donné l'observation D. Si $P(h_1|D) > P(h_2|D)$ l'objet sera prédit comme appartenant à la classe 1, dans le cas contraire l'objet sera prédit comme appartenant à la classe 2.

La probabilité à priori des données D étant constante pour un jeu de données on peut éliminer P(D) de l'équation 1 en comparant $P(h_1|D)$ avec $P(h_2|D)$.

Il serait tentant de calculer directement les probabilités de h1 ou h2 directement sachant que l'exemple est décris par une série de descripteurs. En pratique, cette estimation est impossible : il n'existe souvent qu'une seule donnée correspondant à un ensemble de valeurs de descripteurs. Le recours à la formule de Bayes, change le problème de forme mais ne le résout pas. En revanche, la nouvelle formulation obtenue, moyennant une *hypothèse d'indépendance*, permet de décomposer le problème en probabilités élémentaires qui peuvent être estimées avec une meilleure précision.

Soit D = (d_1, d_2, d_3,…,d_n) où d_n correspond au descripteur n ; étant donné l'hypothèse d'indépendance des descripteurs utilisés dans cette méthode on a :

Eq. 1-2: $P(d1, d2, d3, \dots, dn|h) = \prod_{i=1}^{i=n} P(di|h)$

Finalement il suffit de comparer $P(h1)\prod_{i=1}^{i=n} P(di|h1)$ et $P(h2)\prod_{i=1}^{i=n} P(di|h2)$ pour déterminer quelle hypothèse est la plus probable étant donné l'observation. L'observation sera alors classée selon l'hypothèse la plus probable.

Dans le cas des descripteurs fragmentaux, par exemple, $P(d_1|h_1)$ correspondra à la fréquence d'apparition du fragment 1 dans les molécules du jeu d'apprentissage appartenant à la classe 1.

Toutefois les probabilités conditionnelles de chaque descripteur pris individuellement sont petites. Le produit des probabilités individuelles (Equation 1-2) tend donc très vite vers 0 et peut devenir numériquement instable. Pour obtenir des résultats numériques plus stables on calcule le logarithme de la probabilité d'appartenance d'un objet à une classe étant donné l'observation. Cela revient à sommer les logarithmes des probabilités conditionnelles de chaque descripteur.

Etant donné que le logarithme est une fonction convexe, la décision prise sur la base de ces logarithmes de probabilités sera la même que celle obtenue sans l'utilisation des logarithmes.

1.4.5. Stochastic QSAR Sampler (SQS)

SQS [24] est une approche de régression utilisant une méthode stochastique se basant sur un algorithme génétique [25].

L'algorithme génétique est une approche très simplifiée du monde de la génétique. Il fonctionne selon l'évolution Darwinienne des populations de taille définie. Chaque individu de la population est représenté par un chromosome codant un ensemble de caractéristiques liées (descripteurs). La codification se fait à travers des « gènes » qui définissent l'expression ou la non-expression d'un caractère.

L'apprentissage se fait en 2 étapes : l'initialisation et l'optimisation. Durant l'initialisation, une population non homogène de taille *Npop* est générée. C'est-à-dire que des sous-ensembles de descripteurs plus ou moins grand sont choisis aléatoirement. Pour chaque individu un modèle est développé et ses performances estimées. Durant l'optimisation, les individus maximisant une fonction d'adaptation F sont sélectionnés. L'algorithme utilise alors 2 opérateurs permettant de diversifier la population au cours des générations. L'opérateur de croisement qui permet d'échanger des chaînes entre 2 chromosomes, et l'opérateur de mutation qui permet de remplacer aléatoirement un gène par un autre. La population évolue ainsi à la suite de successions d'étapes de sélection et de création d'individus (sélection des variables, construction et validation de modèles de régression multilinéaire).

Ainsi, l'algorithme cherche à optimiser la fonction $F=R^2-\alpha.n$, où R^2 est le coefficient de détermination et α un paramètre de régularisation qui contraint les modèles obtenus à utiliser un faible nombre de variables n. L'algorithme ne garanti pas de trouver le meilleur modèle, mais il garanti de trouver un échantillon optimal des modèles les plus performants considérant la fonction d'adaptation. L'ensemble des étapes précédentes est répété jusqu'à ce que le modèle de régression souhaité soit obtenu. Le modèle qui est conservé dépend du nombre de paramètres calculés et du coefficient de détermination R^2 calculé en validation.

Dans le cadre d'une classification binaire, le label de classes recherchées est associé aux valeurs -1 et 1. On admettra alors que si le résultat du modèle de régression construit est supérieur à 0, l'objet est prédit 1 ; dans le cas inverse l'objet sera prédit comme appartenant à la classe -1.

1.4.6. JRip

JRip est une classe qui implémente un algorithme d'apprentissage de règles nommé RIPPER [26](Repeated Incremental Pruning to Produce Error Reduction) et proposé par William W. Cohen. C'est une méthode qui utilise l'approche diviser pour régner afin de construire itérativement des règles qui couvrent les exemples encore non couvert pas les règles précédentes. Chaque règle est construit en ajoutant des conditions jusqu'à ce qu'il n'y ait plus aucun exemple de la classe négative qui soit couvert. A chaque étape la condition qui permet d'obtenir le plus grand gain d'information est ajoutée à la règle. La génération des règles s'effectue de la classe la moins peuplée à la classe la plus peuplée. Avec cette approche il est donc aisé d'apprendre des règles sur les classes minoritaires. Les règles obtenues sont en général nombreuses et complexes. C'est pourquoi elles sont soumises à une heuristique de simplification comme pour les arbres : un élagage.

Plus précisément, RIPPER divise le jeu de données en un ensemble d'apprentissage et d'élagage pour décider si la dernière condition d'une règle doit être supprimée. Il intègre aussi une heuristique basée sur la taille minimale de description (en bits) comme critère d'arrêt : si l'ensemble des règles générées a une taille supérieure de 64 bits au moins au plus petit ensemble de règles trouvées alors l'apprentissage s'arrête. Une fois l'ensemble de règles obtenues, le jeu entier est à nouveau divisé aléatoirement en un jeu d'apprentissage et d'élagage, et une nouvelle procédure d'élagage plus importante est mise place. Cette procédure revisite voir remplace chaque règle individuelle initiale de façon à réduire l'erreur de l'ensemble de règles sur le jeu d'élagage. L'algorithme décide alors s'il conserve la règle initiale ou s'il prend la règle revisitée ou remplacée en fonction de la taille de description de l'ensemble de règles.

L'algorithme est fourni en annexe 10.1.

1.4.7. Perceptron Votant (VP)

Le perceptron [27] est la forme la plus simple du réseau de neurones servant à résoudre les problèmes linéairement séparables. Ce type de méthode s'inspire du fonctionnement du cerveau humain pour simuler la réponse d'un neurone vis-à-vis des stimulations qui sont opérées. Dans le cas présent les stimulations correspondent aux vecteurs de descripteurs. Chaque stimulation sera reliée au neurone de sortie à l'aide d'un certain poids. Ces poids sont des paramètres adaptatifs dont la valeur est à déterminer en fonction du problème via un algorithme d'apprentissage. Le résultat de la somme des stimulations pondérées par leur poids est alors soumis à une fonction d'activation qui déterminera la réponse du neurone de sortie. Dans le cas d'une fonction d'activation simple, si cette somme est supérieure à un certain seuil d'activation, défini par la fonction d'activation, alors la réponse du neurone de sortie sera positive.

Dans le cas d'un problème de classification à 2 classes, la valeur des descripteurs sera donc pondérée et la réponse du neurone de sortie sera généralement -1 ou 1.

Soit f une fonction d'activation, w_i le poids affecté au descripteur i et x_i la valeur du descripteur i ; le fonctionnement du perceptron se traduit par l'équation suivante :

$$\text{Eq. 1-3: } f(-1,1) = \begin{cases} 1, & \sum_{i=1}^{n} w_i x_i \geq 0 \\ -1, & sinon \end{cases}$$

Le fonctionnement du perceptron peut finalement être schématisé de la façon suivante :

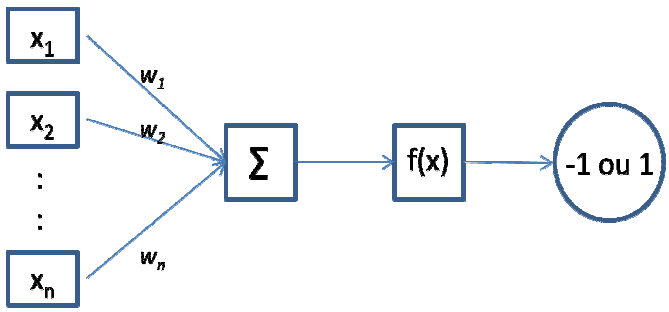


Figure 1-8. Schéma général du perceptron.

Dans le cas du perceptron votant tous les états intermédiaires du perceptron pendant l'apprentissage sont stockés. Les votants obtenus votent à la majorité mais leur poids dépend de leur durée de vie : ceux qui ont pu prédire beaucoup de cas pendant l'apprentissage, avant d'être mis à jour, ont plus de poids.

1.4.8. Réseaux de neurones

Contrairement au simple perceptron, le réseau de neurones [28] peut résoudre des problèmes non linéairement séparables. Du point de vue de l'architecture on retrouve des neurones d'entrée et un ou plusieurs neurones de sortie. La différence avec le perceptron repose sur le fait que le réseau de neurones possède une ou plusieurs couches cachées entre les neurones d'entrée et sortie. C'est pourquoi le réseau de neurones est aussi appelé perceptron multicouches. Chaque neurone de la couche d'entrée envoi un signal pondéré à tous les neurones de la couche cachée qui eux-mêmes envoient un signal pondéré à tous les neurones de la couche de sortie. Les poids de ces signaux sont optimisés à l'aide d'un algorithme d'apprentissage. L'architecture classique d'un réseau de neurones est présenté Figure 1-9.

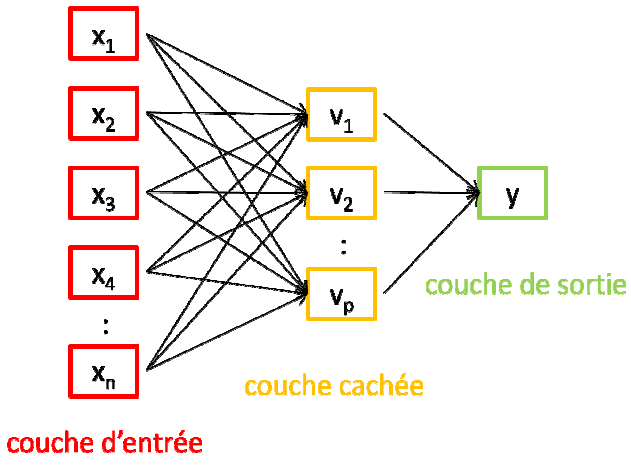


Figure 1-9. Schéma général d'un réseau de neurones à 3 couches.

Les réseaux de neurones permettent de faire de l'apprentissage mono-tâche (Single Task Learning) lorsqu'un seul paramètre doit être prédit, et de l'apprentissage multitâches (Multi Task Learning) lorsque plusieurs paramètres de sortie doivent être

prédit simultanément. Cette dernière approche doit permettre d'améliorer les performances de la modélisation lorsqu'il existe une dépendance entre les paramètres.

2 programmes ont été utilisés pour faire du STL et du MTL : l'association de réseaux de neurones (ASNN) et AMORE.

L'ASNN [29, 30] développé par le Dr. Igor Tetko correspond à la moyenne des prédictions de 100 réseaux de neurones développés à partir du même jeu d'entraînement et des mêmes descripteurs. La subtilité de l'approche vient du fait qu'une correction supplémentaire est faite sur la valeur moyenne trouvée pour les 100 modèles. Cette correction utilise le principe des k voisins les plus proches sur les modèles. Le coefficient de corrélation de Pearson mesure la similarité des vecteurs de prédictions entre les objets afin de déterminer le nombre de voisins les plus proches pour chaque nouvel objet prédit.

La prédiction moyenne d'un nouvel objet *i* est alors corrigée selon la formule suivante :

Eq. 1-4: $\overline{y'}_i = \bar{y}_i + \frac{1}{k}\sum_{j \in N_k(i)}(y_{exp,j} - \bar{y}_j)$

$Y_{exp,j}$ correspond à la valeur expérimentale de la propriété, et la sommation est faite sur les k objets voisins les plus proches de i. Le facteur correctif est donc calculé à partir du jeu d'entraînement.

Le jeu de données est découpé aléatoirement, et à part égale, en un jeu d'entraînement et un jeu de validation interne. Le jeu de validation interne a pour but d'arrêter l'apprentissage avant de faire du sur-apprentissage.

AMORE (A MORE flexible neural network package) est un paquet fonctionnant sous l'environnement R. Il permet de construire des réseaux de neurones flexibles dans le sens où les paramètres du réseau sont directement accessibles et les fonctions disponibles sont personnalisables. Comme les réseaux de neurones classiques il est possible de construire un réseau multicouche avec le nombre de

neurones souhaité dans chaque couche et d'optimiser les poids de chaque connexion entre les neurones.

1.4.9. K-Means

K-Means [31] est l'un des algorithmes d'apprentissage non supervisé les plus simples proposant de faire du clustering. La méthode permet de créer un certain nombre k de clusters défini par l'utilisateur.

Pour chaque cluster un centroïde est défini à partir du jeu d'apprentissage de façon à avoir k centroïdes les plus éloignés possible. Les autres exemples du jeu d'apprentissage sont alors placés dans le cluster avec le centroïde le plus similaire. Quand tous les exemples sont classés, le centroïde de chaque cluster est redéfini comme le barycentre de celui-ci. Après avoir défini les k nouveaux centroïdes, les exemples du jeu d'apprentissage sont à nouveau placés dans le cluster pour lequel la similarité avec le centroïde est la plus grande. On répète les étapes précédentes jusqu'à ce que les centroïdes ne bougent plus.

Finalement l'algorithme a pour but de minimiser la fonction F tel que :

$$\text{Eq. 1-5: } F = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2$$

$||x_i^j - c_j||^2$ correspond à une mesure de la distance entre un exemple x appartenant au cluster j et le centroïde de ce cluster.

1.4.10. Cartes de Kohonen

Les cartes de Kohonen [32] ou cartes auto-organisatrices ont été développées en 1984 par Teuvo Kohonen. Cette approche est une méthode d'apprentissage non supervisée de quantification vectorielle. Il s'agit de regrouper les informations de l'espace d'entrée en classes sur une carte tout en respectant la topologie de l'espace d'observation. La carte est en général à 2 dimensions et de topologie (carrée, rectangulaire) et voisinage (rectangulaire, hexagonale) variable. Comme les réseaux de neurones, chaque neurone de la couche d'entrée est relié à chaque neurone de la

carte de Kohonen par un certain poids. Un neurone de la carte possède donc des coordonnées fixes sur la carte, et des coordonnées adaptables dans l'espace d'entrée original. L'apprentissage adapte les poids entre les neurones de telle manière que des exemples proches dans l'espace des descripteurs soient associés au même neurone ou a des neurones proches sur la carte.

La procédure de l'algorithme est la suivante : avant de commencer l'entraînement, les poids W sont aléatoirement assignés. A chaque itération, un exemple X(t) pris au hasard dans le jeu de données est présenté à la carte. Le neurone de la carte dont le vecteur poids W(t) est le plus proche du vecteur d'entrée X(t) est défini comme le neurone gagnant. Les poids du neurone gagnant ainsi que de ses neurones voisins sont mis à jour. L'adaptation des poids est d'autant plus forte que les neurones voisins sont proches du neurone gagnant. La modification des poids dépend d'un coefficient d'apprentissage qui est une fonction linéaire décroissante. Ce coefficient a une valeur qui décroit aussi pour les neurones voisins, de manière à les spécialiser un peu moins que le neurone vainqueur. Le principe de la carte de Kohonen est schématisé ci-dessous :

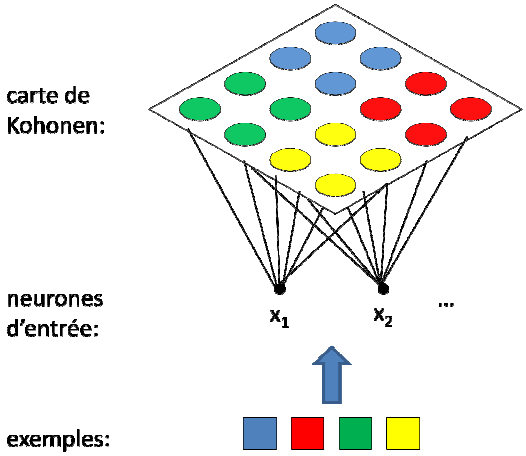


Figure 1-10. Principe d'une carte de Kohonen.

1.4.11. Modèles consensus

Les prédictions faites entre plusieurs modèles peuvent parfois être différentes malgré le fait que les modèles atteignent des performances semblables. Afin d'obtenir de meilleures performances finales et d'obtenir des prédictions plus robustes, un modèle consensus peut être construit. Dans le cas de problèmes de

classification, le modèle consensus correspond à un vote à la majorité entre un certain nombre de modèles choisis. Il existe plusieurs moyens de créer des modèles consensus, mais trois sont particulièrement utilisés dans ce travail :

- Le vote à la majorité est fait à partir de modèles obtenus sur le même jeu d'entraînement, le même jeu de descripteurs, et le même algorithme d'apprentissage. Les seules variations entre les modèles sont les paramètres propres à l'algorithme d'apprentissage comme l'initialisation des poids dans les réseaux de neurones (cf. ASNN).
- Le vote à la majorité est fait à partir de modèles obtenus sur le même jeu d'entraînement, le même algorithme d'apprentissage, mais des jeux de descripteurs différents. Cette technique est souvent appliquée dans le cas de modèles obtenus à partir de descripteurs fragmentaux variés.
- Le vote à la majorité est fait à partir de modèles obtenus sur le même jeu d'entraînement, mais des jeux de descripteurs différents et des algorithmes d'apprentissages différents.

1.5. Critères d'évaluation des modèles

Afin de juger les performances des modèles de classification développés, des critères d'évaluation doivent être utilisés. Il s'agit de comparer les valeurs prédites par les modèles avec les valeurs expérimentales. Dans le cas de la classification les valeurs correspondent aux labels des classes.

La plupart des critères statistiques sont calculés à l'aide de la matrice de confusion. C'est une matrice qui recense pour chaque classe le nombre d'erreurs faites ainsi que le nombre de prédictions correctes. Dans le cas d'une classification binaire, la matrice de confusion s'établit comme suit :

		Classe estimée	
		Positive (1)	Négative (0)
Classe réelle	Positive (1)	Vrais Positifs	Faux Négatifs
	Négative (0)	Faux Positifs	Vrais Négatifs

Figure 1-11. Matrice de confusion pour 2 classes.

Il est évident que le but est d'avoir un maximum de Vrais Positifs (VP) et de Faux Négatifs (VN), et un minimum de Faux Positifs (FP) et de Faux Négatifs (FN). A partir de cette matrice de confusion on peut calculer un certain nombre d'indicateurs en fonction d'importances relatives des différents types de succès et d'erreurs aux yeux du modélisateur :

Le rappel des exemples de la classe négative/positive, encore appelé spécificité/sensibilité, représente la capacité d'un modèle à prédire correctement tous les objets d'une classe. Il est calculé selon la formule suivante :

$$\text{Eq. 1.6: } Rappel(0) = \frac{VN}{VN+FP}$$

$$\text{Eq. 1.7: } Rappel(1) = \frac{VP}{VP+FN}$$

L'équation 1.7 est aussi appelé taux de vrais positifs. La valeur 0 est un échec total et la valeur 1 est un succès total pour retrouver les positifs.

Il est possible de calculer le rappel moyen des 2 classes pour avoir un critère plus global sur cette statistique. Ce rappel moyen est nommé Précision Balancée (PB) et prend des valeurs comprises entre 0,5 et 1.

$$\text{Eq. 1.8: } PB = \frac{Rappel(0)+Rappel(1)}{2}$$

La précision représente la capacité du modèle à prédire dans une classe uniquement les objets appartenant à cette classe. La valeur 0 est un échec total et la valeur 1 est un succès total.

$$\text{Eq. 1.9: } Précision\ (0) = \frac{VN}{VN+FN}$$

$$\text{Eq. 1.10: } Précision\ (1) = \frac{VP}{VP+FP}$$

Il est possible de définir une précision globale pour le modèle :

$$\text{Eq. 1.11: } Précision = \frac{VP+VN}{VP+VN+FP+FN}$$

Quand le modèle fournit un score d'appartenance (par exemple une probabilité) à une classe, 2 stratégies peuvent être employées. La première stratégie revient à nominaliser le score : on définit un seuil au dessus duquel l'objet appartiendra à la classe positive et au dessous duquel l'objet appartiendra à la classe négative. Les critères d'évaluation décrits précédemment peuvent alors être utilisés.
La deuxième stratégie consiste à utiliser de nouveaux critères statistiques : l'aire sous la courbe ROC (AUC) et le critère IAP (Independent Accuracy of Prediction).

La courbe ROC (Receiver Operating Characteristics) correspond au tracé de la proportion des vrais positifs en fonction de la proportion des faux positifs lorsque l'on fait varier le seuil du score d'appartenance à une classe. On peut alors calculer l'aire sous cette courbe (AUC). Une valeur AUC élevée reflète une nette distinction du modèle entre les objets des 2 classes. Une AUC égale à 0,5 correspond à des performances du modèle équivalentes à un tirage au hasard des objets. Une AUC inférieure à 0,5 signifie que le modèle fait moins bien que le hasard. La Figure 1-12 illustre ces propos.

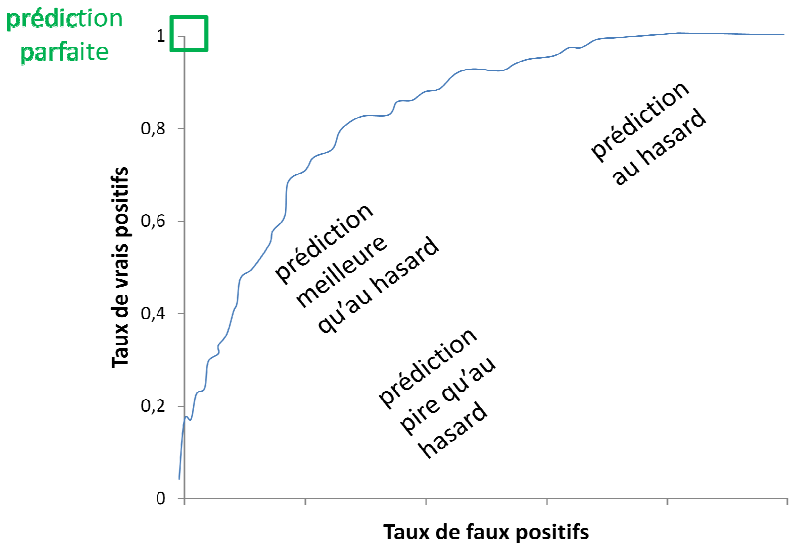


Figure 1-12. Interprétation de la courbe ROC.

Le cas idéal est bien évidemment que l'aire sous la courbe ROC soit égale à 1. On a dans ce cas une prédiction parfaite.

Le critère IAP (Independent Accuracy of Prediction) est une statistique variant entre 0 et 1 et qui estime, comme la valeur AUC, la capacité du modèle à donner une probabilité plus forte aux vrais positifs par rapport aux faux positifs. L'IAP se calcule selon l'équation suivante :

$$\text{Eq. 1.12: } IAP = \frac{Nombre\ de\ P(C_+)>P(C_-)}{(Nombre\ de\ C_-)\times(Nombre\ de\ C_+)}$$

Nombre de $P(C_+)>P(C_-)$ est le nombre de cas pour lesquels un vrai positif a une probabilité supérieure à un faux positif. Toutes les combinaisons vrai-faux positifs sont comparées. Le *Nombre de C-* et le *Nombre de C+* sont respectivement le nombre de faux et vrais positifs. Un exemple est proposé

Tableau **1-1** afin d'illustrer le calcul de ce critère.

Tableau 1-1. Exemple de calcul du critère IAP.

rang	label réel	probabilité prédite
1	1	0,7
2	0	0,6
3	1	0,4
4	0	0,3

IAP = (2+1)/(2x2) = 0,75

Dans le cas de modèles de classification non supervisée on utilisera le calcul de la pureté globale donnée 1.13. La pureté globale permet d'avoir une estimation quant à la pureté des différents nœuds/clusters du modèle. Elle se calcule comme la moyenne pondérée de la pureté de chaque nœud/cluster du modèle. La pureté d'un cluster sert à déterminer si une seule classe d'objets est représentée dans le cluster ou si le contenu de celui est hétérogène.

$$\text{Eq 1.13: } pureté = \sum_{i=1}^{K} \frac{m_i}{m} \times max\{p_{ij}\}_{j\in m_i}$$

- m_i : nombre d'objets dans le cluster i
- m : nombre total d'objets
- K : nombre de clusters
- pij : rapport entre le nombre d'objets dans le cluster i appartenant à la classe j et le nombre total d'objets dans le cluster i.

1.6. Validation des modèles

Les modèles créés sur le jeu d'apprentissage doivent constamment être validé [33] sur un jeu de test. En effet, il est possible d'obtenir de très bonnes performances sur le jeu d'apprentissage et des performances beaucoup moins bonnes sur le jeu de test. Ce phénomène est principalement dû à ce qu'on appelle le sur-apprentissage. La procédure la plus courante utilisée pour tester les modèles et éviter le sur-apprentissage est la validation croisée. Une procédure supplémentaire nommée scrambling doit aussi servir à valider le modèle afin de vérifier que le modèle obtenu n'est tout simplement pas du à la chance.

1.6.1. Validation croisée à n paquets

La validation croisée à n paquets (n-folds cross-validation) est une technique qui consiste à découper aléatoirement un jeu de données en n paquets contenant sensiblement le même nombre d'objets. Un jeu d'entraînement de n-1 paquets est alors utilisé pour construire le modèle et le paquet restant sert à estimer les performances du modèle. Tour à tour, tous les paquets du jeu initial découpé sont prédits. Cette technique permet donc de prédire tous les objets du jeu initial.
Un exemple d'une validation croisée à 5 paquets est présenté sur la Figure 1-13.

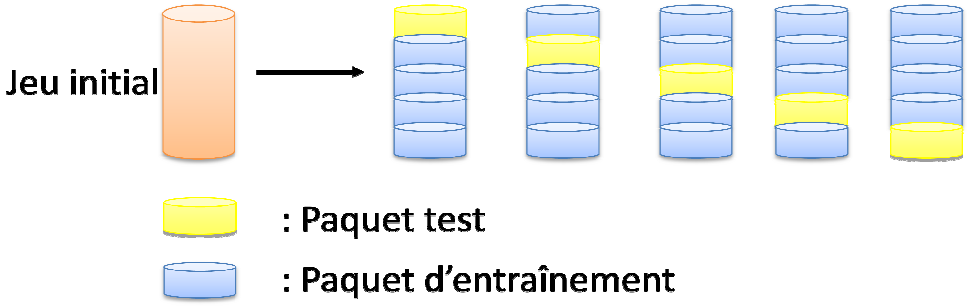


Figure 1-13. Procédure de validation croisée à 5 paquets.

1.6.2. Scrambling ou Y-randomization

Dans certains cas, les modèles construits ont des performances moins bonnes sur le jeu de test malgré le fait que le modèle ne fasse apparemment pas de sur-apprentissage. Ces cas peuvent s'expliquer par l'obtention de modèles dits fortuits. Il

existe en fait une corrélation fortuite entre les descripteurs et la propriété qui permet d'obtenir de bons modèles.

Afin d'évaluer la part de chance dans les modèles construits, une technique appelée scrambling ou encore y-randomization est utilisée. Cette procédure consiste à mélanger aléatoirement la propriété à modéliser dans le jeu initial et à recréer des modèles en validation croisée. Si les modèles sont affectés par des corrélations fortuites on devrait voir apparaître de bonnes performances en validation croisée malgré le fait d'avoir mélangé la propriété au hasard. Dans le cas contraire la performance des modèles devrait être équivalente à celle d'un modèle faisant des prédictions au hasard.

Statistiquement cela revient à tracer la loi normale qui approche au mieux la distribution du critère statistique choisi pour estimer les performances des modèles construits sur le scrambling. Il suffit alors de déterminer la valeur de ce critère statistique pour la borne supérieure d'un intervalle de confiance, par exemple à 95%. Si la valeur est supérieure à celle obtenue par le modèle avant le scrambling, la part de chance est trop importante pour utiliser le modèle initial sans risque. Au contraire, si la valeur est bien inférieure à celle obtenue par le modèle avant le scrambling, le modèle initial peut être utilisé pour faire des prédictions. Une figure présentant le résultat des étapes précédentes est proposé Figure 1-14.

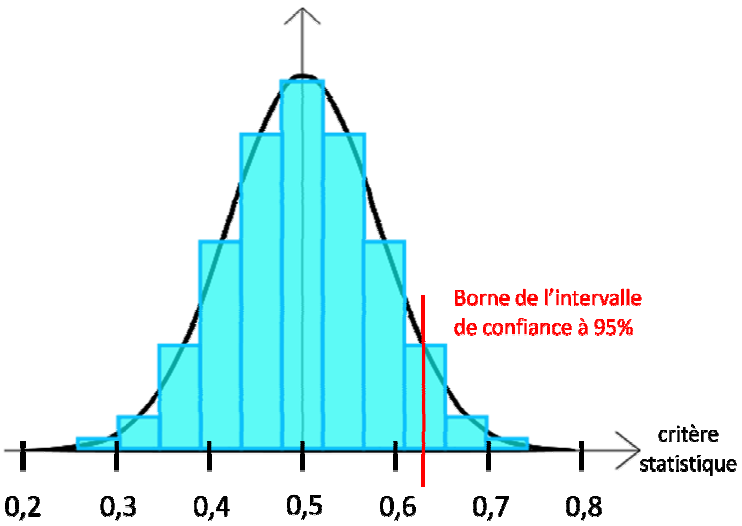


Figure 1-14. Loi normale de la distribution des critères statistiques obtenus en scrambling.

1.7. Domaine d'applicabilité

Un modèle idéal est un modèle capable d'émettre une prédiction pour n'importe quelle molécule imaginable. Cependant cela est souvent loin d'être possible. Les modèles sont construits dans un espace chimique limité dû à la taille limitée du jeu d'entraînement. Lorsqu'une molécule se situe en dehors de cet espace chimique, la prédiction qui est faite sur celle-ci n'est alors plus fiable. Pour prévenir ce type de problème un domaine d'applicabilité est souvent utilisé. Cette stratégie permet d'éliminer du jeu de test les molécules se situant en dehors de l'espace chimique du jeu d'entraînement.

Le domaine d'applicabilité « contrôle des fragments » est généralement utilisé dans cette thèse. Comme son nom l'indique, ce DA s'applique dans le cas de descripteurs fragmentaux. Si une molécule du jeu de test possède de nouveaux fragments par rapport au jeu de fragments générés pour le jeu d'entraînement alors la molécule n'est pas prédite. Les molécules non prédites dépendront donc du jeu de descripteurs choisi pour construire le modèle.

1.8. Conclusion

Dans cette première partie, les différentes techniques permettant de construire un modèle QSAR ont été détaillées. Après nettoyage préalable des données, les descripteurs calculés à partir de la structure des molécules doivent permettre de trouver des relations entre celles-ci et les propriétés souhaitées. Cette relation est identifiée par une fonction calculée à l'aide de différentes méthodes d'apprentissage. On note les méthodes dites quantitative (régression) et qualitative (classification). La procédure de validation croisée permet d'estimer les performances des modèles tandis que le scrambling permet de rendre compte de la part de chance dans la construction des modèles. Finalement, afin de prédire de nouvelles molécules avec fiabilité, un domaine d'applicabilité peut être mis en place pour éliminer les molécules trop exotiques.

1.9. Références

1. Young, D., et al., *Are the Chemical Structures in Your QSAR Correct?* QSAR & Combinatorial Science, 2008. **27**(11-12): p. 1337-1345.
2. Fourches, D., E. Muratov, and A. Tropsha, *Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research.* J. Chem. Inf. Model., 2010. **50**(7): p. 1189-204.
3. Allinger, N.L., et al., *Chimie Organique - volume II: réactions*. 1979.
4. Hoonakker, F., *Graphes condensés de réactions, applications à la recherche par similarité, la classification et la modélisation.*, in Chimie2008, Université de Strasbourg. p. 252.
5. Todeschini, R., et al., *Handbook of Molecular Descriptors*. 2000: Wiley-VCH.
6. Wiener, H., *Structural Determination of Paraffin Boiling Points.* J. Am. Chem. Soc., 1947. **69**(1): p. 17-20.
7. Randic, M., *Characterization of molecular branching.* J. Am. Chem. Soc., 1975. **97**(23): p. 6609-6615.
8. Pearlman, R.S. and K.M. Smith, *Metric Validation and the Receptor-Relevant Subspace Concept.* J. Chem. Inform. Comput. Sci., 1999. **39**(1): p. 28-35.
9. Norinder, U., *Recent progress in CoMFA methodology and related techniques.* Perspect. Drug Discovery Des., 1998. **12-14**(0): p. 25-39.
10. von Itzstein, M., et al., *Rational design of potent sialidase-based inhibitors of influenza virus replication.* Nature, 1993. **363**(6428): p. 418-423.
11. Varnek, A., et al., *ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors.* Curr. Comput. Aided Drug Des., 2008. **4**(3): p. 191-198.
12. Varnek, A., et al., *Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures.* J. Comput. Aided Mol. Des., 2005. **19**(9-10): p. 693-703.
13. Ruggiu, F., et al., *ISIDA Property-Labelled Fragment Descriptors.* Mol. Inf., 2010. **29**(12): p. 855-868.
14. *ChemAxon Pmapper*. 2009; Available from: http://www.chemaxon.com/jchem/doc/user/PMapper.html.
15. *Molecular Operating Environment (MOE)*, 2009, Chemical Computing Group.

16. Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. 2000: Cambridge University Press.
17. Vapnik, V., *Statistical learning theory*. 1998: Wiley.
18. Platt, J. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. in ADVANCES IN LARGE MARGIN CLASSIFIERS. 1999.
19. Kingsford, C. and S. Salzberg, *What are decision trees?* Nature Biotechnology, 2008. **26**(9): p. 1011-1013.
20. Breiman, L., *Random Forests.* Mach. Learn., 2001. **45**(1): p. 5-32.
21. Elkan, C., Naive Bayesian Learning. Adapted from Technical Report No. CS97-557, Department of Computer Science and Engineering, University of California, San Diego, 1997.
22. John, G. and P. Langley. *Estimating Continuous Distributions in Bayesian Classifiers*. in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. 1995.
23. Zhang, H. *The Optimality of Naive Bayes*. in Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004*)*. 2004. AAAI Press.
24. Horvath, D., et al., *Stochastic versus Stepwise Strategies for Quantitative Structure−Activity Relationship Generation - How Much Effort May the Mining for Successful QSAR Models Take?* J. Chem. Inf. Model., 2007. **47**(3): p. 927-939.
25. Goldberg, D., *Genetic Algorithms in Search, Optimization, and Machine Learning*. 1989: Addison-Wesley Professional.
26. Cohen, W. *Fast Effective Rule Induction*. in *In Proceedings of the Twelfth International Conference on Machine Learning*. 1995.
27. F, F.R., *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychol. Rev., 1958. **65**(6): p. 386-408.
28. Haykin, S., *Neural Networks: A Comprehensive Foundation (2nd Edition)*. 1998: Prentice Hall.
29. Tetko, I.V., *Associative Neural Network.* Neural Process. Lett., 2002. **16**(2): p. 187-199.

30. Tetko, I.V., *Neural Network Studies. 4. Introduction to Associative Neural Networks.* J. Chem. Inf. Comput. Sci., 2002. **42**(3): p. 717-728.
31. Hartigan, J.A., *Clustering Algorithms (Probability & Mathematical Statistics)*. John Wiley & Sons Inc.
32. Kohonen, T., *The self-organizing map.* Proceedings of the IEEE, 1990. **78**(9): p. 1464-1480.
33. Tropsha, A., P. Gramatica, and V. Gombar, *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.* QSAR Comb. Sci., 2003. **22**(1): p. 69-77.

2. Bases de données.

Cette 2e partie a pour but de présenter les bases de données en ligne exploitées dans le cadre de cette thèse.

2.1. Kyoto Encyclopedia of Genes and Genomes (KEGG)

La KEGG [1, 2] (Kyoto Encyclopedia of Genes and Genomes) est une collection de base de données en ligne regroupant des données portant sur le génome (virus, règne animal et végétal), les voies enzymatiques et les molécules chimiques à intérêt biologique. La KEGG est considérée par ses développeurs comme une représentation informatique des systèmes biologiques. Elle peut être utilisée pour la modélisation et la simulation, la navigation et la récupération de données. On note 3 bases de données principales dans la KEGG : Ligand, Genes et Pathway. La première a été principalement utilisée dans le cadre de cette thèse pour récupérer un ensemble de réactions métaboliques. Elle contient les informations sur les composés biochimiques ainsi que les réactions et les enzymes associées. A ce jour, KEGG Ligand répertorie 16844 composés et 9107 réactions. L'information génétique est quant à elle stockée dans la base de données KEGG *Genes*, elle représente une collection de catalogues génétiques pour tous les génomes complètement séquencés et pour certains qui le sont partiellement. L'information fonctionnelle est stockée dans la base de données *Pathway* qui contient des représentations graphiques des processus cellulaires, comme le transport membranaire, la transduction de signal, la métabolisation des xénobiotiques etc. Une capture d'écran du site de la KEGG est donné Figure 2-1.

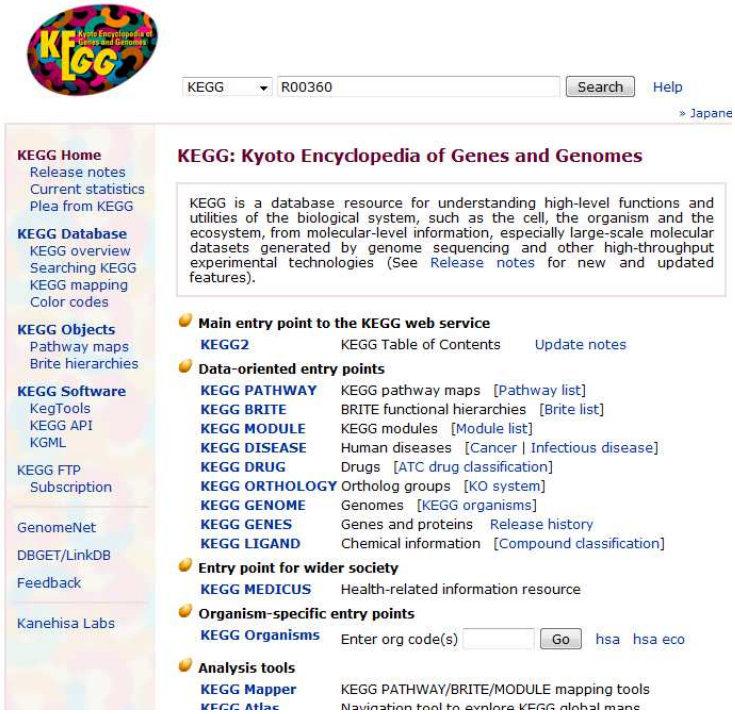


Figure 2-1. Capture d'écran du site de la KEGG.

Dans notre cas, les réactions métaboliques extraites sont identifiées par un nombre appelé le nombre EC [3] (Enzyme Commission). Ce nombre est souvent employé simultanément comme un identifiant de réactions, d'enzymes et gènes d'enzyme. L'assignement des nombres EC aux nouvelles enzymes est réalisé en se basant sur les données expérimentales publiées qui incluent l'entière caractérisation des enzymes et de leur fonction catalytique.

Le nombre EC est constitué de quatre nombres. Le 1er indique le type de réaction catalysé (voir tableau 1), le 2e le substrat général impliqué dans la réaction, le 3e le substrat spécifique impliqué et le 4e le numéro de série de l'enzyme.

Tableau 2-1. Signification du premier chiffre des 4 nombres de l'EC.

1er nombre	réaction catalysée
X = 1: oxydoréductases	oxydoréduction
X = 2: transférases	transfert de groupes
X = 3: hydrolases	hydrolyse
X = 4: lyases	addition de groupe à des atomes engagés dans des doubles liaisons
X = 5: isomérases	isomérisation (de position de groupe ou de fonction)
X = 6: ligases	condensation de deux molécules

Le nombre 1.1.3.X par exemple signifiera qu'on a une réaction de type oxydoréductase avec un groupement >CH-OH comme donneur d'hydrogène et O_2 comme accepteur d'hydrogène. Un exemple du résultat de la recherche dans la KEGG d'une telle réaction est donné Figure 2-2. Outre le nombre E.C., chaque réaction est définie par un nombre propre à la KEGG sous la forme Rxxxxx, où x est un chiffre. La base de données est gratuitement téléchargeable depuis le site sous condition d'être enregistré. Les composés sont disponibles au format *mol* et les réactions sont énumérées dans un fichier texte avec le sens de la réaction et l'identifiant des composés composant une réaction. Il a donc fallu créer un script qui permet de créer des fichiers de réactions *rxn/rdf* à partir des fichiers mol de chaque composé.

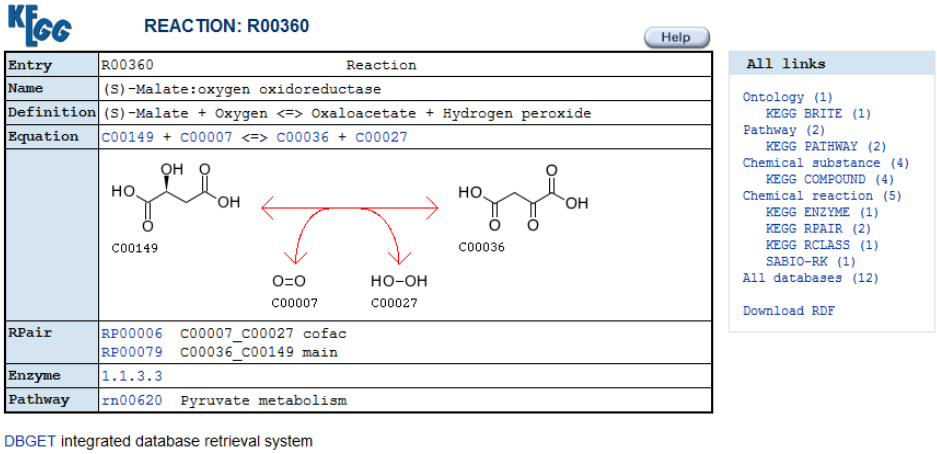


Figure 2-2. Réaction d'oxydation du (S)-Malate en Oxaloacetate (R00360).

2.2. Metabolite

La MDL Metabolite Database est une base de données appartenant à la société Accelrys [4]. Elle correspond à une collection de voies métaboliques pour les xénobiotiques et les biotransformations (en particulier pour les médicaments) ainsi que des données expérimentales in vivo et in vitro issues de la littérature, des conférences, et des « New Drug Applications ». Ce dernier correspond aux informations remises par une entreprise à la FDA [5](Food and Drug Administration) dans le cadre de la mise sur le marché d'un médicament. La base de données contient environ 50'000 composés issus de 20'000 composé parents pour un total de

100'000 transformations. La base de données est interrogée sous la forme de requêtes visant à déterminer l'espèce considérée, l'isoenzyme, le type de biotransformations etc.

Un exemple d'une recherche de biotransformations de type hydroxylation aromatique induites par l'isoenzyme CYP1A2 chez l'homme est donnée Figure 2-3. Les résultats d'une requête sont facilement extractibles en fichier *sdf/rdf*.

Figure 2-3. Exemple d'une recherche de réactions métaboliques par critères dans Metabolite.

2.3. Toxnet

Toxnet [6] est un site internet regroupant des bases de données orientées vers la toxicologie, les produits chimiques dangereux, la santé environnementale et les rejets toxiques. Le site constitue la principale ressource d'informations de l'*U.S. National Library of Medecine (NLM)* mais est aussi composé de bases de données additionnelles d'autres sources. Parmi les bases de données disponibles sur le site on considère tout particulièrement :

- La HSDB (Hazardous substances data bank) qui contient pour chaque molécule présente 150 champs représentant des informations sur les effets sur la santé chez l'homme, la toxicité chez l'animal, des données de

métabolisme/pharmacocinétiques, etc. La base de données est une compilation de plusieurs sources de la littérature allant de la compilation de bases de données déjà existantes aux articles scientifiques et documents gouvernementaux. La particularité de cette base de données est qu'elle est en plus manuellement revue par des experts dans les domaines spécifiés.

- La CTD (Comparative Toxicogenomics Database) qui est une base externe à la NLM qui renseigne sur la relation entre les composés chimiques, les gènes, et les maladies chez l'homme. Cette base de données développée à l'université de Caroline du Nord contient environ 300'000 liens composé-maladie pour 6000 composés chimiques. Parmi les liens proposés on note ceux qui ont été vérifiés/corrigés par un expert des données et ceux qui ne l'ont pas encore été. Un composé peut avoir un lien thérapeutique avec une maladie ou être considéré comme un mécanisme entrainant la maladie.

- TOXLINE (Toxicology Literature Online) qui contient la plus grande collection d'information bibliographique en ligne de la NLM. Cette collection couvre les effets biochimiques, pharmacologiques, physiologiques, et toxicologiques des médicaments et autres composés chimiques. On décompte environ 3 millions de citations bibliographiques avec le numéro CAS des composés qui y sont référencés.

La recherche dans TOXNET se fait à l'aide de mots clés se référant à des composés chimiques, des maladies ou symptômes, les espèces concernées, etc. Un exemple de recherche de médicaments induisant des dommages au foie est donné Figure 2-4.

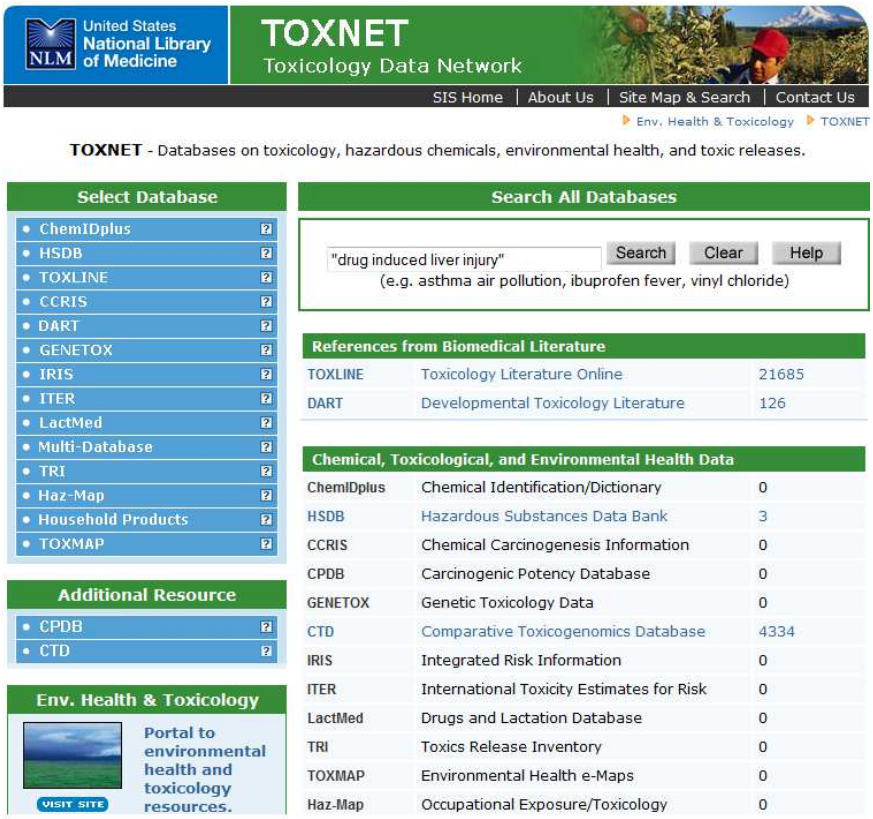


Figure 2-4. Exemple d'une recherche de médicaments induisant des dommages au foie via TOXNET.

2.4. Références

1. Kanehisa, M., et al., KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res., 2012. 40(D1): p. 109-114.
2. Kanehisa, M. and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res., 2000. 28(1): p. 27-30.
3. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes by the reactions they catalyse. cited April 2009; Available from: http://www.chem.qmul.ac.uk/iubmb/enzyme/.
4. Accelrys, Inc., Accelrys Metabolite, San Diego, 2009 (http://accelrys.com).
5. FDA, U.S. Food and Drug Administration. Available from: http://www.fda.gov/.

6. Wexler, P., TOXNET: An evolving web resource for toxicology and environmental health information. Toxicology, 2001. 157(1–2): p. 3-10.

DEUXIEME PARTIE : Modélisation des réactions métaboliques en utilisant les GCRs.

Introduction

Ce chapitre se concentre sur le traitement et la modélisation de propriétés liées aux réactions métaboliques diverses. Les réactions biochimiques mettent en jeu des échanges de matières et/ou d'énergie au sein des cellules d'un corps. L'ensemble de ces réactions forme un réseau de voies métaboliques caractérisant le devenir des molécules, appelées métabolites, au sein de l'organisme. La plupart des réactions du corps humain sont des réactions enzymatiques. Les enzymes permettent aux réactions chimiques de s'effectuer à plus grande vitesse et avec une spécificité limitant l'obtention de produits secondaires.

Dans les deux premières parties, les réactions enzymatiques générales se produisant au sein d'un organisme seront modélisées. Pour ce faire les réactions seront représentées sous la forme de graphes condensés de réactions. L'obtention de tels graphes nécessite dans un premier temps de trouver la correspondance entre les atomes dans les réactifs et ceux dans les produits. Cette procédure visant à assigner un même nombre aux mêmes atomes des deux côtés de la flèche réactionnelle est appelé « atom-to-atom mapping ». Cette procédure peut être automatisée à l'aide de logiciels mais ne délivre cependant pas un « mapping » satisfaisant dans l'intégralité des cas. Nous nous attarderons tout d'abord à identifier les cas de mapping « incorrects » d'un jeu de réactions métaboliques à l'aide d'un modèle QSAR.

Une fois les réactions métaboliques transformées en GCRs, une classification visant à les séparer en 3 différents grands groupes de réactions (oxydoréduction, hydrolyse et transfert de groupes) sera entreprise.

Enfin dans la 3^{e} partie les sites d'oxydation de différents substrats seront prédits chez l'homme pour l'isoenzyme CYP1A2 et 3A4 en utilisant les GCRs.

3. Détection de mapping atomique incorrect généré par un logiciel automatique en utilisant les GCRs.

La transformation de biotransformations en GCRs nécessite un mapping atomique parfait sous peine de se retrouver avec un graphe non représentatif des transformations qui ont eu lieu durant la réaction. Le « *atom-to-atom mapping »* est une procédure qui attribue un même numéro à un même atome du côté des réactifs et des produits. Malheureusement, aucun logiciel proposant d'effectuer ce mapping de façon automatique n'est fiable, surtout lorsqu'il s'agit de réactions biochimiques. Cela s'explique par le fait que le mapping est un problème dégénéré : plusieurs solutions sont optimales pour des réactifs donnés mais une seule dans des conditions réactionnelles précises. Le but de ce travail est de développer un modèle de classification des réactions avec un mapping correct/incorrect effectué avec le logiciel ChemAxon (l'un des plus réputé en chémoinformatique).

La modélisation a été réalisée à l'aide d'un jeu de données contenant 95 réactions biochimiques mal mappées que l'on a divisé en un jeu d'entraînement de 61 réactions et un jeu de test de 33 réactions. Pour chaque jeu les réactions mal mappées ont été manuellement corrigées et ajoutées au jeu correspondant. On a donc pour le jeu d'entraînement 61 réactions mal mappées et 61 réactions bien mappées, et pour le jeu de test 33 réactions mal mappées et 33 réactions bien mappées.

Les GCRs des réactions pour les 2 jeux ont été générés et les descripteurs fragmentaux ISIDA ont été calculés pour construire des modèles SVM et JRip. Un modèle très simple et facilement interprétable a pu être construit. Constitué de 6 règles correspondant à la présence ou à l'absence de motifs sous structuraux ce modèle a permis d'atteindre un rappel de 100% pour les réactions mal mappées et un rappel de 91% pour les réactions bien mappées sur le jeu de test. A ce jour cette approche de détection de mapping incorrect est unique et pourrait être employée pour détecter les cas de mapping incorrect pour n'importe quel algorithme de mapping.

Les résultats ainsi que les méthodes d'apprentissage utilisées sont décrits plus en détail dans l'article qui a récemment été publié dans le *Journal of Chemical Information and Modeling* :

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Models for Identification of Erroneous Atom-to-Atom Mapping of Reactions Performed by Automated Algorithms

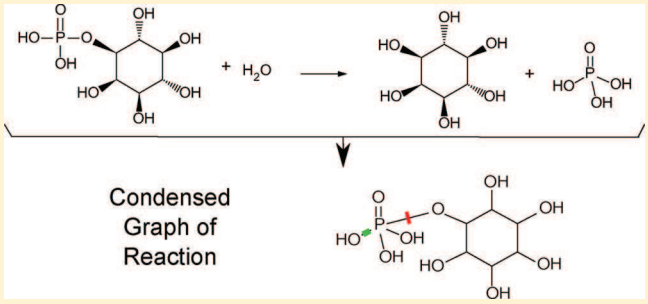
Christophe Muller,[†] Gilles Marcou,[†] Dragos Horvath,[†] João Aires-de-Sousa,[‡] and Alexandre Varnek*[,†]

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

[‡]CQFB, REQUIMTE, Departamento de Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Supporting Information

ABSTRACT: Machine learning (SVM and JRip rule learner) methods have been used in conjunction with the Condensed Graph of Reaction (CGR) approach to identify errors in the atom-to-atom mapping of chemical reactions produced by an automated mapping tool by ChemAxon. The modeling has been performed on the three first enzymatic classes of metabolic reactions from the KEGG database. Each reaction has been converted into a CGR representing a pseudomolecule with conventional (single, double, aromatic, etc.) bonds and dynamic bonds characterizing chemical transformations. The ChemAxon tool was used to automatically detect the matching atom pairs in reagents and products. These automated mappings were analyzed by the human expert and classified as "correct" or "wrong". ISIDA fragment descriptors generated for CGRs for both correct and wrong mappings were used as attributes in machine learning. The learned models have been validated in *n*-fold cross-validation on the training set followed by a challenge to detect correct and wrong mappings within an external test set of reactions, never used for learning. Results show that both SVM and JRip models detect most of the wrongly mapped reactions. We believe that this approach could be used to identify erroneous atom-to-atom mapping performed by any automated algorithm.



INTRODUCTION

A chemical reaction is the transformation of particular bonds of reactants resulting in formation of products. Identification of such bonds is possible if a correspondence of atoms in reactants and products (*atom-to-atom mapping*, AAM) is known. The automatized AAM procedures are implemented in most of the chemical database management systems like SYMYX, ChemAxon, and ACD/Labs. However, they do not always correctly identify related atoms in reactants and products. The point is that the mapping should account for the reaction mechanism, which is not easy to implement into the algorithm. Esters hydrolysis is a typical example: in order to perform AAM one should know to which product − acid or alcohol − belongs to the bridging oxygen atom. Since the 1980s, various methods of automated AAM were reported in the literature. Thus, Jochum et al.[1] suggested the AAM algorithm based on the empirical principle that "most chemical reactions proceed along a pathway of minimum chemical distance, i.e. with a redistribution of a minimum number of valence electrons". Unfortunately this simple strategy is not always adequate, because not all chemical reactions follow the minimal chemical distance principle or, in some cases, a combinatorial explosion of possible solutions may occur. Körner and Apostolakis' algorithm to map metabolic reactions[2] assumes that a correct mapping follows the lowest imaginary transition state energy (ITSE), which is computed as the sum of the weights of reacting bonds. The inconvenience of this method is that for some reactions it finds several alternative mappings or no correct solution. ChemAxon, a popular chemoinformatics software provider, uses an algorithm based on the search of isomorph subgraphs in reactant/product pairs.[3] It is not clear whether complementary information about reaction mechanisms is taken into account by it.

To sum up, nowadays there is no unique AAM algorithm which correctly maps all possible chemical reactions. Developing a new algorithm incorporating the knowledge on reaction mechanisms is an extremely challenging task. More realistically and readily feasible, as will be shown in this article, a first step toward more reliable automated AAM could be the learning of models able to distinguish correctly mapped from erroneously mapped reactions. Agreed, these may not tell what the correct mapping should have been, but they may significantly reduce the time invested by a human expert in charge of AAM, by focusing his or her attention on the relevant, problematic cases only.

The latter scenario is considered in this paper using as example a set of metabolic reactions from the KEGG database.[4] These data have been selected because any metabolic reaction can be unmistakably mapped thanks to the reactant pairs approach.[5] Here, for a selected set of metabolic reactions we compare a manual mapping with automated AAM performed with the

Received: September 1, 2012

   dx.doi.org/10.1021/ci300418q | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

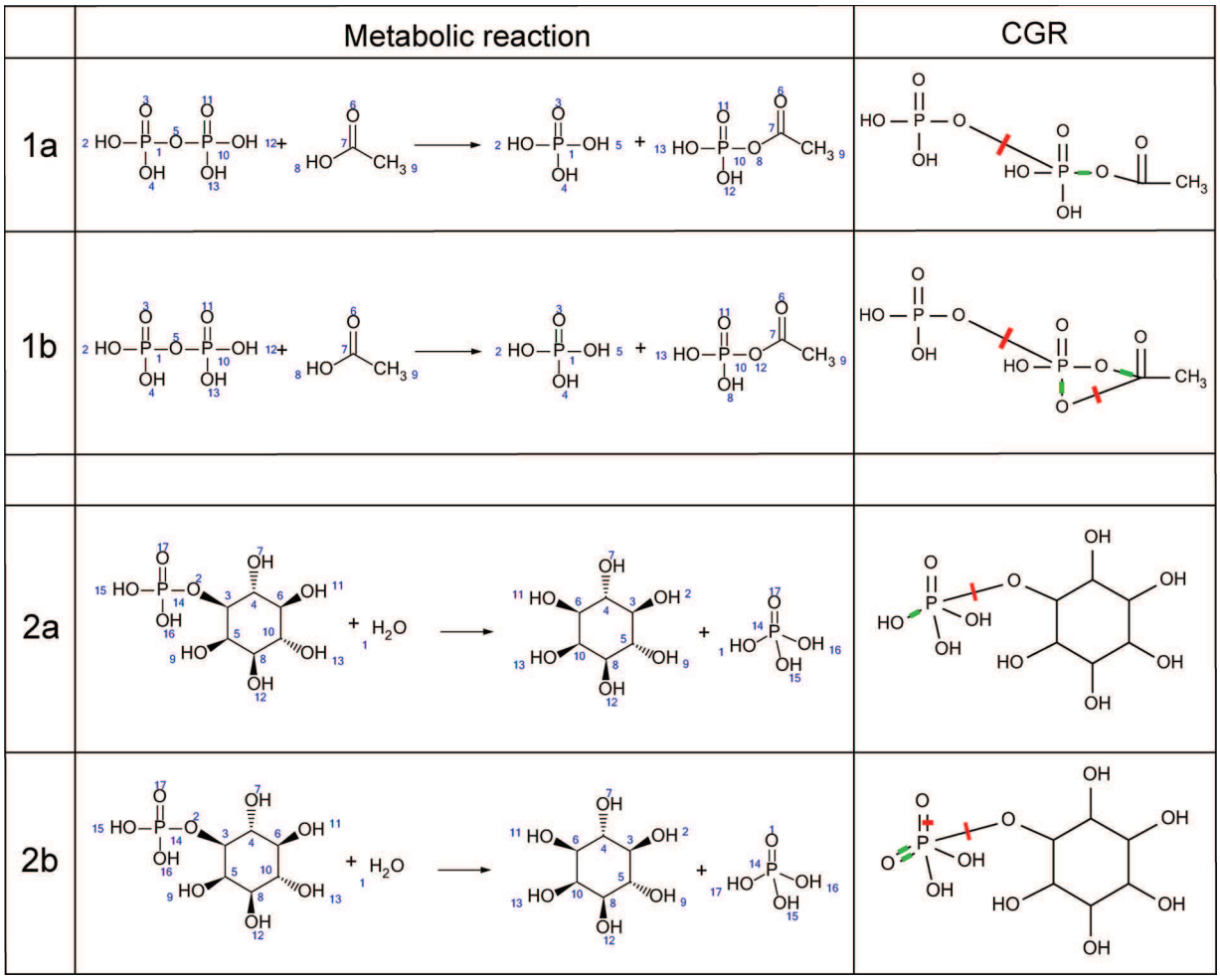


Figure 1. Examples of correct (a) and erroneous (b) atom-to-atom mapping leading to different Condensed Graphs of Reaction. CGR involves both conventional (single, double, ..., etc.) and dynamical bonds describing chemical transformations. Here, — and — correspond to broken and created single bonds, respectively. One can see that CGRs corresponding to correct AAM contain less dynamical bonds than CGRs corresponding to wrong AAM.

ChemAxon Standardizer software (release 2009).[6] Reactions for which automated and manual mapping differ represent the "wrong" AAM class of a modeling data set, serving to learn classification models.

Typically in structure−property modeling, machine-learning methods are applied to individual molecules represented by descriptor vectors. Since any chemical reaction involves several molecular species (at least, two), at the first step they were transformed into Condensed Graphs of Reaction (CGR) − some sort of pseudomolecules condensing information about all reactants and products[7] (Figure 1). Atoms' connectivity in CGR depends on atom−atom mapping, thus different mappings produce different graphs. At the second step, ISIDA fragment descriptors[8,9] have been generated for CGRs, followed by their use in model building (Figure 2). Finally, SVM[10] and JRip[11] models distinguishing correct and wrong AAM were obtained and validated.

METHOD

Condensed Graphs of Reaction. The "Condensed Graph of Reaction" (CGR) is a pseudomolecule which results from the superposition of reactant(s) and product(s) molecular graphs into one single graph.[12] The chemical transformations are explicitly taken into account through dynamic bonds. The latter represent covalent bonds being broken, formed, or transformed during the reaction (Figure 1). Atom-to-atom mapping must be performed in order to identify superposed atoms in reactants and products. AAM is a crucial stage of CGR construction: different mappings lead to different graphs (Figure 1).

Data Processing. 850 metabolic reactions from the three first enzymatic classes - EC 1 (oxidoreductases), EC 2 (transferases), and EC 3 (hydrolases) have been retrieved from the KEGG LIGAND Database.[4] All reactions were standardized with ChemAxon Standardizer (release 2009) tool to remove explicit hydrogens and stereochemistry. Stoichiometry of all reactions has been respected. Moreover, a reactant needs to be duplicated in the reaction if it enters into the product twice. Certain reactions have been discarded either by ChemAxon or manually. This concerns the following:

1. Too complex reactions for which ChemAxon Standardizer gives a warning "Reaction is too complex to be automapped". In case of metabolic reactions, it may happen if a sum of atoms in the reagents and products is larger than 130.
2. Reactions having an unbalanced number of atoms or an unbalanced total charge.
3. Reactions involving metals.
4. The duplicates have been discarded.

dx.doi.org/10.1021/ci300418q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

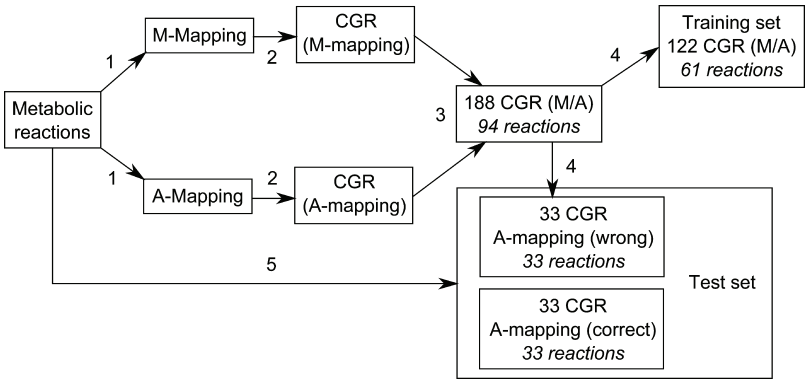


Figure 2. Dataflow. 1. Selected set of metabolic reactions is mapped both manually (M-mapping) or automatically (A-mapping). 2. Two sets of Condensed Graphs of Reactions corresponding to different mapping procedures are generated. 3. Their intersection resulted in a limited number of reactions for which A- and M-mapping differ. The resulting set of CGR (M/A) forms a modeling set which is then split into training and external test sets. 4. 122 CGRs (M/A) are moved to the training set and remaining CGRs of automatic erroneous mapping are added to test set. 5. CGRs of automatic correctly mapped reactions from initial set are generated and added to test set.

5. Reactions involving only displacement of hydrogen atoms. These transformations are not visible on hydrogen suppressed graphs.

6. Reactions containing highly symmetrical reactants or products, because the exact positioning of reaction center could not be deduced by the reaction pairs approach (see below).

7. Reactions in which AAM algorithm assigns to products' atoms numbers which were not assigned to reactants' atoms.

The remaining 630 reactions were both automatically mapped with ChemAxon Standardizer and manually mapped using reactant pairs approach (Figure 2). Manual mapping of metabolic reactions has been performed using the KEGG RPAIR database, which allows users to perform an alignment of reactant pairs. The latter is defined by Kotera et al.[13] " as pairs of compounds that have atoms or atom groups in common on two sides on a reaction". In most cases, the information about reactant pairs is sufficient to perform an unambiguous manual AAM. All manually mapped reactions were considered as correct AAM. If ChemAxon's mapping differed from the manual one, it was considered as erroneous.

All reactions (RD file) have been transformed into CGR using the ISIDA-CGR Designer program and saved as an SD file. Thus, 2 subsets of CGRs corresponding to ChemAxon and manual AAM, respectively, have been generated (Figure 2). These subsets were merged into one SD file in which the duplicates, coming from correct ChemAxon mapping perfectly matching manual AAM, were identified with the EdiSDF program[14] freely available at http://infochim.u-strasbg.fr/.

Modeling Data Sets. Comparison of manual and automated AAM has shown that 94 out of 630 reactions were wrongly mapped by ChemAxon Standardizer, which represents 15% of the initial data set. These reactions and their correctly mapped counterparts have been transformed in 188 Condensed Graphs of Reaction which formed a modeling set. Thus, for each metabolic reaction, 2 CGRs corresponding to correct and erroneous AAM were generated. All reactions forming the modeling set have been split into training and test sets in proportion 2:1. The training set contained 122 CGRs issued from 61 reactions, including 4 reactions of the class EC 1, 29 of the class EC 2, and 28 of the class EC 3. An external test set was composed of 66 CGRs issued from 66 reactions automatically mapped by ChemAxon, namely 33 out of 94 wrongly mapped reactions, and 33 reactions initially correctly mapped. This included 3 reactions of the class EC 1, 32 of the class EC 2, and 31 of the class EC 3. Thus, there is no particular reaction type which represents a problem for ChemAxon AAM: wrongly mapped reactions were detected for all studied classes.

Descriptors. ISIDA fragment descriptors[8,9] were used for the modeling. Each fragment represents a subgraph of a CGR, whereas its occurrence is a descriptor value. Molecules were represented with implicit hydrogen atoms. For each class, four subtypes are defined AB, A, B, and AP. Sequences represent the shortest paths between two selected atoms and may include both atoms and bonds (AB), atoms only (A), and bonds only (B). For the Atoms Pairs (AP) subtype, only terminal atoms and the topological distance between them are represented explicitly. An "extended augmented atom" represents a selected atom with its environment including sequences of AP, AB, A, and B types issued from this atom. Only fragments having at least one dynamical bound were considered. In extended augmented atoms, the branches not containing dynamic bonds were omitted.

For every subclass, the minimal (n_{min}) and maximal (n_{max}) numbers of atoms varied from 2 to 10 for the sequences and from 2 to 6 atoms for the augmented atoms. For any combination of n_{min} and n_{max} all intermediate shortest paths with n atoms ($n_{min} \leq n \leq n_{max}$) are also generated (Figure 3). Varying n_{min} and n_{max}, as well as different subclasses, 240 descriptors' pools were generated with the ISIDA Fragmentor program.[15] Fragmentation types leading to the most performing models have been identified in a cross-validation procedure.

Obtaining and Validation of Models. Two machine learning methods - Support Vector Machine[10] and a propositional rule learner JRip - have been used to build and validate the models distinguishing correct and incorrect AAM considered as two different classes. SVM modeling was performed with the LibSVM[16] software. Tanimoto similarity coefficient was used as kernel. The cost parameter was optimized to achieve the best class separation.

The JRip method realizes the RIPPER algorithm[11] (Repeated Incremental Pruning to Produce Error Reduction) implemented in the Weka software.[17] In RIPPER, the training set is randomly split into growing and pruning sets. An initial rule set is generated on a growing set followed by its further simplifications by applying pruning operation. The pruning stops when simplification yields to an increase of the error on the pruning set. Default parameters from Weka were kept for building models.

Balanced accuracy (BA) was used to assess the performance of the models. BA is calculated as a function of true positive (TP),

dx.doi.org/10.1021/ci300418q | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

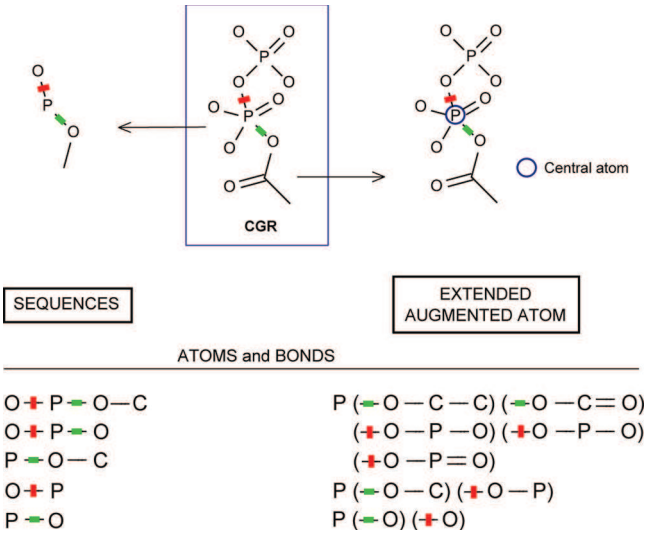


Figure 3. Different types of ISIDA fragment descriptor used in this study: only sequences containing at least one dynamic bond (red if broken, green if created) are considered. This example shows atoms/bonds sequences (subclass AB) of length from 2 to 4 atoms are generated and extended augmented atoms of length from 2 to 3 atoms. Fragments of subclasses A (atoms only) and B (bonds only) could be derived from AB fragments by omitting symbols of atoms or bonds, respectively.[7]

false positive (FP), true negative (TN), and false negative (FN) examples retrieved with the model:

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Selection of Optimal Types of Fragment Descriptors and Model Validation. The goal of this stage was to select among different initial descriptor pools the one providing the highest predictive performance of the models. The workflow is shown in Figure 4. At the first step, few fragmentation types performing the best in 5 × 5-folds cross-validation (5 × 5 CV) have been selected according to BA. They have been tested additionally in 20 × 5 CV. For each calculation, the average value <BA> and the standard deviation Δ(BA) have been assessed. Smaller Δ(BA) corresponds to more stable models. Finally, *Y*-scrambling has been performed in order to check a chance correlation problem. At this step, 5-fold cross-validation has been performed on a data set with the reshuffled labels "correct/erroneous AAM". This procedure has been repeated 20 times. For robust models, BA (scrambling) is expected to be much lower than <BA>.

Selected fragmentations have been used to build the models on the entire training set followed by their validation on the external test set. Applicability domain has not been used because all reactions belong to the three first enzymatic classes.

RESULTS AND DISCUSSION

In SVM calculations, sequences of bonds of length 2 to 3 atoms perform better than other fragments. In 20 × 5 CV, the models built on these descriptors result in <BA> = 88.5% and Δ(BA) = 1.1. In scrambling calculations, BA varied from 38.7 to 58.1% which is significantly smaller compared to <BA>. In JRip modeling, the optimal fragment descriptors were sequences of bonds of length 2 to 8 atoms. In 20 × 5 CV, it performs similarly to SVM: <BA> = 88.7%, Δ(BA) = 3.4, and BA (scrambling) = 40.4 to 56.5%.

Finally, the models for correct and wrong AAM has been obtained on the entire training set using optimal descriptor pools for SVM and JRip. Both machine-learning methods perform well on the external test set achieving BA = 0.94. Notice that SVM retrieves correctly mapped reactions slightly better than JRip, whereas the opposite trend is found for the retrieval of wrongly mapped reactions where JRip model performs slightly better (Table 1).

Although SVM and JRip models have similar predictive performances, the latter approach has some additional benefits.

Table 1. Number of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) Examples Retrieved with SVM and JRip Models from the External Test Set[a]

	TP	FP	TN	FN
SVM	31	2	31	2
JRip	30	1	32	3

[a]Here, reactions with correct and erroneous AAM correspond, respectively, to positive and negative examples.

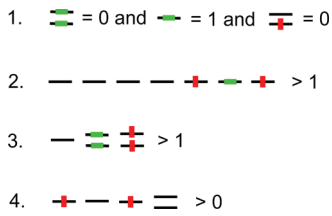


Figure 5. Rules generated by JRip to filter correctly mapped metabolic reactions. In selected structural patterns, only bonds are represented explicitly.

Table 2. True Positives (TP) and False Positives (FP) Retrieved by Particular JRip Rules on the Training and Test Sets

	training set		test set	
rule #	FP	TP	FP	TP
1	1	45	0	28
2	0	5	0	0
3	1	4	1	2
4	0	2	0	0

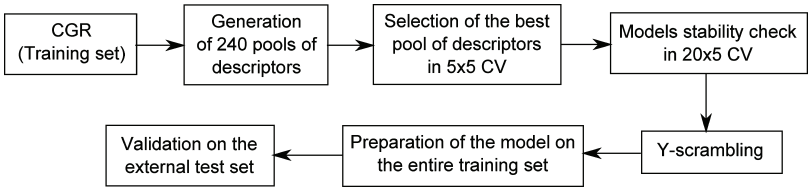


Figure 4. Modeling workflow used in this study.

dx.doi.org/10.1021/ci300418q | *J. Chem. Inf. Model.* XXXX, XXX, XXX–XXX

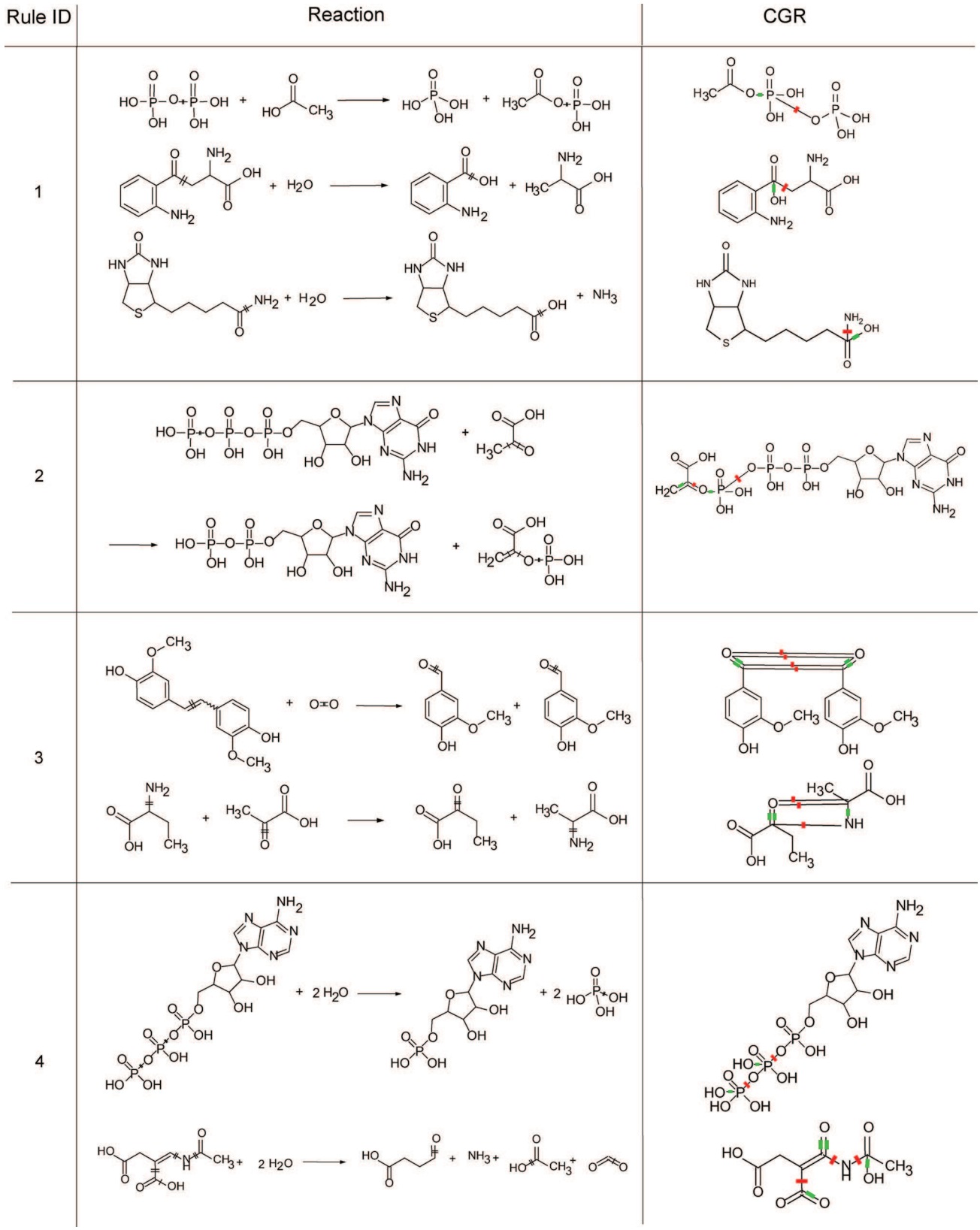


Figure 6. Examples of reactions retrieve by each rule generated by JRip to filter correctly mapped metabolic reactions. Crossed bonds in reactants and products structures represent the ChemAxon annotation of chemical transformations.

Generally, JRip models could be easily interpreted. In this study, it involves four simple rules based on occurrences of particular structural patterns (Figure 5). For example, the first rule states that correctly mapped reaction *must not* involve formation of a double bond AND transformation of double bond to single bond AND *must* involve formation of one single bond. Notice that this rule allows one to retrieve more than 74% and 84% of correctly mapped reactions in the training and test set, respectively (Table 2).

dx.doi.org/10.1021/ci300418q | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

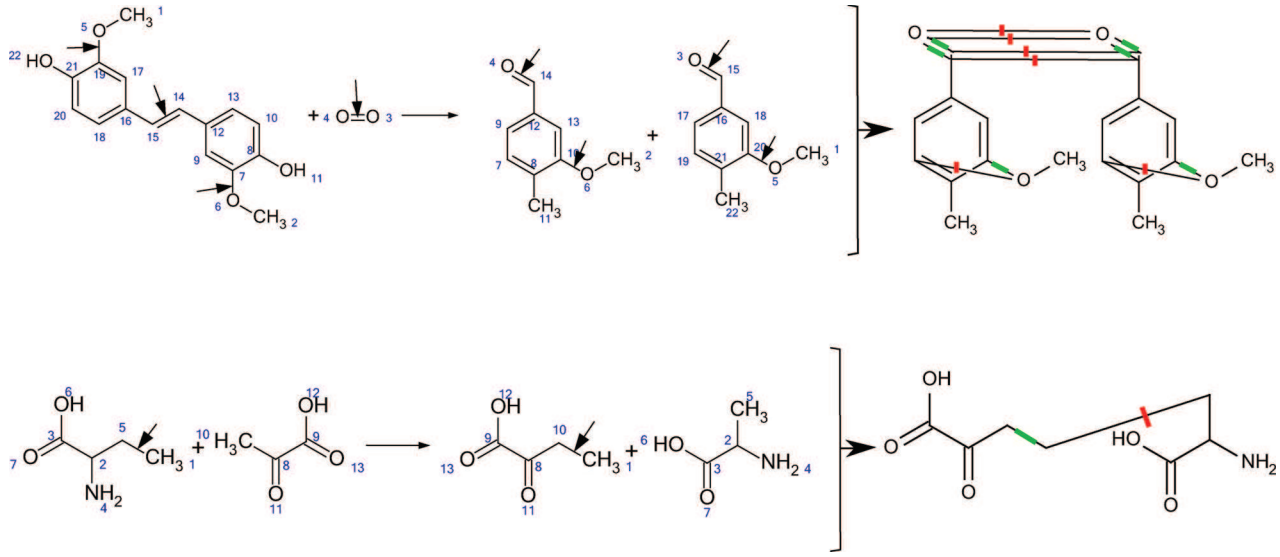


Figure 7. Reactions from training set erroneously classified by JRip model as correctly mapped ones. Black arrows point to the reacting centers.

Examples of reactions retrieve by the JRip rules are given in Figure 6. Rule 1 retrieves both transferase and hydrolase reactions. Rule 2 retrieves exclusively transferase reactions of transfer of phosphate groups using alcohol or a carboxyl groups as acceptor. Rule 3 concerns oxidoreductases and transferases reactions. Finally, the last rule retrieves only hydrolases reactions.

Another advantage of JRip models concerns the possibility of completing them by manually designed rules. In Table 2, two False Positives in the training set correspond to wrongly mapped reactions given in Figure 7. There are very few similar reactions in the training set, and therefore the model has not captured their structural patterns. Visual inspection of CGRs corresponding to the above reactions suggests two additional rules C − CH_2 + C > 0 and C − O + C > 0 which improve the model's performance on the external test sets (BA = 0.95).

CONCLUSION

Automated atom-to atom mapping of chemical reactions represents a difficult task because in some cases the computer algorithm based on the graph theory is not sufficient and a reaction mechanism must be taken into account. Nowadays, there exists no automated algorithm perfectly performing AAM. In this paper, we demonstrated that statistical models built on fragment descriptors issued from Condensed Graphs of Reactions represent an efficient way to detect erroneous AAM made by automated algorithms. Although, any machine-learning methods can be used for the modeling, those deriving associative rules (e.g., JRip) have some preference because automatically derived rules can be completed by manually derived custom ones.

ASSOCIATED CONTENT

Supporting Information

630 enzymatic reactions used in the modeling. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Jochum, C.; Gasteiger, J.; Ugi, I. The principle of minimal chemical distance. *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495−505.

(2) Korner, R.; Apostolakis, J. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J. Chem. Inf. Model.* **2008**, *48* (6), 1181−1189.

(3) *AutoMapper*, version 5.1.1; ChemAxon: Budapest, Hungary, 2009.

(4) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40* (D1), D109−D114.

(5) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003**, *125* (39), 11853−11865.

(6) *Standardizer*, version 5.1.1; ChemAxon: Budapest, Hungary, 2009.

(7) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19* (9−10), 693−703.

(8) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 191−198.

(9) Horvath, D.; Bonachera, F.; Solovev, V.; Gaudin, C.; Varnek, A. Stochastic versus stepwise strategies for quantitative structure-activity relationship generation − how much effort may the mining for successful QSAR models take? *J. Chem. Inf. Model.* **2007**, *47* (3), 927−939.

(10) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cristianini, N., Shawe-Taylor, J., Eds.; Cambridge University Press: Cambridge, United Kingdom, 2000.

(11) Cohen, W. W. Fast Effective Rule Induction. In *Machine learning: proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, July 9−12, 1995*; Prieditis, A., Eds.; The Morgan Kaufmann series in machine learning; Morgan Kaufmann Publishers: Burlington, 1995; pp 115−123.

(12) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. *Int. J. Artif. Intell. Tools* **2011**, *20* (2), 253−270.

(13) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **2004**, *126* (50), 16487−16498.

(14) de Luca, A.; Horvath, D.; Marcou, G.; Solov'ev, V.; Varnek, A. Mining chemical reactions using neighborhood behavior and condensed

graphs of reactions approaches. *J. Chem. Inf. Model.* **2012**, *52* (9), 2325−2338.
(15) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA property-labelled fragment descriptors. *Mol. Inf.* **2010**, *29* (12), 855−868.
(16) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intelligent Systems Technol.* **2011**, *2* (3), 27:1−27:27.
(17) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explorations* **2009**, *11* (1), 10−18.

4. Classification des réactions enzymatiques de la KEGG.

4.1. Introduction

Au sein d'une cellule du corps humain, de multiples molécules interagissent perpétuellement aux cours de réactions biochimiques. Ces réactions peuvent être considérées individuellement mais la complexité des systèmes biologiques nous laisse appréhender des groupes de réactions. Chacune des réactions formant un groupe concoure à un même objectif représenté par un chemin ou une voie métabolique. Un chemin métabolique est défini par une succession de réactions enzymatiques ou non, dans lequel, chaque réaction peut-être catalysée par plusieurs de ces enzymes.

Des méthodes usuelles de Chémoinformatique peuvent être appliquées afin de comparer et d'effectuer des classifications automatiques de ces réactions enzymatiques pour :

- valider les systèmes de classification,
- comparer ou reconstruire des chemins métaboliques,
- classifier les mécanismes enzymatiques.

Dans le cas présent, la comparaison de ces réactions repose sur le nombre EC (Enzyme Commission). Ce nombre EC est souvent employé simultanément comme un identifiant de réactions, d'enzymes et gènes d'enzyme. Cependant l'utilisation de ce nombre possède des limitations [1]. En effet on note des problèmes concernant les réactions réversibles [2], lorsqu'une enzyme catalyse plus d'une réaction, ou lorsqu'une même réaction est catalysée par plusieurs enzymes [3, 4].

Parmi les méthodes qui tentent de déterminer le nombre EC associé à une enzyme ou une réaction, ou de comparer celui-ci à la classification proposée, on compte les méthodes qui se basent sur la similarité des interactions ligand-protéine pour déterminer la fonction d'une nouvelle enzyme [5], celles qui comparent la structure des protéines [6], ou encore celles qui ne se basent que sur la comparaison de la structure des réactifs et produits des réactions métaboliques et par extension

des mécanismes réactionnels [4, 7-11]. Cette dernière approche nous intéressant plus particulièrement, nous allons détailler un peu plus certaines de ces méthodes.

Récemment, une étude [8] utilisant les cartes de Kohonen a permis de classer des réactions de la base de données KEGG et ainsi de retrouver le premier chiffre du nombre EC de chacune de ces réactions. Dans cette approche 7 descripteurs empiriques physico-chimiques sont calculés pour chaque liaison de chaque molécule participant à la réaction. Chaque réaction est représentée numériquement, sans assigner de centre réactionnel, en calculant la différence entre les descripteurs moléculaires des produits et ceux des réactifs représentés sous la forme de fingerprint appelé Molmap [9]. Cette soustraction permet de ne garder que les informations sur les liaisons transformées durant la réaction. Une carte de Kohonen utilisant les fingerprints de réactions est finalement construite.

Dans une autre étude Gasteiger et al. [10] proposent de classer automatiquement les réactions enzymatiques de type hydrolase. Pour ce faire ils définissent le(s) centre(s) réactionnel(s) de chaque réaction, et calculent 6 propriétés physico-chimiques empiriques pour chaque liaison rompue. Chaque réaction est ensuite projetée dans un espace à deux dimensions à l'aide dune carte de Kohonen.

J.-L. Faulon et al. [11] proposent aussi une méthode où l'identification des centres réactionnels n'est pas nécessaire. Ils calculent la signature réactionnelle pour chaque réaction métabolique. Cela revient à générer pour chaque atome des réactifs et des produits, les fragments d'atomes et liaisons traduisant l'environnement proche de l'atome considéré (comparable aux atomes unis augmentés dans ISIDA). Les descripteurs des produits sont alors soustraits aux descripteurs des réactifs pour obtenir la signature réactionnelle. Des machines d'apprentissage automatique peuvent alors être utilisées pour comparer les différentes réactions métaboliques entre elles.

Le but de cette partie est de démontrer de façon succincte la capacité des graphes condensés de réactions à représenter les réactions métaboliques et ainsi permettre la classification de différents types de réactions à l'aide d'une carte de Kohonen. Pour ce faire nous comparerons les résultats de notre classification à celle du 1er chiffre du nombre EC seulement, celui-ci étant caractéristique du type de réaction catalysée.

4.2. Matériel et méthodes

4.2.1. Jeu de données

Les réactions de la KEGG appartenant aux 3 premières classes enzymatiques ont été extraites automatiquement. Seules ces 3 classes ont été choisies car ce sont ces types de réactions qui semblent le mieux se séparer à l'aide d'une SOM d'après la publication de João Aires-de-Sousa [8]. L'extraction des réactions à été réalisée grâce à un script utilisant le navigateur lynx. Les fichiers *.mol* relatifs aux réactions ont été rapatriés et mis en ordre dans un fichier *.rdf* afin de respecter la stœchiométrie des réactions. Par soucis de temps les réactions ayant des longueurs de chaînes variables (à ajuster manuellement) ont été écartées ainsi que les réactions pour lesquelles la charge totale des réactifs était différente de la charge totale des produits. Les réactions ont ensuite été automatiquement mappées par le logiciel Standardizer de ChemAxon. Le mapping a été manuellement vérifié puis corrigé et les réactions ont été transformés en GCRs. Compte tenu de l'ampleur du travail que représente la vérification et la correction du mapping, seules 627 réactions ont été conservées pour constituer le jeu de données. Parmi celles-ci on compte :

- 242 réactions de type oxydoréductase EC = 1.X.X.X
- 243 réactions de type transférase EC = 2.X.X.X
- 142 réactions de type hydrolase EC = 3.X.X.X

Pour ces réactions, les hydrogènes n'ont pas été pris en compte étant donné que le programme de mapping refuse de mapper les réactions contenant trop d'atomes. De plus il n'est pas toujours évident de localiser un même hydrogène des deux côtés de la flèche réactionnelle.

4.2.2. Descripteurs, méthodes d'apprentissages et critères d'évaluation des modèles.

Pour chaque GCR les descripteurs calculés sont les descripteurs ISIDA SMF. Seuls les fragments contenant au minimum une liaison dynamique ont été générés.

Au total 180 fragmentations ont été générées, cela comprend les séquences d'atomes (IA), les séquences de liaisons (IB), les séquences d'atomes et liaisons (IAB) ainsi que les atomes unis (IIIA, IIIB, IIIAB). La longueur minimale et maximale des fragments générés a été variée de 2 à 10 pour les séquences, et de 2 à 6 pour les atomes unis.

L'algorithme de clustering K-Means a été utilisé afin de générer des clusterings rapidement. 33 clusterings, allant de 3 à 99 clusters par pas de 3, ont été effectués pour chaque fragmentation. La métrique utilisée pour calculer la similarité entre les vecteurs de descripteurs est la métrique euclidienne. Les autres paramètres ont été pris par défaut dans le logiciel Weka. Une carte de Kohonen a aussi été construite à l'aide du logiciel SOM_PAK. La dimension de la carte est choisie de sorte à ce que le nombre de neurones final soit proche du nombre de clusters du modèle K-Means le plus performant associé à ce jeu de descripteurs. La topologie de la carte est un toroïde rectangulaire. Les autres paramètres ont été laissés par défaut.

Les performances des clusterings ont été évaluées en calculant la pureté des différents modèles.

4.3. Résultats

4.3.1. K-Means

En tout 5940 clusterings ont été réalisés. Les modèles ont été classés selon la pureté des clusters obtenus. Il faut retenir que seuls les fragments contenant au moins une liaison dynamique ont été générés, c'est-à-dire seul le cœur de réaction ainsi que son environnement proche sera décrit. La performance du meilleur modèle obtenu pour chaque type de fragmentation est donné Tableau 4-1 et Figure 4-1. Ces meilleurs modèles sont généralement obtenus lorsque 99 clusters sont utilisés.

Tableau 4-1. Pureté du meilleur modèle de clustering obtenu avec K-Means pour chaque type de fragmentation.

fragmentation	IA(2-3)	IB(2-5)	IAB(2-4)	IIIA(2-3)	IIIB(3-3)	IIIAB(2-2)
pureté	0,946	0,963	0,931	0,917	0,901	0,903

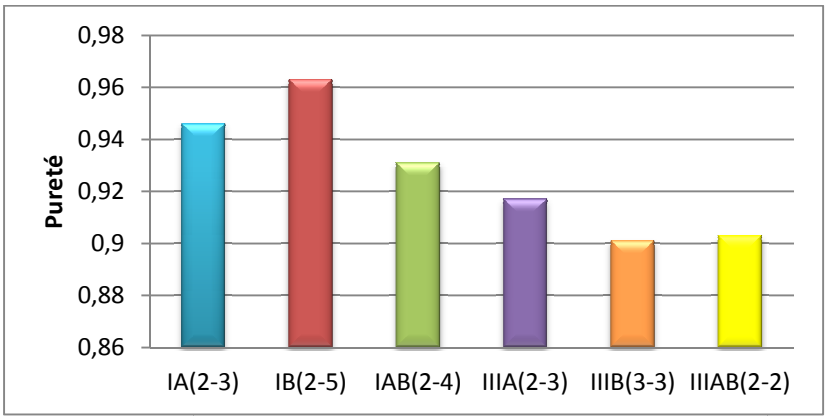


Figure 4-1. Graphique représentant la pureté du meilleur clustering obtenu pour chaque type de fragmentation.

En considérant uniquement le meilleur modèle obtenu pour chaque type de fragmentation on s'aperçoit que le meilleur clustering est obtenu pour les fragments représentant uniquement des séquences de liaisons (pureté=0,963). Néanmoins les autres types de descripteurs permettent eux aussi d'obtenir des clusterings aux performances proches de celle du meilleur modèle obtenu. Il semblerait que l'emploi de séquences permet d'atteindre des performances plus élevées que l'emploi d'atomes unis. De plus l'emploi de liaisons uniquement permet d'obtenir le meilleur modèle pour les séquences, mais aussi le moins bon dans le cas des atomes unis.

Un graphique représentant la pureté moyenne des 15 meilleurs modèles (correspondant à 15 fragmentations de longueurs différentes mais de mêmes type et sous-type) a été tracé pour déterminer si un sous-type de descripteurs fragmentaux est plus approprié qu'un autre (voir Figure 4-2). Les valeurs exactes des puretés moyennes sont données Tableau 4-2. Les séquences de liaisons uniquement possèdent encore le meilleur score, mais cette fois-ci l'écart avec les autres fragmentations est plus important. Les séquences permettent toujours d'obtenir de meilleurs modèles que les atomes unis. Cependant on constate que l'emploi de liaisons uniquement permet aussi d'obtenir le meilleur score parmi les fragmentations de type atomes unis.

Tableau 4-2. Pureté moyenne des 15 meilleurs modèles de clustering obtenu avec K-Means pour chaque type de fragmentation.

fragmentation	IA	IB	IAB	IIIA	IIIB	IIIAB
pureté moyenne	0,889	0,941	0,865	0,834	0,855	0,831

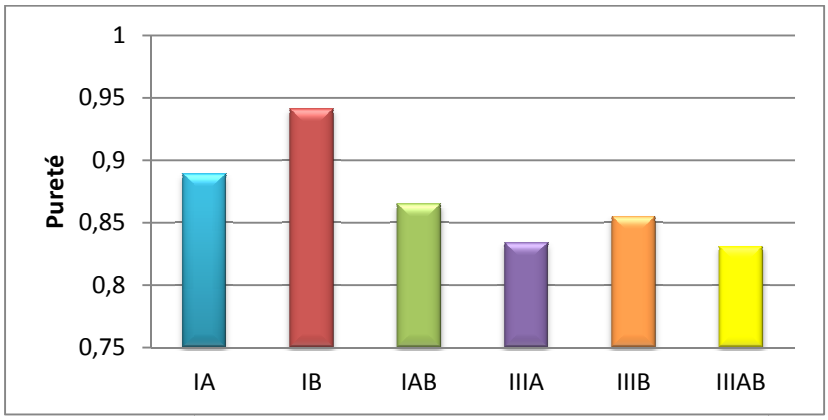


Figure 4-2. Graphique représentant la pureté moyenne des 15 meilleurs clusterings pour chaque sous-type de fragmentation.

Il n'est pas surprenant que l'usage de liaison uniquement entraine les meilleures performances. Les réactions enzymatiques d'une même classe peuvent entrainer les mêmes transformations de liaisons mais les atomes qui définissent ces liaisons peuvent être différents. Un exemple est proposé sur la Figure 4-3. Les 2 réactions sont catalysées par une enzyme de type hydrolase. Dans les deux cas la liaison reliant le groupe oxo avec l'atome d'intérêt (N ou O) casse et l'oxygène de l'eau vient s'attacher à la place pour former un groupe carboxyle. Les GCRs de ces réactions permettent de mieux visualiser les liaisons transformées (voir Figure 4-4).

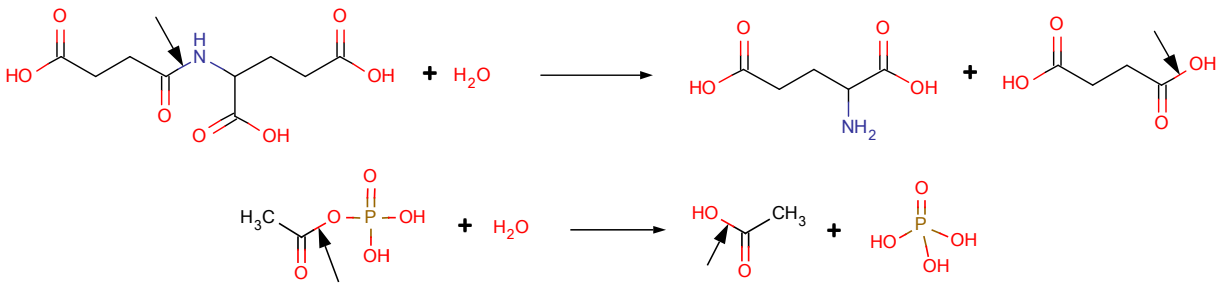


Figure 4-3. Réactions d'hydrolases avec de haut en bas les réactions R00411 et R00317. Les flèches noires pointent vers les liaisons cassées ou formées.

En générant uniquement les séquences de liaisons, et pour un environnement proche du centre réactionnel, les descripteurs entre ces 2 réactions seront identiques. A l'inverse, en incluant les atomes dans les descripteurs, il est évident que les vecteurs de descripteurs seront différents entre les deux réactions pour les

plus petits fragments possibles générés (c.à.d. contenant 2 atomes). Pour R0411 on verra par exemple apparaitre le fragment CN absent de R00317.

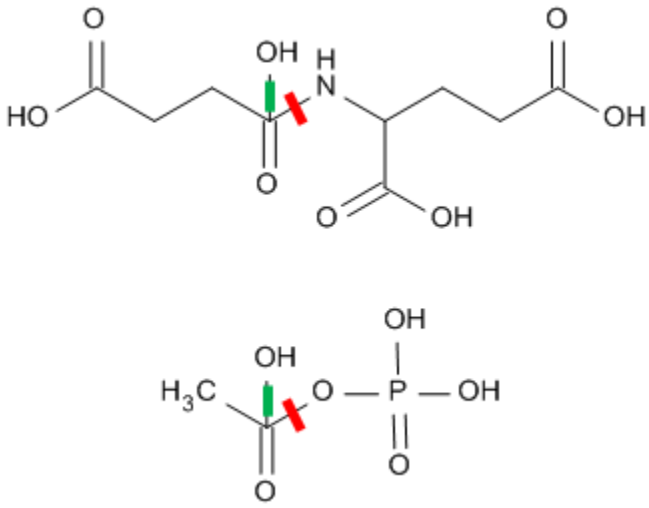


Figure 4-4. Graphes condensés des réactions R00411 et R00317 proposées Figure 4-3.

Les séquences d'atomes engendrent logiquement de moins bons clusterings car l'information sur les transformations opérées sur les liaisons n'est plus disponible. Et donc il est possible d'imaginer que des environnements d'atomes similaires mais possédant des liaisons différentes seront affectés par différentes enzymes bien que les descripteurs basés sur les séquences d'atomes sont similaires.

Les séquences d'atomes et liaisons engendrent de moins bons clusterings que les séquences d'atomes uniquement malgré la présence d'information sur les liaisons. Ceci peut s'expliquer par le fait que l'information codée par les descripteurs est trop spécifique dans notre situation. En effet, 2 réactions seront considérées comme semblables si les mêmes transformations ont lieu, mais aussi si l'environnement proche des centres réactionnels en termes d'atomes est similaire. Lorsque que 2 réactions seront considérées comme similaires par ce type de descripteurs elles appartiendront à la même classe de réactions enzymatiques mais elles possèderont aussi généralement des substrats similaires. Ce type de descripteurs peut alors être utilisé non pas uniquement pour prédire le 1er chiffre du nombre E.C. mais pour prédire les chiffres suivants propres aux substrats impliqués dans la classe réactionnelle.

Etant donné que nous nous attachons ici à séparer les réactions en classes enzymatiques, l'utilisation de séquences de liaisons seule est préférée. Les

performances de ce type de fragmentation sont présentées Figure 4-5 pour chaque combinaison de longueurs de fragments allant de 2 à 10 atomes.

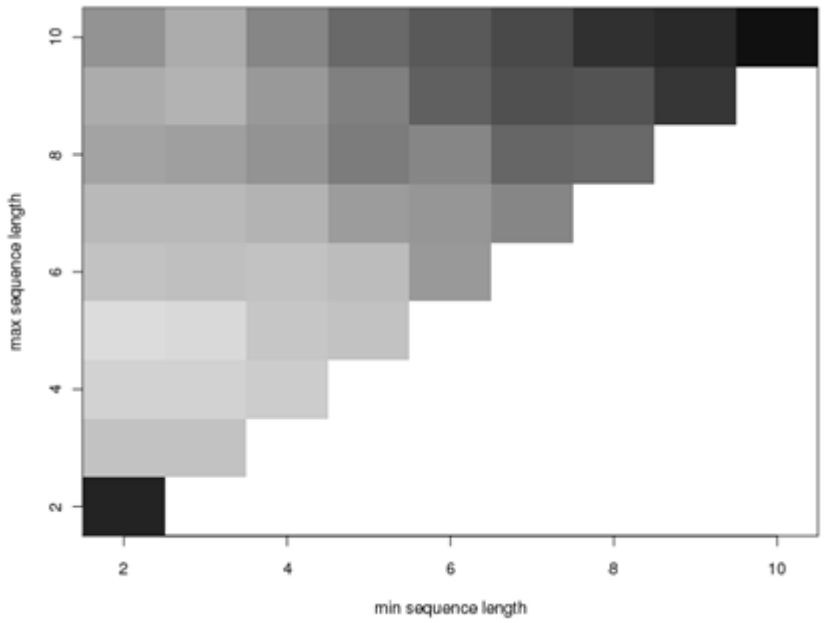


Figure 4-5. Pixel map de la pureté.
La pixel map représente la pureté du meilleur clustering trouvé pour chaque longueur de fragment pour les séquences de liaisons. Une case sombre correspond à un mauvais score alors qu'une case claire correspond à un bon score.

Les performances semblent inégales selon la taille des fragments considérés. Néanmoins on observe une certaine continuité dans le fait que les fragments amenant de bonnes performances tendent lentement vers des descripteurs moins efficace lorsque la longueur des fragments varie, et non de façon spontanée.

On s'aperçoit que les performances les moins élevées se situent pour les fragments très longs (8 à 10 atomes) d'une part, et les fragments extrêmement courts (2 atomes) d'autre part. Pour les fragments très longs l'explication est simple, les descripteurs sont beaucoup trop spécifiques. La conséquence de l'utilisation de tels descripteurs sera l'apparition de descripteurs uniques à une réaction possédant une faible chance d'apparition dans d'autres GCRs. La similarité entre les GCRs devient alors très faible et presque tous sont considérés comme dissimilaires. D'autant plus que les liaisons se situant à plus de 8 liaisons du centre réactionnel ne sont généralement pas réactives et n'influent pas sur le centre réactionnel. Pour les fragments de longueur 2 c'est le problème inverse qui intervient. Les fragments ne

sont plus assez spécifiques et leur utilisation correspond juste à un décompte des liaisons dynamiques dans le GCR.

Les performances les plus hautes correspondent aux mélanges de fragments de longueurs courtes et moyennes (2 à 5 atomes). C'est-à-dire les descripteurs décrivant l'environnement proche des liaisons transformées sans pour autant être trop spécifique. Ce clustering « gagnant » comporte 99 clusters.

4.3.2. SOM

Une carte de Kohonen a alors été construite en utilisant les séquences de liaisons uniquement de longueur allant de 2 à 5 atomes (1 à 4 liaisons). La taille de la carte a été choisie de manière à avoir un nombre de clusters proche de celui du meilleur modèle K-Means. Ainsi une carte de taille 10x10 a été construite. La carte est représentée sur la Figure 4-6.

A première vue il semble y avoir une séparation nette des différentes classes enzymatiques. Sauf exception les réactions d'une même classe se regroupent dans un même espace de la carte plus ou moins grand. Cette séparation se retrouve accentuée par la présence de clusters vides entre celles-ci. On retrouve néanmoins des clusters « poubelles » où des réactions de classes différentes sont présentes.

En regardant plus en détail on s'aperçoit que les réactions d'oxydoréductases ne sortent pas dans le même cluster selon qu'elles oxydent ou réduisent les substrats. Cela n'est pas surprenant étant donné que les descripteurs ISIDA SMF comportant des liaisons dynamiques seront différents selon le sens de la réaction. Prenons l'exemple des réactions R01093 (neurone 3J) et R01758 (neurone 4D) toutes les deux catalysées par la même enzyme (voir Figure 4-7). Dans le premier cas (R01093) on a une oxydation. Le fragment décrivant la réaction comportera une liaison dynamique de type simple liaison cassée, alors que dans le cas de la réduction R01758 le fragment décrivant la réaction comportera une liaison dynamique de type simple liaison créée. Cette simple différence permet de classer ces deux réactions dans des clusters différents.

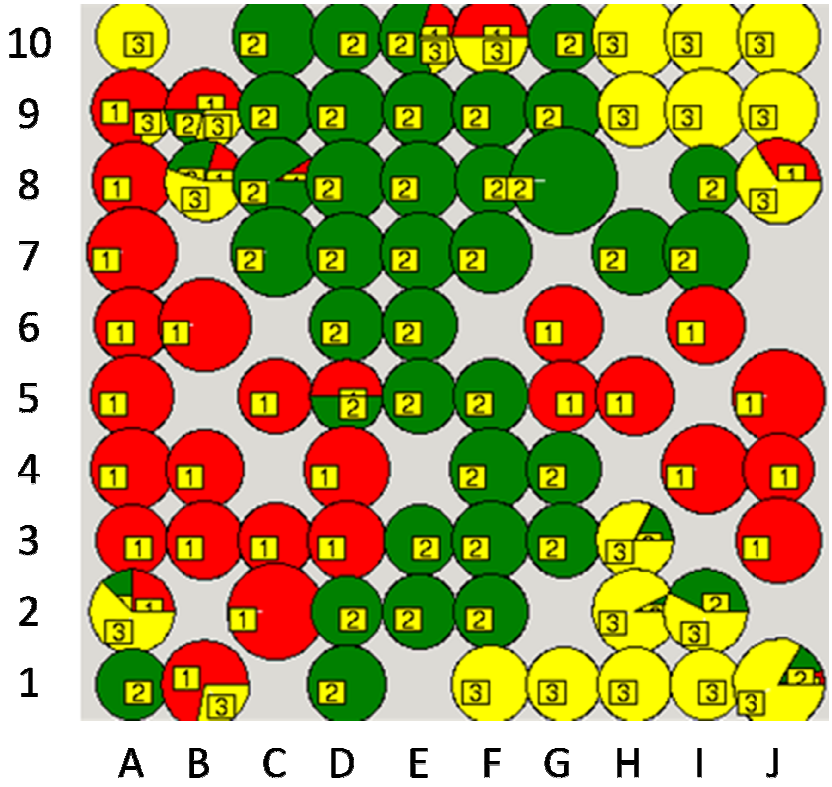


Figure 4-6. Carte de Kohonen 10x10 séparant les 3 premières classes enzymatiques de la KEGG.
Seules les séquences de liaisons ont été utilisées. En rouge les oxydoréductases, en vert les transférases et en jaune les hydrolases.

On trouve d'autres cas pour lesquels des réactions catalysées par une même enzyme se trouvent dans des clusters différents. En effet les réactions catalysées par l'enzyme 1.10.3.1 sortent dans des clusters différents (neurones 9B et 5C), et pour cause la transformation observée est différente (Figure 4-8). Contrairement au cas précédent il ne s'agit pas juste d'une inversion du sens de la réaction. Les 2 réactions sont considérées comme des oxydations où l'oxygène joue le rôle d'accepteur. La coupure de la double liaison du dioxygène intervient dans les 2 cas mais le devenir des oxygènes est différent. Dans le 1er cas on observe 3 liaisons dynamiques et dans le 2^{e} cas 7 liaisons dynamiques. Il n'est donc pas étonnant que ces 2 réactions soient placées à des endroits différents sur la carte.

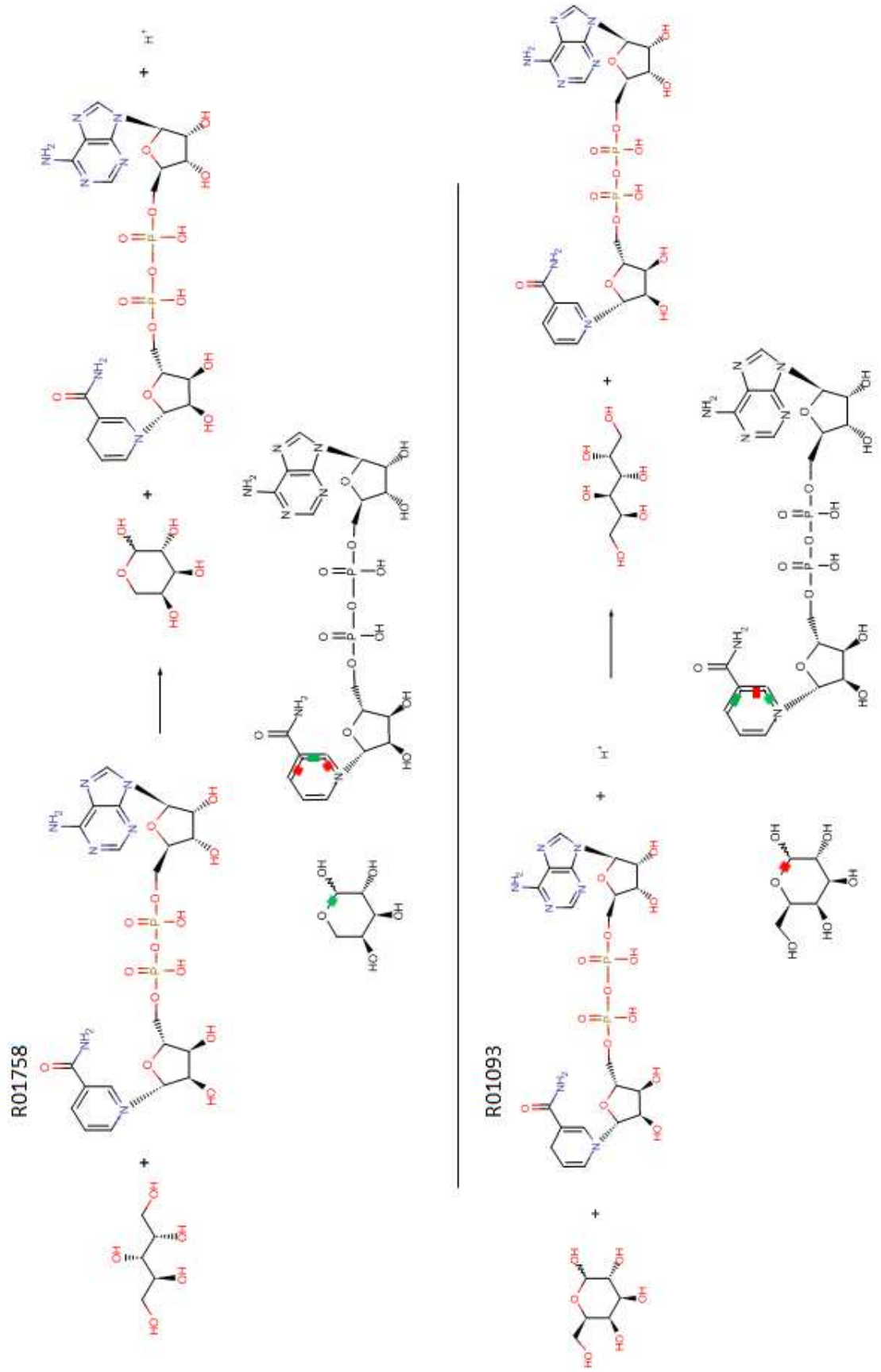


Figure 4-7. Réactions R01758 et R01093 avec leur GCR respectifs.
La partie haute correspond à la réaction R01758 et la partie basse à la réaction R01093 de la KEGG.

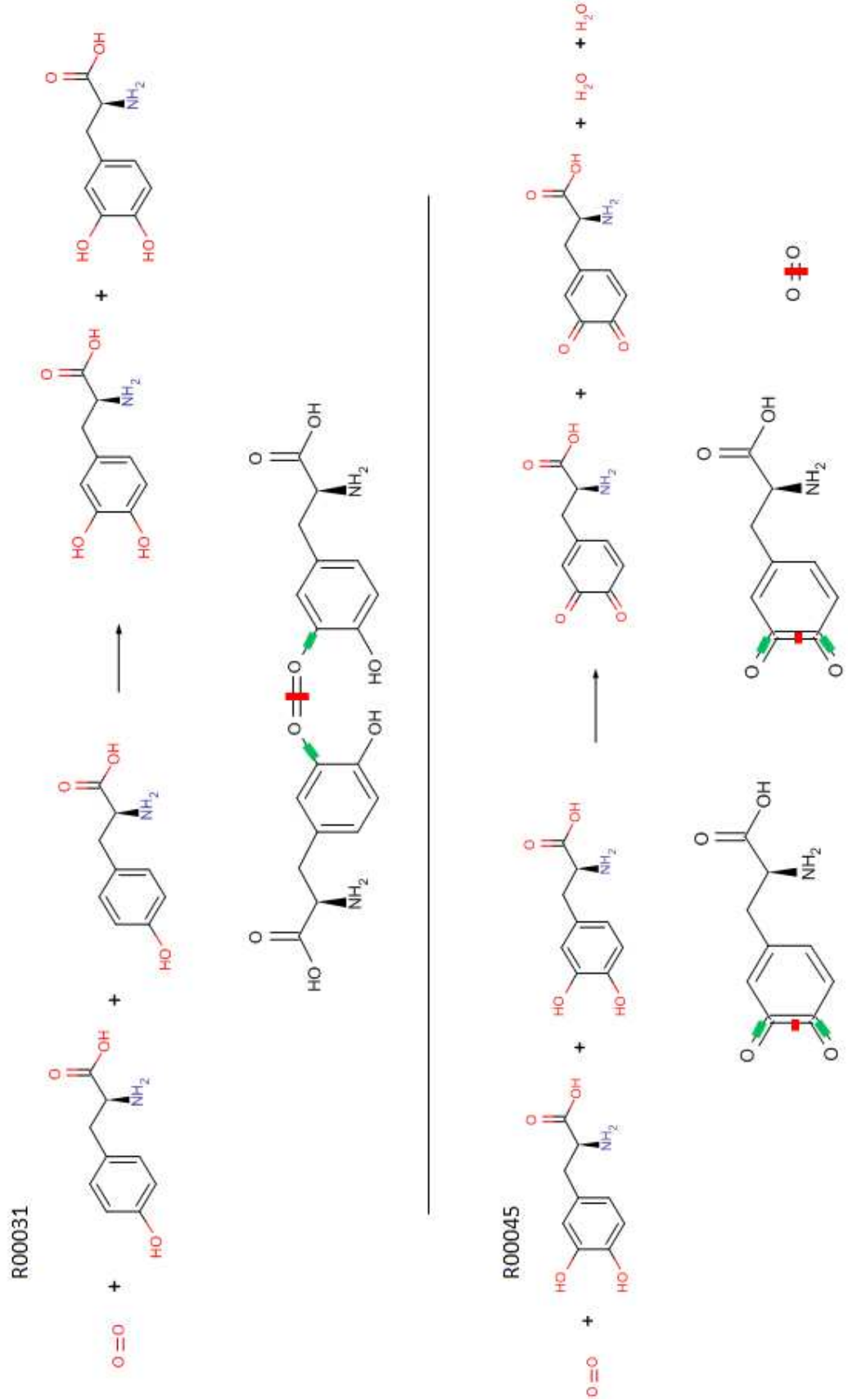


Figure 4-8. Réactions R00031 et R00045 avec leur GCR respectifs.
La partie haute correspond à la réaction R00031 et la partie basse à la réaction R00045 de la KEGG.

La clusterisation est donc bien basée sur la réaction mise en jeu ce qui illustre certaines limites de la nomenclature E.C.

Dans d'autres neurones on peut retrouver des mélanges de plusieurs réactions ayant des identifiants E.C. différents. C'est le cas par exemple du neurone 1B où l'on retrouve des oxydoréductases ainsi que des hydrolases. En regardant de plus prêt on constate que les transformations qui ont lieux sont très similaires. Par exemple les réactions R00123 et R00447 (Figure 4-9) sont catalysées par des enzymes différentes mais les GCRs respectifs laissent apparaître un certain nombre de fragments semblables et seulement 1 liaison dynamique différente due à l'oxydation du dioxygène. Etant donné qu'un seul fragment représente la transformation du dioxygène, la prise en compte de cette différence sera minimale dans le calcul de similarité. On peut voir là une faiblesse dans le calcul de la similarité due à la présence de petits substrats qui de par leur taille ne permettent pas forcément de différencier entre une oxydoréduction et une hydrolase. D'un autre côté le fait d'omettre les substrats secondaires des réactions pourrait permettre de mettre en évidence des réactions semblables vis-à-vis des transformations faites sur le substrat principal, et ainsi d'identifier des chemins métaboliques parallèles permettant d'arriver au même résultat mais de façon différente.

On trouve d'autres clusters contenant des mélanges de réactions pour lesquels la distance calculée entre les réactions est élevée mais qui ont été classées ensemble par défaut. Cela peut être du à l'absence de réactions similaires dans le jeu de données.

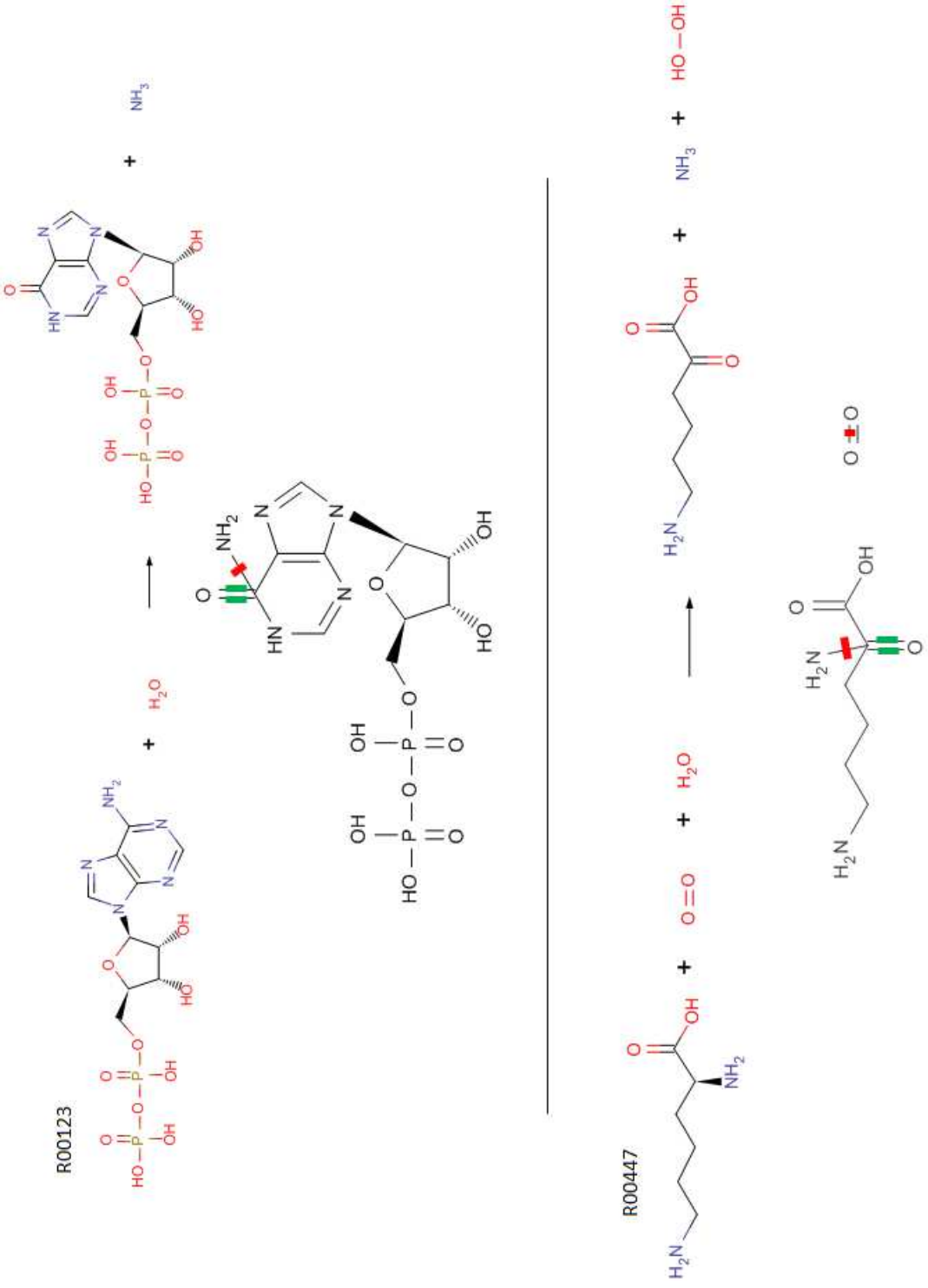


Figure 4-9. Réactions R00123 et R00447 avec leur GCR respectifs.
La partie haute correspond à la réaction R00123 et la partie basse à la réaction R00447 de la KEGG.

4.4. Conclusion

Dans ce chapitre, les graphes condensés de réaction ont été utilisés pour classer 627 réactions métaboliques appartenant à 3 classes réactionnelles distinctes. 180 jeux de descripteurs ont été générés. Lors d'une première étape, des clusterings de type K-Means ont été entrepris pour déterminer le jeu de descripteurs et le nombre de clusters entrainant les meilleures performances. L'emploi simultané de séquences de liaisons impliquant des longueurs de fragments allant de 2 à 5 atomes et d'un nombre de 99 clusters entraine la meilleure séparation. Une carte de Kohonen de taille 10x10 basée sur le jeu de descripteurs sélectionné a alors été construite. L'observation de cette carte montre une nette séparation des réactions selon le type de réaction catalysée.

Cela confirme donc que les graphes condensés de réactions représentent correctement l'information réactionnelle présente dans une réaction métabolique. On remarquera que, malgré le fait que l'approche utilisée ici est différente de celle des travaux de Gasteiger [10] et Aires-de Sousa [8], l'information sur les liaisons seule a permis d'obtenir nos meilleurs modèles. Quelque soit l'approche, l'information sur les atomes n'est donc pas utilisée.

Notre méthode se distingue des 2 dernières par le fait que les liaisons prises en compte ne sont pas uniquement les liaisons réactives mais aussi celles se situant dans l'environnement proche de celles-ci. On peut donc penser qu'indirectement les fragments générés prennent en compte les possibles effets de conjugaisons etc. reflétés par les descripteurs empiriques physico-chimiques calculés dans les 2 autres études.

Comme perspectives de ce travail, il serait intéressant pour un plus grand jeu de données de déterminer les 2^e et 3^e chiffres du nombre E.C. Le 4^e chiffre est plus compliqué à déterminer étant donné qu'à ce niveau hiérarchique il n'y a souvent qu'une seule réaction qui correspond à cet E.C. Mais à défaut de déterminer le 4^e chiffre, la réaction pourrait être placée dans un cluster isolé sur la carte de Kohonen. Pour commencer on pourrait envisager de traiter les 6 classes enzymatiques de la KEGG au lieu de se limiter uniquement aux 3 premières.

4.5. Références

1. Babbitt, P.C., *Definitions of enzyme function for the structural genomics era.* Curr. Opin. Chem. Biol., 2003. **7**(2): p. 230-237.
2. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Evolution of function in protein superfamilies, from a structural perspective.* J. Mol. Biol., 2001. **307**(4): p. 1113-1143.
3. Green, M.L. and P.D. Karp, *Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers.* Nucleic Acids Res., **33**(13): p. 4035-4039.
4. Kotera, M., et al., *Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions.* J. Am. Chem. Soc., 2004. **126**(50): p. 16487-16498.
5. Izrailev, S. and M.A. Farnum, *Enzyme classification by ligand binding.* Proteins: Struct., Funct., Bioinf., 2004. **57**(4): p. 711-724.
6. Dobson, P.D. and A.J. Doig, *Predicting Enzyme Class From Protein Structure Without Alignments.* J. Mol. Biol., 2005. **345**(1): p. 187-199.
7. O'Boyle, N.M., et al., *Using Reaction Mechanism to Measure Enzyme Similarity.* J. Mol. Biol., 2007. **368**(5): p. 1484-1499.
8. Latino, D., Q.-Y. Zhang, and J. Aires-de-Sousa, *Genome-Scale Classification of Metabolic Reactions and Assignment of EC Numbers with Self-Organizing Maps.* Bioinformatics, 2008. **24**(19): p. 2236-2244.
9. Zhang, Q.-Y. and J. Aires-de-Sousa, *Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers.* J. Chem. Inf. Model., 2005. **45**(6): p. 1775-1783.
10. Sacher, O., M. Reitz, and J. Gasteiger, *Investigations of Enzyme-Catalyzed Reactions Based on Physicochemical Descriptors Applied to Hydrolases.* J. Chem. Inf. Model., 2009. **49**(6): p. 1525-1534.
11. Faulon, J.-L., et al., *Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor.* Bioinformatics, 2008. **24**(2): p. 225-233.

5. Prédictions des sites d'oxydation pour les substrats de l'isoenzyme CYP1A2 et CYP3A4 chez l'homme.

5.1. Introduction

Actuellement le développement de nouveaux médicaments n'est plus uniquement basé sur la perpétuelle augmentation de l'activité vis-à-vis d'une cible spécifique. En effet les paramètres ADMET [1] (absorption, distribution, métabolisme, excrétion, toxicité) doivent être pris en compte dès les premières étapes de la recherche afin d'éviter l'échec des médicaments dans des étapes plus avancées du développement et ainsi de minimiser les pertes que cela induit.

Le devenir des composés dans l'organisme humain, et plus spécifiquement des xénobiotiques, est une des propriétés les plus compliquées à prédire actuellement. Les composés sont dégradés chez l'homme par plusieurs enzymes et peuvent donner lieu à plusieurs nouveaux composés appelés métabolites. Le métabolisme des médicaments est généralement séparé en 2 étapes. Durant la première étape les molécules sont principalement oxydées pour former des composés plus polaires et donc plus solubles. Ce processus est souvent suivi d'une 2^{e} étape de conjugaison. Les métabolites issus de ces différentes réactions peuvent avoir des propriétés différentes comparées au composé initial. De nombreux effets secondaires peuvent alors apparaître au travers des métabolites formés. Les composés initiaux peuvent devenir toxiques, inactifs ou au contraire actifs après qu'ils aient été dégradés. Il est donc nécessaire de connaître à l'avance le devenir d'un composé à l'intérieur de l'organisme humain. Cependant déterminer le profil métabolique complet pour toutes les molécules dans les étapes initiales de recherche est impossible. Des méthodes in sillico doivent donc être développées. Un résumé des différentes approches est donné dans une revue récente [2].

Plusieurs études proposent déjà de prédire la localisation des sites possibles de métabolisme. On compte principalement les méthodes basées sur la connaissance simultanée de la structure de l'enzyme et du substrat (docking, champs d'interactions moléculaires) [3-5], et celles basées sur la structure du substrat uniquement (énergie d'activation, QSAR, ligand-based) [6-13].

5.1.1. Les cytochromes P450

Les enzymes responsables de la métabolisation des composés dans l'organisme appartiennent à la superfamille d'enzymes du cytochrome P450 (CYP) et sont exprimées dans le foie humain et l'intestin grêle. Ces enzymes ont toutes le même complexe porphyrine-hème leur servant de centre catalytique. La différence entre celles-ci vient du fait qu'elles possèdent des séquences d'acides aminés différentes leur conférant des topologies différentes au niveau du site actif.

Le métabolisme de chaque isoenzyme est déterminé par 3 facteurs :

- La topologie du site actif.
- La facilité d'accès du site potentiel d'oxydation au complexe fer-oxygène de l'hème.
- La facilité avec laquelle les électrons et hydrogènes sont arrachés des différents carbones et hétéroatomes du substrat.

Toutefois le mécanisme de catalyse des P450 est constant pour pratiquement toutes les isoenzymes. Durant l'oxydation du substrat, le cytochrome P450 interagit avec un carbone ou hétéroatome particulier. Il en résulte un produit intermédiaire pouvant être soit un radical (résultant de l'abstraction d'un hydrogène) soit d'un atome chargé (résultant d'un transfert d'un électron). Les recombinaisons d'atomes et liaisons qui découlent de cette molécule intermédiaire dépendent de la nature chimique de la substance oxydée. Les hydroxylations sur un carbone sont un exemple de transformation donnant lieu à un produit intermédiaire de type radicalaire. A l'inverse, les oxygénations d'hétéroatomes donnent un produit intermédiaire pour lequel l'hétéroatome est chargé positivement. La Figure 5-1 illustre ces deux types de réactions.

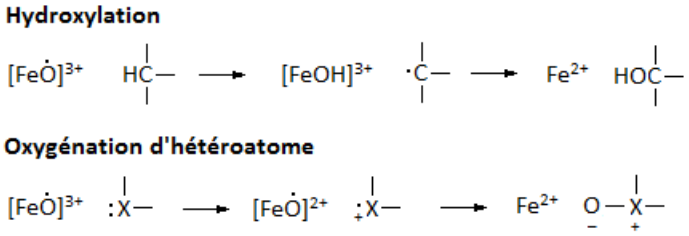


Figure 5-1. Deux types de réactions sont présentés. La réaction d'hydroxylation illustre la formation d'un produit intermédiaire résultant de l'abstraction d'un hydrogène. La réaction d'oxygénation d'un hétéroatome illustre la formation d'un produit intermédiaire via le transfert d'un électron.

5.1.2. Propriétés des substrats et isoenzymes CYP1A2 et CYP3A4.

Le site actif de l'enzyme CYP1A2 est composé de plusieurs groupements aromatiques et possède une forme et une taille restreinte ne permettant qu'aux structures planes d'occuper le site de liaison. Les substrats de cette enzyme seront donc préférentiellement hydrophobes et possèderont des groupements aromatiques plans capables de faire des interactions π-π avec ceux du site actif. Un exemple d'interaction entre le CYP1A2 et l'alpha-naphthoflavone est proposé sur la Figure 5-2.

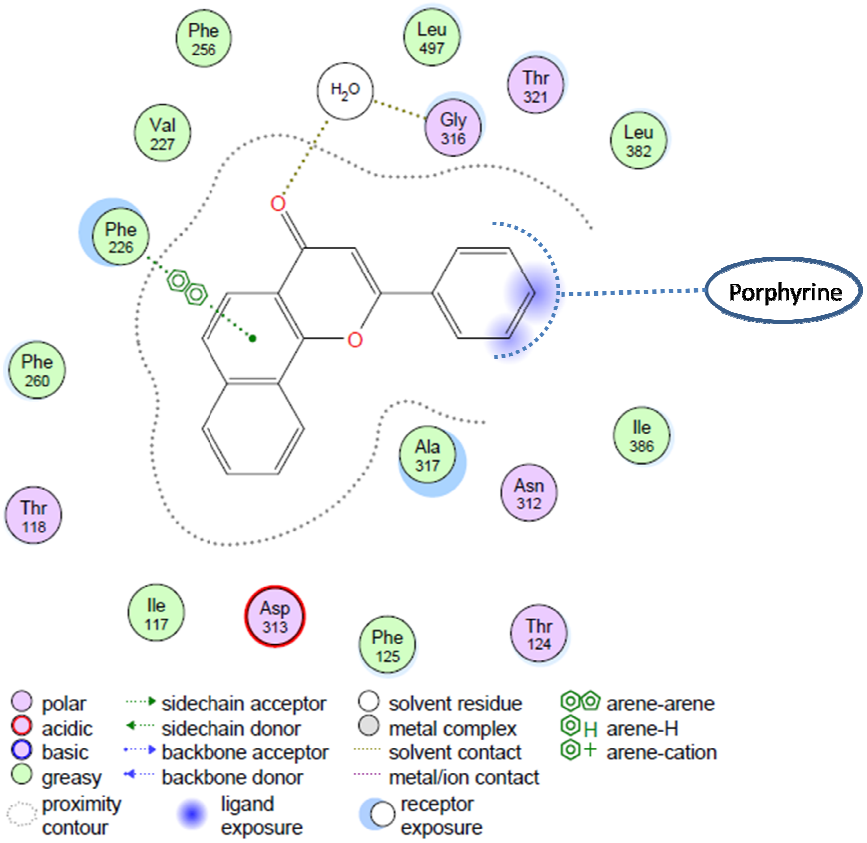


Figure 5-2. Schéma 2D de l'interaction ligand-récepteur de l'alpha-naphthoflavone située dans le site actif du CYP1A2 chez l'homme. La zone d'interaction possible de la porphyrine avec le ligand est schématisée.

L'enzyme CYP3A4 est certainement l'enzyme la moins comprise malgré le fait qu'elle soit l'une des plus répandue dans le foie humain. Environ 50% des médicaments sur le marché sont métabolisés par cette enzyme [14]. Le site actif peut accueillir une large variété de classes de substrats allant de petites molécules (ex : testostérone MW=288) à des molécules plus larges (ex : cyclosporine A, MW=1202).

L'enzyme CYP3A4 se lie essentiellement aux substrats grâce aux interactions lipophiles. Il apparaît que la regiosélectivité des sites d'oxydation dépend plus de la réactivité chimique des sites que de la présence d'interactions spécifiques entre le substrat et le site actif. Les sites de métabolisme sont donc largement dictés par l'aisance avec laquelle les hydrogènes seront arrachés dans le cas d'hydroxylation sur un carbone, ou avec laquelle les électrons seront arrachés dans le cas des réactions d'oxygénation d'hétéroatomes. Un exemple d'interaction entre le CYP3A4 et l'erythromycin A est proposé sur la Figure 5-3.

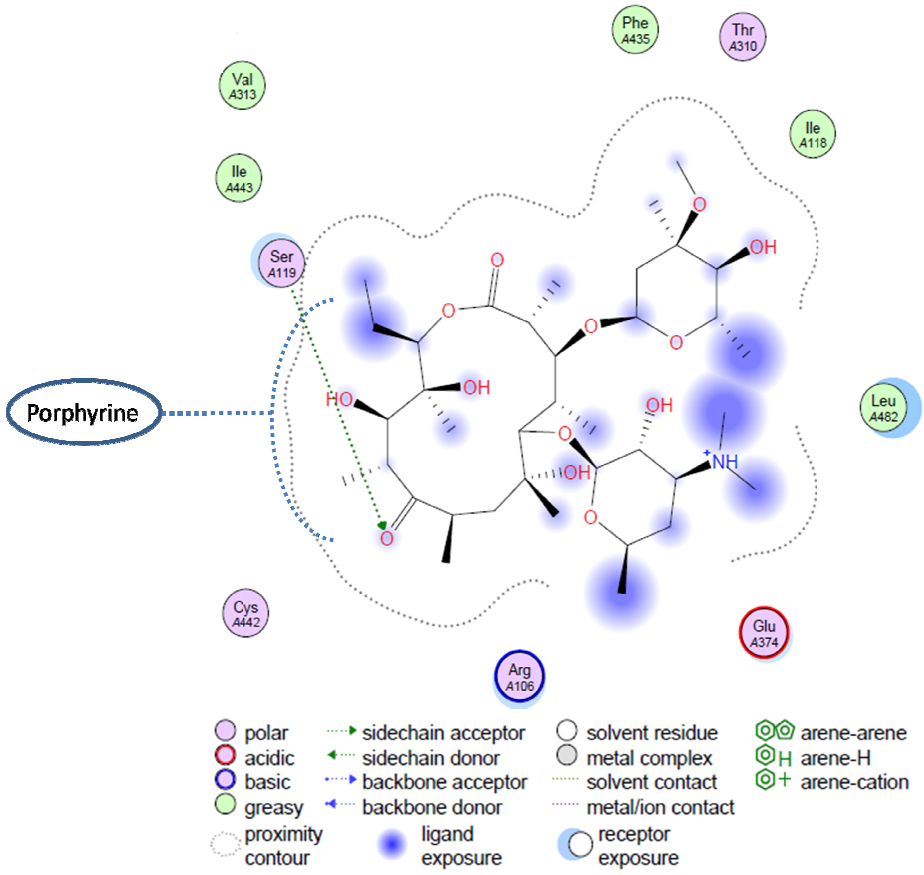


Figure 5-3. Schéma 2D de l'interaction ligand-récepteur de l'erythromycin A située dans le site actif du CYP3A4 chez l'homme. La zone d'interaction possible de la porphyrine avec le ligand est schématisée.

5.2. Prédictions des sites d'hydroxylation aromatique pour les substrats de l'isoenzyme CYP1A2 chez l'homme.

5.2.1. Introduction

Nous proposons de développer dans un premier temps une méthode pour prédire la régiosélectivité des transformations d'hydroxylation aromatique pour les substrats de l'isoenzyme CYP1A2 chez l'homme. Les hydroxylations aromatiques expérimentalement observées ont été extraites de la base de données Metabolite. Pour tous les sites potentiels, les biotransformations d'hydroxylation aromatique ont été générées afin d'avoir aussi un jeu de réactions non observées expérimentalement. L'ensemble des réactions a alors été convertit en GCRs, les descripteurs ISIDA ont été générés, et des modèles NB, RF et SVM ont été construits. Les meilleurs modèles ont montré de bonnes performances comparés au logiciel MetaSite.

5.2.2. Méthodologie

5.2.2.1. Données

Les transformations expérimentalement observées d'hydroxylation aromatique pour l'isoenzyme CYP1A2 chez l'homme ont été extraites de la base de données Metabolite. Seules les réactions pour lesquelles un seul site aromatique à la fois est hydroxylé ont été conservées. A partir de là, les substrats de chaque réaction ont été extraits. Les formes tautomères et la structure aromatique des substrats a été canonisée avec Standardizer [15] de ChemAxon. Les doublons ont ensuite été recherchés et éliminés.

Finalement 96 substrats ont été conservés. A partir de ceux-ci, 588 transformations potentielles d'hydroxylation aromatique ont été générées afin de former un jeu d'entraînement. Parmi ces réactions, 136 ont été expérimentalement observées alors que les 452 autres non.

Les réactions ont alors toutes été transformées en graphes condensés de réaction. Le processus de création du jeu de données est résumé Figure 5-4.

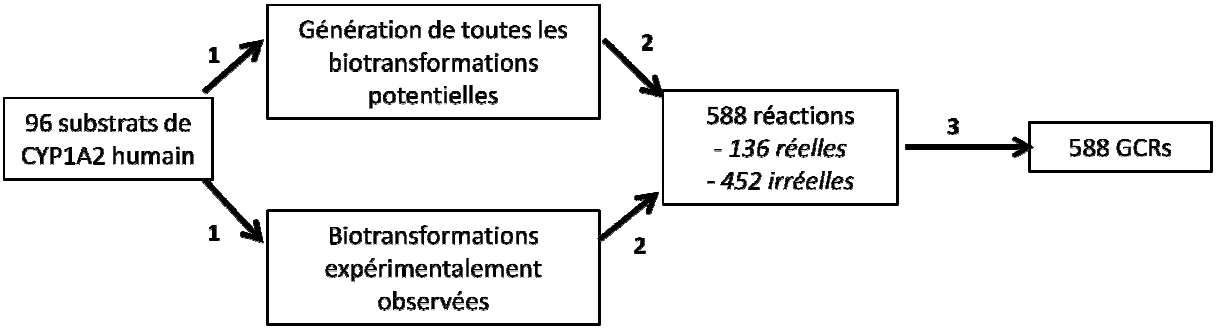


Figure 5-4. Flux de données.

1. Pour les substrats de l'isoenzyme CYP1A2 de l'homme impliqués dans des biotransformations expérimentalement observées, toutes les biotransformations potentielles sont générées.
2. L'intersection des biotransformations observées et non observées permet d'identifier les sites d'hydroxylation aromatique expérimentalement observés et non observés.
3. L'ensemble des réactions est transformé en Graphes Condensés de Réactions.

5.2.2.2. Génération des transformations potentielles

Pour tous les substrats, les biotransformations potentielles ont été générées par le « Laboratory of Structure-Function Based Drug Design » de l'académie des sciences médicales de Moscou dirigé par le professeur V. Poroikov à l'aide de la méthode appelée P450CARBOX (« **P450 carb**on **ox**idation ») [6]. Cette méthode consiste à hydroxyler un carbone aromatique dès qu'un tel site est identifié. Un exemple de génération de biotransformations potentielles est donnée Figure 5-5.

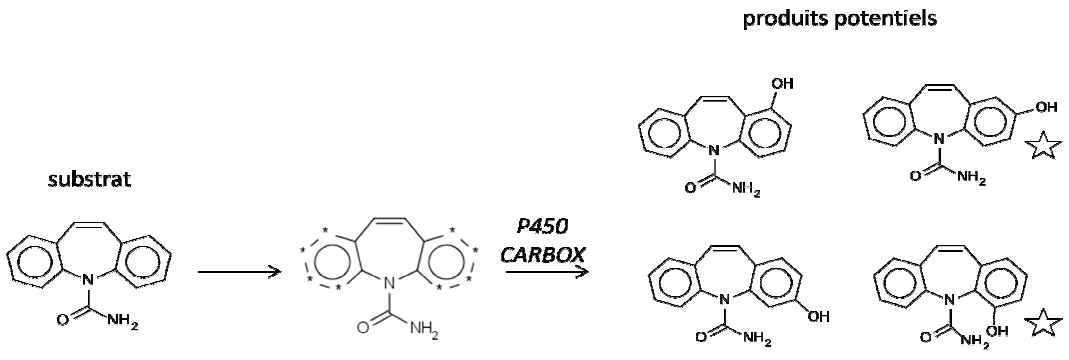


Figure 5-5: Exemple de biotransformations possible pour un substrat.
Les astérisques indiquent les sites potentiels d'hydroxylation aromatique. Les étoiles indiquent les produits observés expérimentalement.

5.2.2.3. Descripteurs

Pour chaque GCR les descripteurs calculés sont les descripteurs ISIDA SMF. Seuls les fragments contenant au minimum une liaison dynamique ont été générés. Au total 248 fragmentations ont été générées, cela comprend les séquences d'atomes (IA), de liaisons (IB), d'atomes et liaisons (IAB), de paires d'atomes (IAP) ainsi que les atomes unis (IIIA, IIIB, IIIAB, IIIAP). La longueur minimale des fragments générés a été variée de 3 à 10 pour les séquences, et de 3 à 6 pour les atomes unis. La longueur maximale des fragments générés a été variée de 3 à 10 pour les séquences d'atomes unis. Etant donné que les substrats de l'isoenzyme CYP1A2 ont une conformation relativement plate dans la cavité [16], l'utilisation de descripteurs 2D semble être judicieuse.

5.2.2.4. Machines d'apprentissages.

Les machines d'apprentissages testées dans cette étude sont le bayésien naïf, la forêt aléatoire et la machine à vecteurs support.

Les paramètres par défaut fournis par Weka ont été conservés pour le bayésien naïf. Pour RF, le nombre d'arbres de la forêt à été fixé à 100 et chaque arbre est construit sur la base de 10 descripteurs sélectionnés aléatoirement.

La construction des modèles SVM a été effectuée grâce au logiciel LIBSVM et le coefficient de similarité de Tanimoto a été utilisé comme noyau. Le coût de la SVM a été optimisé en validation croisée à 5 paquets pour chaque jeu d'entraînement afin d'obtenir les meilleures PB pour chaque jeu de descripteurs. Le coût a été varié de 1 à 91 avec un pas de 10. Le poids des objets appartenant à la classe des transformations observées a été modifié de façon à équilibrer chaque jeu de modélisation. Afin d'avoir des modèles probabilistes, pour comparer notre approche à celle de MetaSite, les modèles SVM ont été reconstruits sur les meilleurs jeux de descripteurs en sélectionnant l'option permettant d'avoir des probabilités en sortie pour chaque site potentiel d'hydroxylation aromatique. Les performances de ces derniers modèles ont été estimées à l'aide du critère IAP en 5-CV et sur les 5 jeux de test.

5.2.2.5. Validation des modèles.

Le jeu de données a été divisé aléatoirement en 5 jeux d'entraînements et de tests. Les performances prédictives des modèles ont été estimées en 5-CV pour chaque jeu d'entraînement et la précision balancée moyenne a été calculée. Il a été fait attention à ce que les oxydations effectuées pour un même substrat aient toutes été mises dans le même paquet lors de la validation croisée.

Pour les meilleurs jeux de descripteurs et pour chaque jeu de modélisation les données ont été soumises à une procédure de scrambling puis à une validation croisée à 5 paquets. Cette procédure a été répétée 3 fois. Si les modèles construits ne sont pas due à une corrélation fortuite entre les descripteurs et la propriété prédite, alors les performances trouvées en scrambling doivent être inférieures à celles trouvées sur le jeu d'origine.

Finalement, un modèle consensus a été construit en utilisant les modèles possédant la meilleure précision balancée et les performances prédictives des modèles ont été testées sur les 5 jeux de tests. La stratégie de modélisation est résumée Figure 5-6.

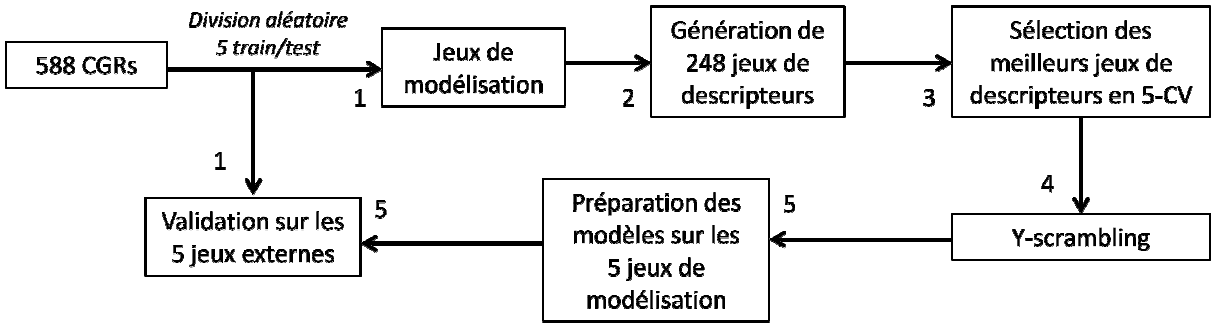


Figure 5-6. Stratégie de modélisation.

1. Les 588 GCRs correspondant aux transformations observées et non observées sont divisés aléatoirement en 5 paires de jeux d'entraînement (4/5 des données) et de test (1/5 des données).
2. 248 jeux de descripteurs ISIDA sont générés pour chaque jeu d'entraînement.
3. Une procédure de validation croisée à 5 paquets est exécutée pour chaque jeu d'entraînement et les descripteurs amenant les meilleurs modèles sont sélectionnés.
4. Une procédure de scrambling est mise en place pour la construction de modèles basés sur les descripteurs sélectionnés. Seuls les descripteurs n'entrainant pas de modèles fortuits sont conservés.
5. Les modèles sont construits pour chaque jeu d'entraînement à partir des descripteurs sélectionnés et appliqués sur le jeu de test correspondant.

5.2.2.6. Etude comparative

MetaSite [4] sera utilisé afin de comparer les performances de nos modèles avec celui-ci. Pour prédire les sites de métabolisme, le logiciel calcule deux ensembles de descripteurs représentés sous la forme de fingerprint, un pour l'enzyme basé sur les champs d'interactions moléculaires (GRID flexible molecular interaction fields), et un pour le substrat basé sur la reconnaissance des propriétés pharmacophoriques de chaque atome (GRID probe pharmacophore recognition). Les deux fingerprints sont alors comparés et une composante d'accessibilité E est calculée pour chaque atome du substrat. Cette composante est fonction de la complémentarité entre les types d'atomes du substrat (hydrophobes, accepteur de liaison hydrogène, …) avec les régions à l'intérieur de la cavité de la protéine (hydrophobe, donneur de liaison hydrogène, …), et en fonction de la distance de l'atome du substrat avec le centre réactionnel de l'hème. Donc un score d'accessibilité maximale pour un atome correspond à placer un substrat dans la cavité de façon à avoir une complémentarité maximale entre les propriétés pharmacophoriques de celui-ci avec les régions d'interactions de la protéine, et que l'atome considéré soit proche du centre réactionnel de l'hème.

Une autre composante, la composante de réactivité R, est aussi calculée pour chaque atome. Pour un atome donné du substrat et un type de mécanisme réactionnel donné (C-hydroxylation, N-dealkylation, dehalogénation, …) la composante de réactivité représente l'énergie d'activation qu'il faut fournir pour former le produit intermédiaire. Cette énergie est calculée à l'aide de calculs quantiques semi-empirique.

La probabilité d'un atome d'être oxydé correspond alors au produit de sa composante d'accessibilité avec sa composante de réactivité : $P(i) = E_i \cdot R_i$, où i correspond à un atome donné.

Dans cette étude seules les prédictions faites par MetaSite pour les sites aromatiques capables d'être hydroxylés sont prises en compte.

5.2.3. Résultats

5.2.3.1. Validation croisée

Dans un premier temps, des calculs utilisant plusieurs machines d'apprentissage ont été entrepris. La précision balancée a été calculée pour chaque jeu de modélisation en validation croisée à 5 paquets et la valeur moyenne a été calculée. La PB moyenne des modèles pour les 3 jeux de descripteurs amenant les meilleures performances et pour chaque méthode d'apprentissage est donnée dans le tableau suivant (Tableau 5-1).

Tableau 5-1. Précision balancée moyenne pour les modèles des 3 meilleures fragmentations de chaque méthode.

NB		**RF**		**SVM**	
descripteurs	**PB**	**descripteurs**	**PB**	**descripteurs**	**PB**
IIIAP(5-7)	0,71	IAB(3-6)	0,737	IIIB(5-6)	0,771
IIIB(5-9)	0,707	IAB(5-6)	0,715	IAB(3-6)	0,768
IIIB(5-6)	0,702	IAB(5-7)	0,713	IIIB(4-6)	0,767

Comme on peut le voir, la SVM permet d'atteindre les meilleures performances. Le 3^{e} meilleur modèle SVM est même plus performant que les meilleurs modèles des deux autres méthodes. On peut tout de même voir quelques similarités entre les 3 méthodes. En effet, on retrouve les mêmes jeux de descripteurs pour les meilleurs modèles SVM dans les meilleures performances des deux autres machines d'apprentissage. A la vue de ces résultats il est logique que la SVM soit choisie pour la suite des investigations de cette étude.

Afin de créer un modèle consensus, la distribution de la précision balancée a été tracée pour la SVM et tous les modèles situés après le pic maximum apparent de la PB ont été conservés. Cela correspond aux modèles ayant une PB supérieure à 0,73, c'est-à-dire 33 modèles. Nous avons testé plusieurs consensus intégrant plus ou moins de modèles mais contenant toujours un nombre impair de modèles (1, 3, 5,…) afin d'être sûr d'obtenir toujours un vote majoritaire concernant les prédictions. Les performances prédictives des différents consensus ont été calculées en 5-CV

pour chaque jeu de modélisation et la valeur moyenne de la PB a été calculée. Les performances des modèles consensus en fonction du nombre de modèles les composant sont données Figure 5-7.

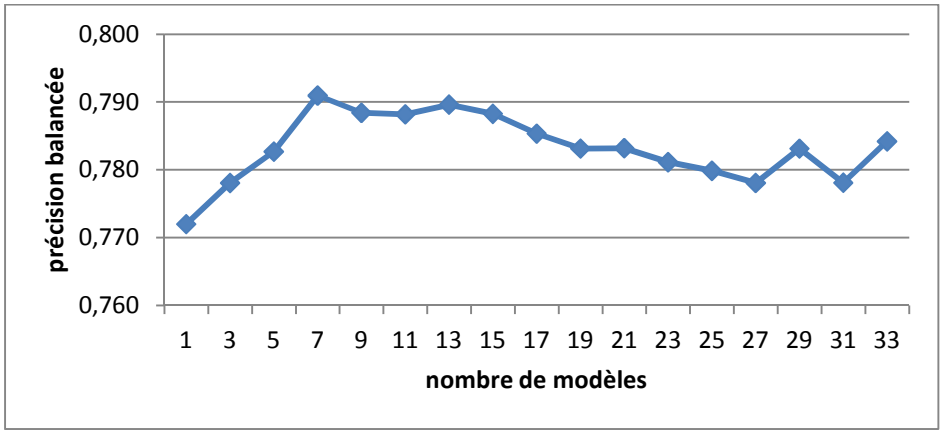


Figure 5-7. Performances des modèles consensus SVM moyennés sur les 5 jeux de modélisation. Chaque ensemble de modèles est constitué des meilleurs modèles.

Le modèle consensus est toujours plus efficace que l'utilisation seule du meilleur modèle. Toutefois, après 15 modèles on voit que la PB a tendance à diminuer pour donner pratiquement des performances comparables à celle du meilleur modèle individuel. Etant donné que la PB semble atteindre un maximum pour 7 modèles inclus dans le consensus (PB≈0,79), il semble judicieux de construire un consensus final composé des 7 meilleurs modèles individuels (0,755 < PB < 0,771) afin de faire des prédictions sur des jeux externes.

Des performances plus détaillées sont données Tableau 5-2 pour les modèles SVM construits sur la meilleure fragmentation et pour les consensus sur les 5 jeux de modélisation. En ne considérant que les liaisons des fragments de type atomes unis augmentés allant d'une longueur de 5 à 6 atomes (IIIB(5-6)) la PB varie pour les 5 jeux de modélisation entre 0,694 et 0,807. Pour les consensus, la PB varie entre 0,728 et 0,826. Pour le 5^{e} jeu de modélisation les performances sont moins bonnes que pour les autres jeux. Sans considérer ce dernier jeu la performance des différents modèles est relativement stable et la PB varie entre 0,773 et 0,807 pour les modèles construits sur les descripteurs IIIB(5-6) et entre 0,792 et 0,826 pour les consensus.

Tableau 5-2. Précisions balancées obtenues pour les modèles SVM construits sur la meilleure fragmentation et pour les modèles consensus sur les 5 jeux de modélisations en 5-CV.

jeu de modélisation	5-CV (PB)	descripteurs
1	0,793	IIIB(5-6)
	0,815	consensus
2	0,807	IIIB(5-6)
	0,826	consensus
3	0,773	IIIB(5-6)
	0,792	consensus
4	0,786	IIIB(5-6)
	0,794	consensus
5	0,694	IIIB(5-6)
	0,728	consensus

5.2.3.2. Y-scrambling

Les performances obtenues en scrambling ont été calculées pour toutes les fragmentations utilisées pour construire les modèles inclus dans le consensus. Pour chaque jeu de modélisation la propriété a été aléatoirement mélangée et une 5-CV a été effectuée. Cette procédure a été répétée 3 fois. Les modèles ainsi construits ont permis d'obtenir un total de 75 valeurs de PB. Il a été calculé que la PB maximale qu'un modèle scramblé puisse atteindre sur un paquet seul est de 0,69. La différence de PB en scrambling avec la PB sur le jeu de données original est suffisamment significative pour affirmer que la performance des modèles sélectionnés n'est pas due à une corrélation fortuite des descripteurs avec la propriété modélisée.

5.2.3.3. Validation externe

Les prédictions sur les jeux externes amènent comme escompté les performances espérées en validation croisée (voir Tableau 5-3). Les PB varient entre 0,705 et 0,82 pour les modèles construits sur les descripteurs IIIB(5-6) et entre 0,708 et 0,872 pour les modèles consensus. On note pour certains jeux de modélisation/test une différence de performance entre la validation croisée et la validation externe. Pour le jeu 2 par exemple, le modèle consensus passe d'une PB de 0,826 à 0,708. Inversement pour le jeu 5 où le modèle consensus passe d'une PB de 0,728 à 0,872. Cette différence peut être expliquée par le fait que les jeux de test sont composés de

15 à 20 substrats, les sites d'hydroxylation aromatique mal prédits pour un substrat entraine de suite une baisse non négligeable des performances, et inversement. On peut penser que pour un test set externe plus large les performances auraient été plus similaires aux performances de la validation croisée.

Tableau 5-3. Précision balancée obtenues pour les modèles SVM construits sur la meilleure fragmentation et pour le consensus sur les 5 validations externes.

jeu de test	PB	descripteurs
1	0,766	IIIB(5-6)
	0,753	consensus
2	0,705	IIIB(5-6)
	0,708	consensus
3	0,769	IIIB(5-6)
	0,769	consensus
4	0,797	IIIB(5-6)
	0,786	consensus
5	0,820	IIIB(5-6)
	0,872	consensus

La Figure 5-8 propose des exemples de substrats pour lesquels tous les sites d'hydroxylation aromatique ont été bien prédits par les modèles consensus. Pour le premier substrat la prédiction n'est pas très difficile étant donné qu'il n'y a que 4 sites possibles d'hydroxylation aromatique. Mais comme on peut le voir, pour certaines molécules (substrat 3) il y a 11 sites possibles et le modèle prédit tout de même correctement lequel est réactif et lesquels ne le sont pas. Pour le substrat 4 on peut constater la réelle force de notre approche pour prédire la régiosélectivité. En fait, la plupart des approches considérées dans la littérature sont capables de proposer uniquement un classement des différents sites en fonction de leur probabilités d'être oxydés ou non sans pour autant prédire le nombre exact de sites réactifs pour le substrat. Au contraire nos modèles prédisent à la fois les sites d'hydroxylation et ceux qui ne le sont pas, et sont donc capables de prédire pour un substrat donné autant de sites d'hydroxylation qu'il est susceptible d'y en avoir.

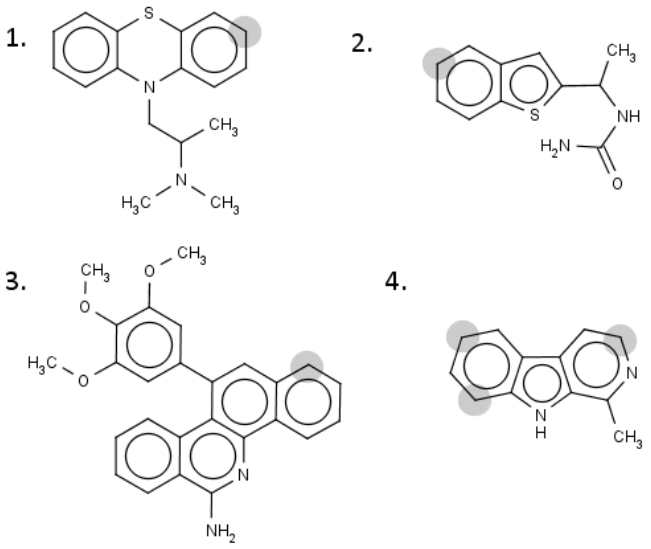


Figure 5-8: Structures chimiques de substrats de l'isoenzyme CYP1A2 des jeux externes.
Les cercles grisés indiquent les sites expérimentaux d'hydroxylation aromatique correctement prédit par les modèles consensus.

Nous allons ensuite observer quelques exemples de substrats pour lesquels les sites d'hydroxylation ont été mal prédits par nos modèles et essayer de comprendre pourquoi. Deux exemples de sites d'oxydation mal classés sont proposés Figure 5-9. On peut voir sur l'exemple **a.** que les structures les plus similaires au substrat proposé sur la gauche de la figure vis-à-vis des vecteurs de descripteurs présentent toutes un site d'hydroxylation aromatique expérimentalement observé alors que ce même site n'est pas réactif pour les structures les plus similaires. On remarque que le groupement éthyle initialement situé 2 atomes après le site considéré n'est plus présent pour les cas les plus similaires. On peut aisément penser que ce simple changement pourrait affecter la réactivité du site étant donné que le changement de structure est, qui plus est, très proche du site regardé. Toutefois l'erreur de prédiction pourrait aussi venir du fait que le métabolite n'a pas encore été expérimentalement observé.

Pour l'exemple **b.**, le site d'oxydation présenté pour le substrat de gauche est expérimentalement vérifié alors que les sites d'oxydation avec l'environnement structural le plus similaire sont non réactifs. Si on regarde de plus près les structures on s'aperçoit qu'il n'y a pas de changement structural entre le substrat regardé et les molécules les plus similaires jusqu'à 6 atomes de distance du site d'oxydation. Il n'est donc pas surprenant que le site d'hydroxylation aromatique ait été

incorrectement prédit pour le substrat étant donné que les vecteurs de descripteurs sont quasiment identiques. En cherchant dans la littérature on s'aperçoit que la formation d'un métabolite hydroxylé en position 4 (site présenté) par l'isoenzyme CYP1A2 chez l'homme n'est pas du tout évidente contrairement à la position 5. On peut donc émettre de sérieux doutes quand à l'existence du métabolite hydroxylé à cette position.

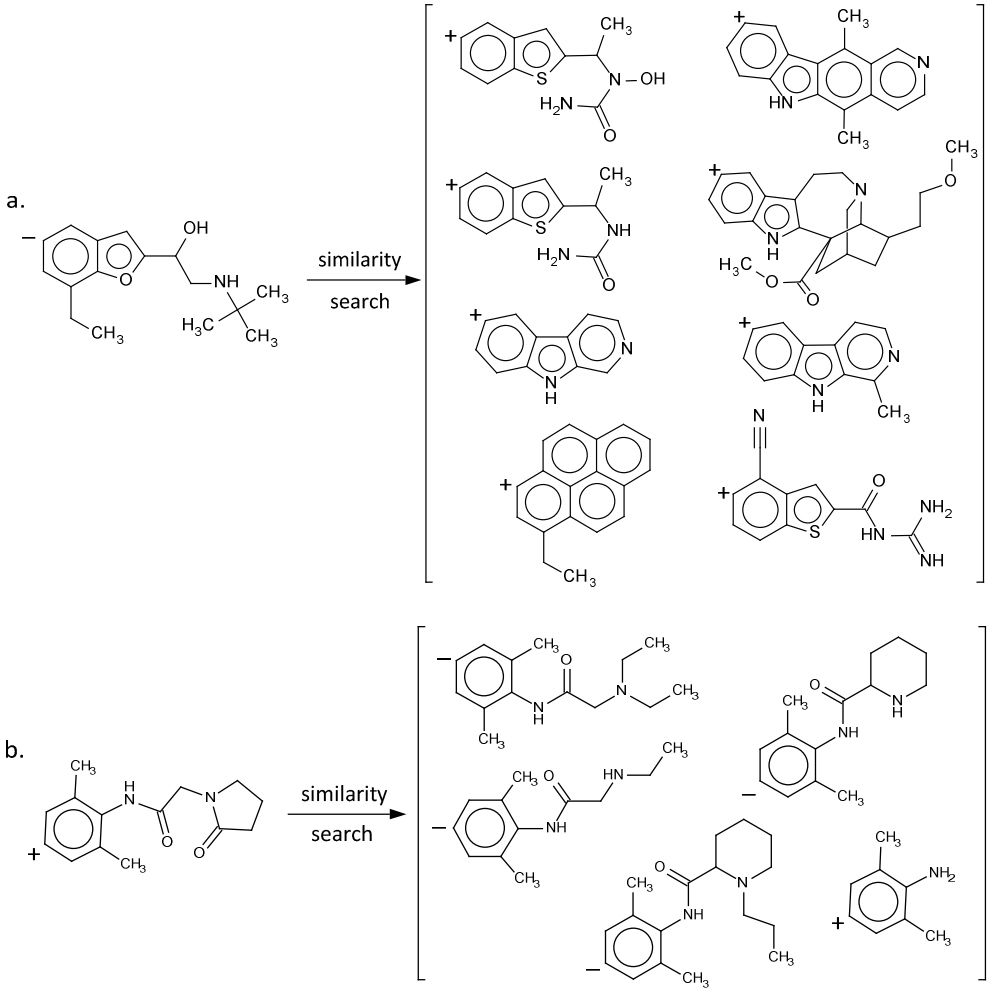


Figure 5-9: Exemples de substrats pour lesquels la prédiction du site d'oxydation est incorrecte**.**
Deux sites d'oxydation (a. and b.) sont présentés sur la gauche de la figure avec à leur droite les sites d'oxydation respectifs les plus similaires vis-à-vis des vecteurs de descripteurs. "+" indique un site d'hydroxylation aromatique expérimentalement observé à l'inverse de "-".

5.2.3.4. Etude comparative

Afin de comparer nos modèles avec ceux de MetaSite nous avons construits des modèles SVM proposant des probabilités pour prédictions au lieu du label de la classe prédite. Les modèles ont été reconstruits sur les mêmes jeux de descripteurs précédemment utilisés et le critère IAP a été calculé sur les 5 jeux externes pour nos modèles ainsi que ceux de MetaSite. Les résultats sont présentés Tableau 5-4.

Tableau 5-4. Statistiques IAP obtenues pour les modèles SVM construits sur la meilleure fragmentation, pour les consensus, et pour MetaSite sur les 5 validations externes.

	Jeux externes (IAP)		
jeu	IIIB(5-6)	consensus	MetaSite
1	0,681	0,805	0,689
2	0,783	0,831	0,743
3	0,917	0,936	0,544
4	0,909	0,912	0,633
5	0,933	0,985	0,477

Les modèles consensus amènent les meilleures performances comparés aux autres modèles. A nouveau on peut voir que les performances du consensus sont aussi les plus stables. D'un point de vue général nos modèles sont plus performants que MetaSite. Néanmoins cette remarque est modérée par le fait que nos modèles se concentrent sur les hydroxylations aromatiques pour CYP1A2 alors que MetaSite fait des prédictions pour tous les types de biotransformations. Le critère IAP est intéressant dans le fait qu'il permet d'avoir une idée de la distance des exemples mal classés par rapport à l'hyperplan séparateur. Ainsi pour un même substrat on peut voir si un site réactif classé comme non réactif sera plus proche de l'hyperplan que les sites réellement non réactifs. Si c'est le cas on devrait alors observer un IAP élevé. Ce phénomène est observé pour le jeu 5. Le critère IAP est très élevé (IAP=0,985) comparé à la PB (PB=0,872). De plus la PB des modèles IIIB(5-6) en validation externe est très proche de celle des modèles consensus, voir parfois meilleure, et pourtant le critère IAP des modèles consensus est à chaque fois supérieur à celui des modèles IIIB(5-6). Donc malgré le fait que la PB des deux types de modèles soit équivalente, les modèles consensus auront tendance à placer les

exemples mal classés plus près de l'hyperplan séparateur que les modèles individuels.

5.2.4. Conclusion

Dans cette partie les graphes condensés de réactions ont été utilisés afin d'identifier les sites d'hydroxylation aromatique des substrats du CYP1A2 chez l'homme. Un jeu de données contenant des biotransformations expérimentalement observées et non observées a été généré et 248 jeux de descripteurs ont été calculés. Des modèles SVM, NB et RF ont été construits à partir de ces jeux de descripteurs. Un consensus a été formé à l'aide des meilleurs modèles individuels SVM obtenus. Les biotransformations expérimentalement observées ont été séparées des biotransformations non observées avec une PB moyenne de 0,78 sur les jeux de test. Les performances de nos modèles ont été comparées au logiciel MetaSite. Il a été montré que nos modèles consensus atteignent de meilleures performances que celui-ci.

La différence majeure, et principal avantage, entre notre approche et celles proposées dans la plupart des autres études repose sur le fait que nous prédisons pour chaque site possible d'hydroxylation aromatique si le site est réactif ou non. Cela permet de prédire quantitativement le nombre potentiel de sites réactifs et ainsi de surmonter le problème qui vise à définir un seuil différenciant les sites réactifs des sites non réactifs dans le cas de modèles probabilistes. De même, notre approche ne nécessite pas de connaître la structure du cytochrome pour être appliquée.

5.3. Prédictions des sites d'oxydation pour les substrats de l'isoenzyme CYP3A4 chez l'homme.

5.3.1. Introduction

Nous allons ici développer une méthode de prédiction des sites d'oxydation de substrats de l'isoenzyme CYP3A4 chez l'homme. Contrairement à beaucoup de méthodes donnant des résultats sous forme de probabilités sans définir de seuil délimitant les sites d'oxydation des sites non réactifs, nous proposons une approche qui permet de déterminer la localisation ainsi que le nombre de sites d'oxydation d'une molécule. Pour ce faire, un groupement hydroxyle a été ajouté pour chaque site d'une molécule sous la forme d'une réaction et chaque réaction a ensuite été convertie GCR. Sous cette forme, les méthodes traditionnelles de fouille de données peuvent être appliquées afin de distinguer les sites de métabolisme des autres sites en regardant l'environnement proche de chaque site. Nous comparons notre approche à la méthode SMARTCyp [17].

5.3.2. Méthodologie

5.3.2.1. Données

Notre jeu de données a été généré à l'aide de la base de données Metabolite [18]. 1385 biotransformations expérimentalement observées et catalysées par l'isoenzyme CYP3A4 ont été extraites de celle-ci. Ces biotransformations comptent des réactions de N-dealkylation, O-dealkylation, attachement d'un groupement oxo sur un carbone, transformation d'aldéhyde en acide carboxylique et inversement, ouverture d'un hétérocycle aliphatique avec détachement d'un atome de carbone, dehalogénation, transformation d'un groupement hydroxyle sur un carbone en cétone/acide carboxylique, hydroxylation aromatique et hydroxylation aliphatique. Un exemple pour chaque type de biotransformation est donné annexe 10.2. Pour ces réactions, les structures ont été aromatisées à l'aide du logiciel *Standardizer [15]* de ChemAxon et les doublons éliminés à l'aide du logiciel *EdiSDF* (disponible sur le site http://infochimie.u-strasbg.fr). De même, seules les réactions pour lesquelles une

biotransformation à la fois est représentée sont conservées. Au final 1211 réactions ont été conservées, ce qui correspond à 779 substrats différents.

Pour chaque substrat présent dans les réactions expérimentalement observées, tous les sites potentiels d'oxydation ont été marqués pour celui-ci. Au total, 49375 hydroxylations ont été générées. Comme précédemment les structures ont été aromatisées, puis les doublons éliminées. Au final, seules 9326 réactions générées ont été conservées. Parmi celles-ci, les sites d'oxydation expérimentalement observés sont identifiés et les sites restant sont considérés comme non oxydés. Le jeu de données contient donc 1211 sites d'oxydation et 8115 sites non oxydés. L'ensemble des réactions d'hydroxylation générées par le logiciel P450 CARBOX a ensuite été transformé en Graphe Condensé de Réactions. Les différentes étapes de création du jeu de données sont résumées Figure 5-10.

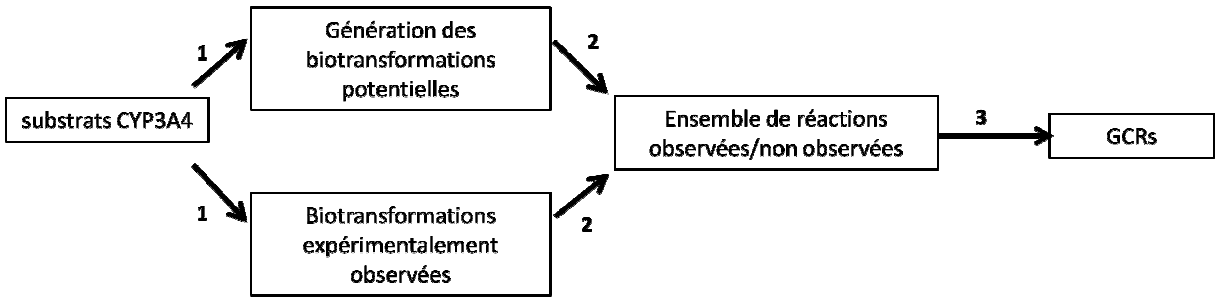


Figure 5-10. Flux de données.

1. Pour les substrats de l'isoenzyme CYP3A4 de l'homme impliqués dans des biotransformations expérimentalement observées, toutes les biotransformations potentielles sont générées.
2. L'intersection des biotransformations observées et non observées permet d'identifier les sites d'oxydation expérimentalement observés et non observés.
3. L'ensemble des réactions est transformé en Graphes Condensés de Réactions.

Un jeu de test externe a également été généré. Les substrats appartenant aux biotransformations expérimentalement observées et catalysées par l'isoenzyme CYP3A4 ont été extraites de la publication de Patrik Rydberg [17]. Seuls les substrats non présents dans notre jeu d'entraînement ont été rapatriés. La même procédure de génération des biotransformations que pour le jeu d'entraînement a été mise en place. Ainsi pour les 56 substrats rapatriés, 733 biotransformations potentielles ont été générées dont 85 correspondant à des sites d'oxydation expérimentalement observés. L'ensemble de ces réactions a alors été transformé en GCRs.

5.3.2.2. Génération des transformations potentielles

Les sites potentiels d'oxydation ont été générés comme pour le CYP1A2 [6]. L'oxydation d'atomes de carbone particulier par le cytochrome P450 pour un xénobiotique est présentée comme l'addition d'un groupement hydroxyle au niveau du site d'oxydation, c'est-à-dire présentée similairement à une hydroxylation. Un exemple de marquage des différents sites potentiels d'oxydation pour un substrat est présenté Figure 5-11.

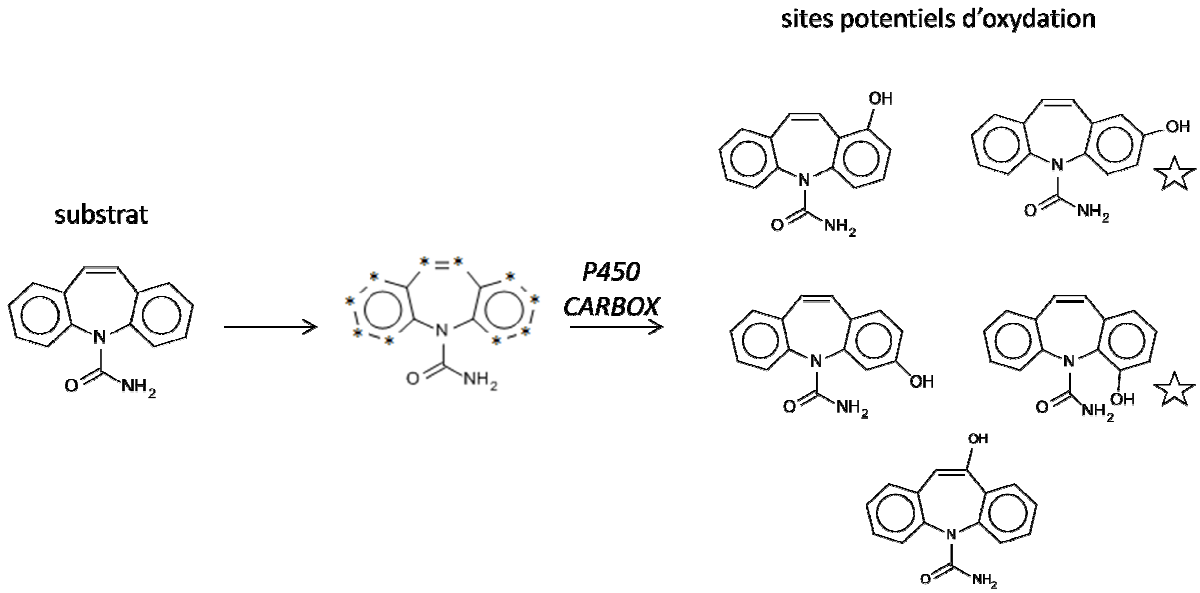


Figure 5-11. Exemple de sites possibles d'oxydation pour un substrat de l'isoenzyme CYP3A4.
Les astérisques représentent les sites d'oxydation possibles. Les étoiles indiquent les biotransformations expérimentalement observées.

Cette façon de représenter les sites d'oxydation correspond en fait (sauf rares exceptions), soit au métabolite final formé dans le cas d'hydroxylation (Figure 5-11), soit à un état intermédiaire comme dans le cas des N-dealkylation par exemple (Figure 5-12) [19]. Des exemples d'oxydation à l'aide de la méthode P450CARBOX sont donnés pour chaque type de biotransformation en annexe 10.2.

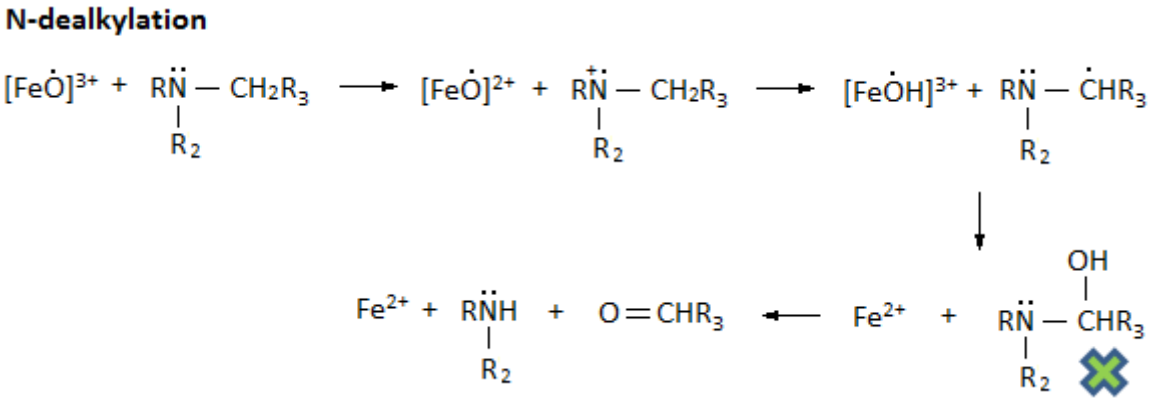


Figure 5-12. Réaction de N-dealkylation via le transfert d'un électron [19].
L'état intermédiaire représenté par P450CARBOX est indiqué par une croix verte.

5.3.2.3. Descripteurs

Pour chaque GCR les descripteurs calculés sont les descripteurs ISIDA SMF. Seuls les fragments contenant au minimum une liaison dynamique ont été générés. Au total 408 fragmentations ont été générées, cela comprend les séquences d'atomes (IA), de liaisons (IB), d'atomes et liaisons (IAB), de paires d'atomes (IAP) ainsi que les atomes unis (IIIA, IIIB, IIIAB, IIIAP). La longueur minimale des fragments générés a été variée de 3 à 10 pour les séquences, et de 3 à 6 pour les atomes unis. La longueur maximale des fragments générés a été variée de 3 à 15 pour les séquences, et de 3 à 10 pour les atomes unis.

Les descripteurs MOE appartenant à la classe 2D ont aussi été calculés. Les descripteurs de chiralité ont été omis étant donné que seule une partie des structures du jeu d'entraînement possédait une information sur la répartition des différents groupements d'une molécule dans l'espace et aucune dans le jeu de test. Le calcul des descripteurs MOE à partir des GCRs revient en fait à calculer, selon le cas, les descripteurs du métabolite final formé, ou d'une molécule intermédiaire d'une réaction d'oxydation comme expliqué plus haut.

Un mélange des descripteurs MOE (vision globale du métabolite) avec les descripteurs ISIDA (vision localisée sur le voisinage proche du site oxydé) est aussi entrepris afin d'observer une éventuelle synergie qui pourrait améliorer les performances.

5.3.2.4. Machines d'apprentissages.

Etant donné que la machine d'apprentissage SVM permettait d'obtenir les performances les plus élevées dans le cas de la prédiction des sites d'oxydation pour les substrats de l'isoenzyme CYP1A2 (voir 3.), cette même méthode a été sélectionnée pour le cas présent. La construction des modèles SVM a été effectuée grâce au logiciel LIBSVM et le coefficient de similarité de Tanimoto a été utilisé comme noyau. Le coût de la SVM a été optimisé en 5x5-CV afin d'obtenir les meilleures performances pour chaque jeu de descripteurs. Le coût a été varié de 1 à 401 avec un pas de 50 dans un premier temps et avec un pas de 10 dans un second temps pour les meilleurs jeux de descripteurs. Le poids des objets appartenant à la classe des transformations observées a été fixé à 6,7 étant donné que notre jeu d'entraînement possède 6,7 fois plus de transformations non observées expérimentalement que de transformations observées. Afin d'avoir des modèles probabilistes pour comparer notre approche à celle de SMARTCyp les modèles SVM ont été reconstruits sur les meilleurs jeux de descripteurs en sélectionnant l'option permettant d'avoir des probabilités en sortie pour chaque site d'oxydation potentiel. Les performances de ces derniers modèles ont été estimées à l'aide du critère IAP en 5x5-CV et sur le test externe.

5.3.2.5. Validation des modèles.

Les performances prédictives des modèles ont été estimées en 5x5-CV et la précision balancée moyenne a été calculée. Il a été fait attention à ce que les oxydations effectuées pour un même substrat aient toutes été mises dans le même paquet lors de la validation croisée. Un protocole de scrambling a été réalisé pour les meilleurs modèles afin de s'assurer que les modèles construits ne sont pas du à une corrélation fortuite entre les descripteurs et la propriété prédite. Finalement un modèle consensus a été construit et les meilleurs modèles construits ont été appliqués sur le jeu de test externe. La stratégie de modélisation est résumée Figure 5-13.

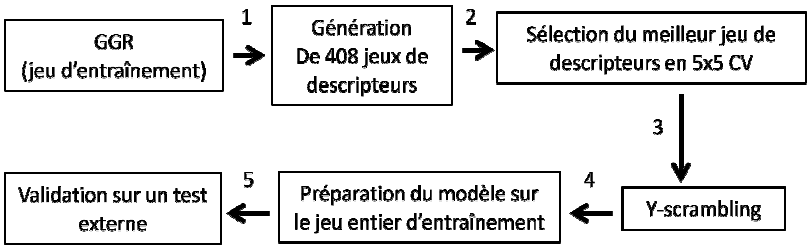


Figure 5-13. Stratégie de modélisation.

1. 408 jeux de descripteurs sont générés à partir des 9326 GCRs correspondant aux transformations observées et non observées.
2. Une procédure de validation croisée à 5 paquets répétée 5 fois après mélange aléatoire du jeu initial est mise en place afin de sélectionner les jeux de descripteurs permettant de construire les meilleurs modèles.
3. Une procédure de scrambling est mise en place pour la construction de modèles basés sur les descripteurs sélectionnés. Seuls les descripteurs n'entrainant pas de modèles fortuits sont conservés.
4. Des modèles basés sur les jeux de descripteurs sélectionnés sont construit sur le jeu d'entraînement complet.
5. Les modèles construits sont appliqués sur le jeu de test externe composé des GCRs de 733 biotransformations potentielles dont 85 expérimentalement observées.

5.3.2.6. Etude comparative

SMARTCyp est une méthode qui prédit les sites d'oxydation d'une molécule dont la structure 2D est connue. L'algorithme de prédiction utilise deux descripteurs : un descripteur de réactivité et un descripteur d'accessibilité. La probabilité d'un site d'être oxydé est donnée par la formule S=E-8A. E est le descripteur de réactivité et correspond à l'énergie requise par l'isoforme CYP3A4 pour réagir à une position donnée, c'est-à-dire l'énergie d'activation de la réaction d'oxydation. Cette énergie est prise dans une table de valeurs d'énergies d'activation précalculées à l'aide de calculs quantiques. Le descripteur d'accessibilité est un nombre variant entre 0,5 et 1 mesurant la distance d'un point de vue topologique entre l'atome considéré et le centre de la molécule. « 1 » signifie que l'atome considéré est à l'extrémité de la molécule, « 0,5 » signifie que l'atome considéré est au centre de la molécule. S correspond au score, plus S est faible plus la probabilité de l'atome considéré d'être oxydé est élevée.

La publication proposant les résultats pour SMARTCyp ne donne que les 3 sites possédant la plus grande probabilité d'être oxydé. Dans un souci de comparaison nous en feront de même et regarderont successivement parmi les 3 sites les plus probables combien correspondent véritablement à des sites d'oxydation.

5.3.3. Résultats

5.3.3.1. Validation croisée

Les modèles SVM ont été construits dans un premier temps en validation croisée à 5 paquets. Pour chaque jeu de descripteurs cette procédure a été répétée 5 fois et les valeurs statistiques obtenues ont été moyennées sur l'ensemble des validations croisées. Les résultats sont donnés Tableau 5-5. Le meilleur modèle obtenu a été construit en utilisant des fragments sous-structuraux contenant des séquences d'atomes et liaisons allant d'une longueur de 3 à 7 atomes. Etant donné que chaque fragment doit contenir au minimum une liaison dynamique, ces séquences décrivent jusqu'à 5 sphères de coordination autour de l'atome oxydé.

Tableau 5-5. Performances en validation croisée 5-CV des modèles SVM pour les meilleurs jeux de descripteurs de chaque type de descripteur.

Descripteurs	Précision[a]	Rappel[a]	Précision[b]	Rappel[b]	P.B.
IAB(3-7)	0,96	0,835	0,409	0,767	0,801
MOE-2D	0,929	0,711	0,496	0,636	0,673
IAB(3-7) + MOE-2D	0,951	0,878	0,459	0,695	0,786

[a]. sites d'oxydation non observés.
[b] sites d'oxydation observés expérimentalement.

On aperçoit d'après les résultats du tableau 1-1 que les modèles construits sur les descripteurs MOE seuls obtiennent les performances les plus faibles (PB=0,673). Au contraire les modèles construits sur les descripteurs ISIDA uniquement obtiennent les meilleures performances (PB=0,801). Le mélange des descripteurs ISIDA et MOE ne permet pas d'améliorer les performances des descripteurs ISIDA seuls. Les descripteurs plus globaux comme MOE ont donc tendance à détériorer la qualité des modèles construits.

En s'intéressant maintenant plus précisément aux modèles construits sur les descripteurs ISIDA seuls, on peut voir que la précision des sites d'oxydation observés expérimentalement paraît assez faible pour la validation croisée. Cela peut être facilement expliqué par le fait qu'il y'a une forte disproportion entre les deux classes présentes dans notre jeu de données. Certes le nombre de sites d'oxydation avérés a été pondéré lors de la construction des modèles SVM, cela a permis d'améliorer le rappel de ces sites mais en contrepartie de nombreux sites d'oxydation non observés ont aussi été prédits comme expérimentalement observés. Toutefois cette erreur est préférable car, en pratique, les métabolites expérimentalement observés dépendent largement des techniques utilisées, on ne pourra donc pas affirmer qu'un métabolite qui n'a pas été observé n'existe véritablement pas. Ainsi la précision balancée mesurant le succès global de la modélisation est tout de même élevé (P.B.=0,801).

Une approche consensus a aussi été testée. Les consensus ont été construits en faisant varier le nombre de modèles de façon à avoir un nombre impair de modèles composant le consensus. Leurs performances ont été estimées à l'aide de 5 validations croisées à 5 paquets. Un graphique récapitulant la précision balancée moyenne calculée sur les 5 CV en fonction du nombre de modèles inclus dans le consensus est donné Figure 5-14.

On voit sur le graphique que le nombre de modèles composant le consensus ne permet pas d'obtenir de meilleures performances que les meilleurs modèles individuels. Les performances ont même tendance à être légèrement moins bonnes. Nous choisissons donc de conserver les modèles construits sur les séquences d'atomes et liaisons de longueur 3 à 7 pour faire des prédictions externes.

CYP3A4

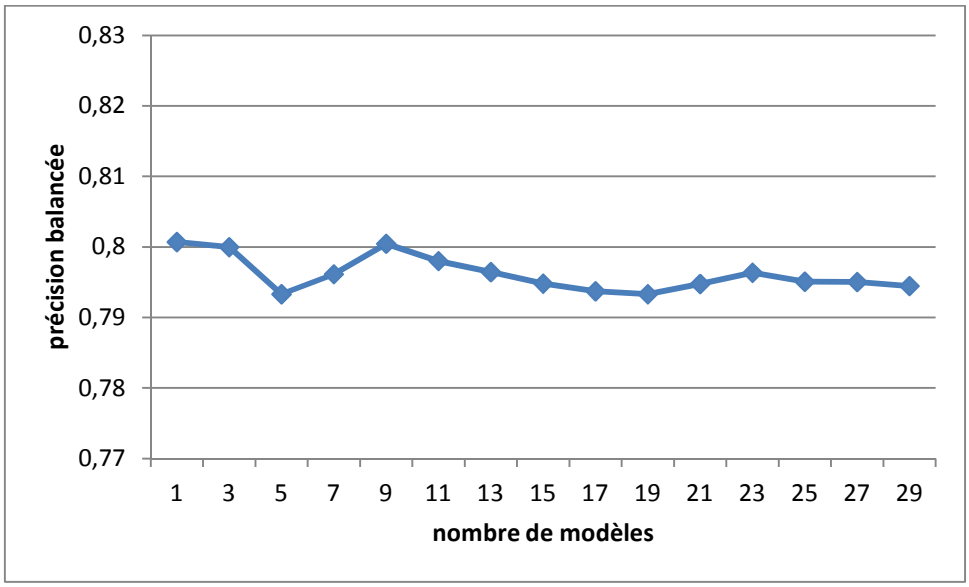


Figure 5-14. Performances des modèles consensus SVM moyennées sur les 5 CV et dépendant du nombre de modèles composant le consensus. Les meilleurs modèles individuels sont choisis pour entrer dans le consensus.

Finalement des modèles SVM proposant des prédictions sous forme de probabilités ont été construits pour le meilleur jeu de descripteurs. Le critère IAP a été moyenné pour les 5 validations croisées. On observe une haute valeur pour cette statistique (Tableau 5-6). Cela montre clairement que nos modèles assignent une probabilité plus élevée pour un site d'oxydation avéré que pour un site non réactif.

5.3.3.2. Y-scrambling

Les performances en scrambling des modèles construits sur les séquences d'atomes et liaisons de longueur allant de 3 à 7 atomes ont été obtenues à l'aide de 20 validations croisées à 5 paquets. Pour chaque validation croisée la propriété a été aléatoirement mélangée dans le jeu de données. Au total 100 valeurs de précision balancée ont été obtenues. Rappelons que les performances pour un modèle qui serait strictement dû au hasard doit avoir une précision balancée de 0,5. Les précisions balancées obtenues pour chaque paquet varient entre 0,466 et 0,540. Etant donné que la statistique est beaucoup plus importante sur le jeu de données originale (BA=0,801) on peut affirmer que les modèles obtenus en validation croisée ne sont pas dus à une corrélation fortuite des descripteurs avec la propriété.

5.3.3.3. Validation externe et étude comparative

Un modèle SVM a alors été construit sur le jeu d'entraînement entier en utilisant le jeu de descripteurs ainsi que les paramètres ayant menés aux meilleures performances. Les performances sur le test set sont données Tableau 1-2. La valeur IAP pour les modèles SVM probabilistes construit sur le même jeu de descripteur est aussi donnée Tableau 5-6. En comparant les valeurs P.B. et IAP obtenues en validation croisée avec celles obtenues sur le test externe on s'aperçoit que celles-ci sont très légèrement inférieures. Cette faible différence atteste de la robustesse de notre modèle à prédire correctement les sites d'oxydation de substrats de l'isoenzyme CYP3A4 chez l'homme.

Tableau 5-6. Performances des modèles SVM pour le meilleur jeu de descripteurs IAB(3-7) en validation croisée et pour le test externe.

	Précision[a]	Rappel[a]	Précision[b]	Rappel[b]	P.B.	IAP
Training set (5-CV)	0,960	0,835	0,409	0,767	0,801	0,877
Jeu de test	0,957	0,850	0,382	0,706	0,778	0,842

[a]. sites d'oxydation non observés
[b]. sites d'oxydation observés expérimentalement

Les performances prédictives de notre modèle ont été comparées à celles de la méthode SMARTCyp ne considérant que les 3 prédictions les plus probables pour chaque substrat. Dans un premier temps nous regarderons pour combien de substrats nous retrouvons un site d'oxydation avéré pour la prédiction la plus probable (top 1), les deux prédictions les plus probables (top 2), et les trois prédictions les plus probables (top 3) (Tableau 5-7). Pour le top 1 les performances de SMARTCyp sont légèrement supérieures aux nôtres. Pour top 2 et top 3 notre modèle permet d'avoir de meilleures performances. Il est intéressant de noter qu'une solution consensus de notre modèle avec la méthode SMARTCyp (top 1 ISIDA + top 1 SMARTCyp) permettrait d'identifier un site d'oxydation dans 43 substrats, ce qui est plus important que de considérer le top 2 pour n'importe laquelle des méthodes.

Tableau 5-7. Résultats sur le test externe du meilleur modèle SVM et de SMARTCyp. Pour top 1, top 2 et top 3 on compte respectivement le nombre de substrats pour lesquels le site le plus probable, les deux sites les plus probables et les 3 sites les plus probables contiennent au moins un site d'oxydation.

	ISIDA	SMARTCyp
Top 1	29	32
Top 2	40	35
Top 3	45	37

Une autre façon de comparer notre modèle avec SMARTCyp est de compter le nombre de sites d'oxydation retrouvés en top 2 et top 3 par rapport au nombre total de sites d'oxydation qu'il serait possible de retrouver pour chaque substrat si toutes les prédictions étaient correctes. Le jeu de test compte 16 substrats pour lesquels il existe plus d'un site d'oxydation. Donc si toutes les prédictions faites étaient correctes on devrait retrouver 72 sites de métabolisme en répertoriant les deux prédictions les plus probables pour chaque substrat. Comme on peut voir dans le Tableau 5-8 le modèle ISIDA retrouve beaucoup plus de sites d'oxydation que SMARTCyp pour les prédictions top 2 et top 3. En effet pour les prédictions top 3 notre modèle retrouve environ 73% de tous les sites d'oxydation qu'il est possible de trouver alors que SMARTCyp n'en retrouve que 60%.

Tableau 5-8. Résultats sur le test externe du meilleur modèle SVM et de SMARTCyp. Pour top 1, top 2 et top 3 on compte respectivement le nombre de sites d'oxydation retrouvés parmi la prédiction la plus probable, les deux prédictions les plus probables et les 3 prédictions les plus probables pour un substrat. Max correspond au nombre maximum de sites d'oxydation qu'il est possible de retrouver si toutes les prédictions sont correctes.

	ISIDA	SMARTCyp	Max
Top 1	29	32	56
Top 2	47	39	72
Top 3	59	48	81

Sur la Figure 5-15 sont donnés des exemples de structures pour lesquelles les sites d'oxydation ont été correctement prédits.

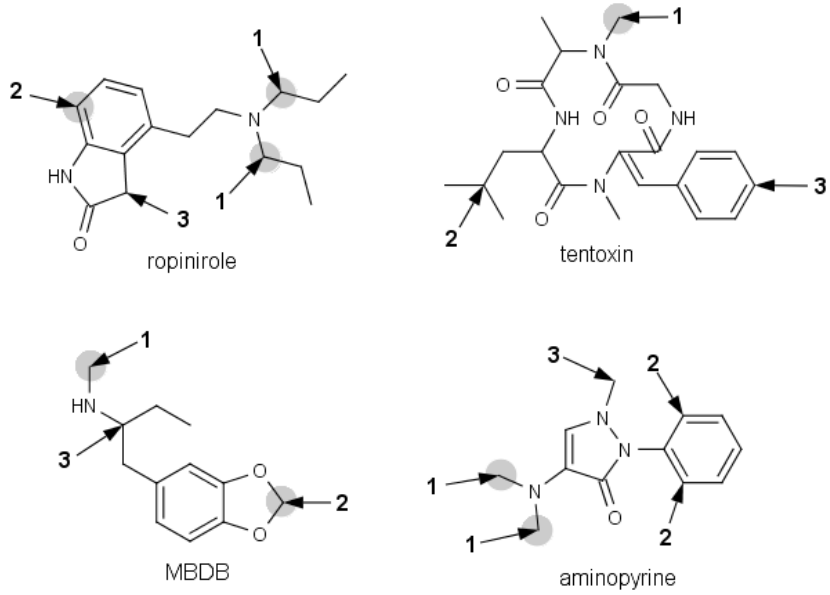


Figure 5-15. Exemples de structures du test set correctement prédites avec le modèle ISIDA.
Les cercles gris correspondent aux sites d'oxydation expérimentalement observés. Les nombres correspondent aux trois sites prédits comme les plus probables.

Il faut noter que pour 11 molécules le modèle ISIDA ne retrouve aucun site d'oxydation parmi les 3 prédictions les plus probables. Des exemples représentatifs de tels cas sont donnés Figure 5-16. En analyse complémentaire nous avons décidé de comparer ces molécules à celles du jeu d'entraînement. Pour ce faire le coefficient de Tanimoto a été calculé entre les vecteurs de descripteurs associés aux sites d'oxydation recherchés avec ceux du jeu d'entraînement.

Il a été trouvé que les descripteurs de l'acide lithocholic (test) et de l'androstérone (entraînement) sont identiques. En fait les deux structures sont très similaires mis à part que le site qui est oxydé pour l'acide lithocholic n'entraîne pas la formation de métabolite pour l'androstérone. Il semblerait donc à la vue de ces informations que les différences présentes entre 2 structures, même loin du centre réactionnel, ont une influence sur la réactivité du site. Cependant, comme l'observation de métabolites dépend de la technique utilisée il est aussi possible que le métabolite correspondant de l'androstérone existe mais n'a simplement pas été expérimentalement observé. Nous avons un autre exemple qui va dans cette direction.

Pour les sites oxydés expérimentalement observés de la phencyclidine et de la loratidine, les cas les plus similaires sont respectivement les sites de la ketamine et de l'amprenavir. A nouveau ces sites sont considérés comme non réactifs dans le jeu d'entraînement à l'inverse du jeu de test. L'environnement proche entre les sites similaires de la phencyclidine et de la ketamine est assez ressemblant mis à part la présence du groupement cétone sur la ketamine. Pour la loratidine et l'amprenavir la similarité entre les environnements proches des sites d'oxydation n'est plus aussi évidente. La présence du groupement tetrahydrofurane peut aisément expliquer la différence de réactivité entre les 2 molécules. Il n'y a donc pas véritablement de structures similaires de la loratidine, ce qui explique la mauvaise prédiction.

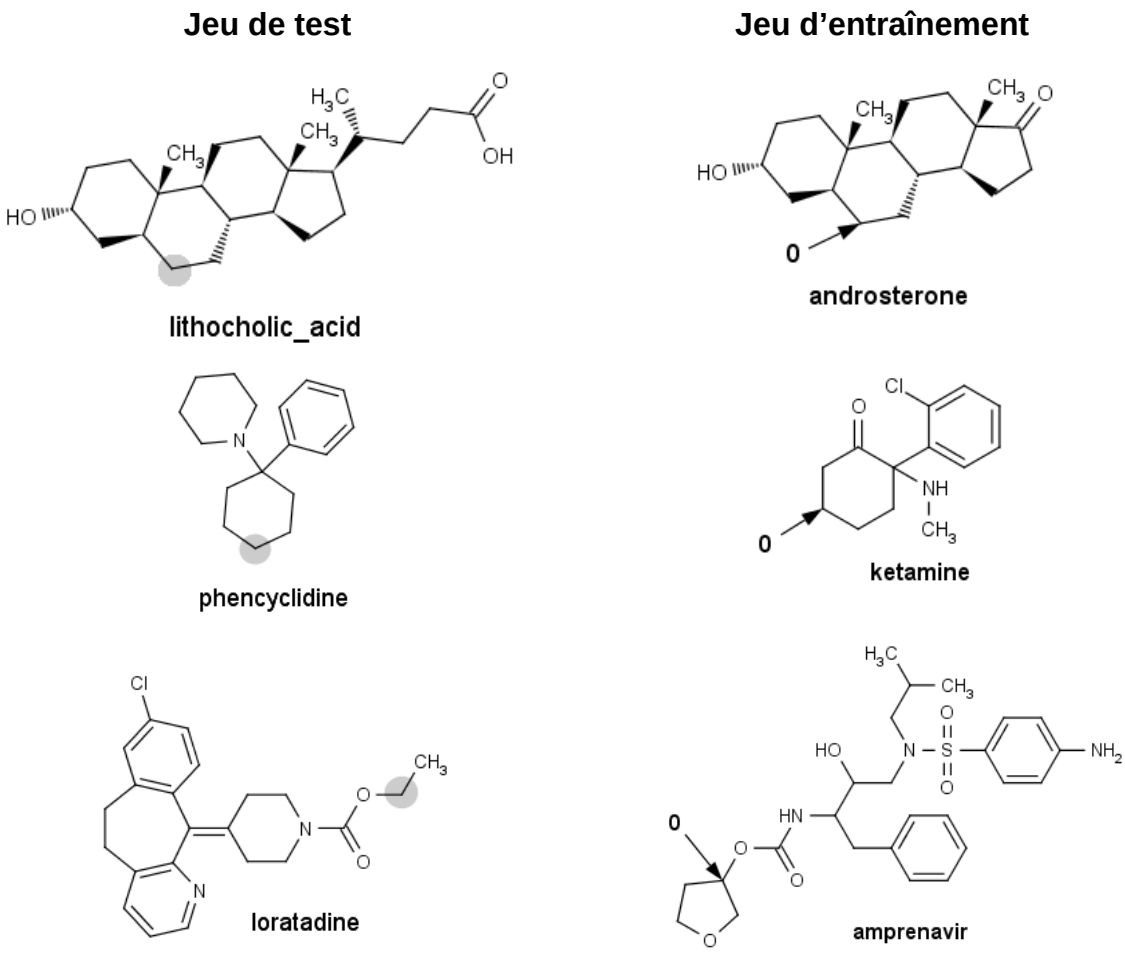


Figure 5-16. Exemples de structures du test set (à gauche) incorrectement prédites avec le modèle ISIDA et leur structure respective la plus similaire dans le jeu d'entraînement (à droite).
Les cercles gris correspondent aux sites d'oxydation expérimentalement observés. « 0 » indique un site similaire non réactif.

5.3.4. Conclusion

Dans cette étude nous avons développés des modèles prédictifs de la régiosélectivité des oxydations induites par le CYP3A4 chez l'homme. Pour ce faire, un jeu de données contenant des biotransformations expérimentalement observées ainsi que des biotransformations non observées a été élaboré. Les descripteurs fragmentaux ISIDA ainsi que les descripteurs MOE 2D ont été générés. Le meilleur modèle SVM a été construit à partir des séquences d'atomes et de liaisons impliquant des longueurs de fragments allant de 3 à 7 atomes. Une précision balancée de 0,78 a pu être obtenue sur le jeu de test dans le cadre de la séparation des biotransformations expérimentalement observées des non observées. Les performances de nos modèles ont été comparées avec celle du logiciel SMARTCyp. Des performances équivalentes ont pu être observées. Dans 80% des cas, au moins un site métabolique a été détecté parmi les 3 sites d'oxydation les plus probables pour une molécule. Ces résultats montrent qu'il y a une nette influence de l'environnement proche du site oxydé dans la prédiction des sites d'oxydation pour d'autres molécules. Très souvent, des environnements similaires entrainent des réactivités similaires.

L'analyse du test set nous a permis d'observer et comprendre les limites de notre modèle. Cependant certaines erreurs viennent aussi du fait que les molécules du jeu de test sont trop dissimilaires par rapport à celles du training set. Dans ce dernier cas aucune garantie n'est faite sur les prédictions effectuées. Pour les molécules très grandes, la description de l'environnement proche du site oxydé peut ne peut pas être suffisante. Les changements structurels lointains du site oxydé peuvent avoir un impact sur l'orientation du substrat complexé. Il faut donc réfléchir au moyen d'améliorer nos descripteurs pour identifier ce type de cas. On pourrait par exemple ajouter des descripteurs qui prennent en compte la taille de la molécule ou encore sa forme. Une étape de docking pourrait aussi être mise en place pour limiter l'application de nos modèles aux zones accessibles par la porphyrine lors de l'oxydation.

5.4. Références

1. Ekins, S., Y. Nikolsky, and T. Nikolskaya, *Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity.* Trends Pharmacol. Sci., 2005. **26**(4): p. 202-209.
2. Crivori, P. and I. Poggesi, *Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs.* Eur. J. Med. Chem., 2006. **41**(7): p. 795-808.
3. Jung, J., et al., *Regioselectivity prediction of CYP1A2-mediated phase I metabolism.* J. Chem. Inf. Model., 2008. **48**(5): p. 1074-1080.
4. Cruciani, G., et al., *MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist.* J. Med. Chem., 2005. **48**(22): p. 6970-6979.
5. Zamora, I., L. Afzelius, and G. Cruciani, *Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450 2C9.* J. Med. Chem., 2003. **46**(12): p. 2313-2324.
6. Borodina, Y., et al., *A new statistical approach to predicting aromatic hydroxylation sites. Comparison with model-based approaches.* J. Chem. Inf. Comput. Sci., 2004. **44**(6): p. 1998-2009.
7. Funatsu, K., K. Hasegawa, and M. Koyama, *Quantitative Prediction of Regioselectivity Toward Cytochrome P450/3A4 Using Machine Learning Approaches.* Mol. Inf., 2010. **29**(3): p. 243 - 249.
8. Singh, S.B., et al., *A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules.* J. Med. Chem., 2003. **46**(8): p. 1330-1336.
9. Haji-Momenian, S., et al., *Comparative molecular field analysis and QSAR on substrates binding to cytochrome p450 2D6.* Bioorg. Med. Chem., 2003. **11**(24): p. 5545-5554.
10. Hennemann, M., et al., *CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory.* Chem. Med. Chem., 2009. **4**(4): p. 657-669.
11. Zheng, M., et al., *Site of metabolism prediction for six biotransformations mediated by cytochromes P450.* Bioinformatics, 2009. **25**(10): p. 1251-1258.

12. Boyer, S., et al., *Reaction site mapping of xenobiotic biotransformations.* J. Chem. Inf. Model., 2007. **47**(2): p. 583-590.
13. Sheridan, R.P., et al., *Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9.* J. Med. Chem., 2007. **50**(14): p. 3173-3184.
14. Guengerich, F.P., *CYTOCHROME P-450 3A4: Regulation and Role in Drug Metabolism.* Annu. Rev. Pharmacol. Toxicol., 1999. **39**(1): p. 1-17.
15. *Standardizer was used for structure canonicalization and transformation, JChem 5.3.5, 2010, ChemAxon (http://www.chemaxon.com)*.
16. Sansen, S., et al., *Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2.* J. Biol. Chem., 2007. **282**(19): p. 14348-14355.
17. Rydberg, P., et al., *SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism.* ACS Med. Chem. Lett., 2010. **1**(3): p. 96-100.
18. *Accelrys, Inc., Accelrys Metabolite, San Diego, 2009 (http://accelrys.com)*.
19. Guengerich, F.P., *Common and Uncommon Cytochrome P450 Reactions Related to Metabolism and Chemical Toxicity.* Chem. Res. Toxicol., 2001. **14**(6): p. 611-650.

TROISIEME PARTIE : Développement de modèles QSAR pour la prédiction de la toxicité.

6. Prédiction de la mutagénicité liée au test biologique d'Ames.

6.1. Introduction

Un composé « mutagène » est un composé capable d'induire des mutations. Cet effet peut résulter de plusieurs mécanismes différents. En effet, un composé réactif envers l'ADN peut établir une liaison covalente avec celui-ci ou encore provoquer des suppressions de bases, ce qui a pour effet de distordre la structure de l'ADN. De même des composés non réactifs peuvent devenir réactifs après métabolisation. Les distorsions de l'ADN peuvent aussi provenir de l'intercalation de composés aromatiques polycycliques entre des paires de base de l'ADN stabilisé par des interactions de π-stacking. Les distorsions de l'ADN empêchent la réplication correcte de celui-ci et des mutations peuvent alors intervenir.

De nos jours l'identification de telles substances est devenue une priorité, en particulier dans le milieu pharmaceutique [1]. En effet, les problèmes sanitaires que ces produits chimiques peuvent potentiellement induire sont redoutés (problèmes d'infertilités, cancers ou mutations transmissibles aux générations futures). Ceci explique pourquoi le test d'Ames [2, 3], dédié à la détection de molécules mutagènes, est utilisé mondialement.

Ce test expérimental est cependant trop long (48 heures) et trop coûteux pour être utilisé à grande échelle. Par ailleurs, le développement de nouveaux composés doit se faire sous la contrainte d'être non-mutagène. La prise en compte de cette propriété doit donc être introduite très en amont du processus de recherche et développement. C'est pourquoi la prédiction de la mutagénicité par des modèles informatiques a attiré une telle attention.

Il a de plus déjà depuis longtemps été établi dans la littérature que la propriété mutagène d'un composé est fortement liée à sa structure chimique [4, 5], et de nombreuses sous-structures présentes principalement dans les composés mutagènes, et appelés toxicophores, ont pu être détectées. Plus récemment, Bailey et al. [6] identifiaient une liste de 33 alertes structurales pendant que Kazius et al. [7] en listaient 29 permettant de retrouver les composés mutagènes avec une forte précision. De même, de nombreux systèmes experts basés sur les descripteurs sous-structuraux émettant des règles ou proposant des modèles probabilistes ont été

commercialisés [8]. En 2007, Mazzatorta et al. [9] ont proposé un modèle QSAR atteignant 0,788 de précision balancée. En combinant les prédictions de leur modèle aux alertes structurales de Kazius et al. [7] ils ont même pu atteindre une précision balancée de 0,852. L'ensemble de ces méthodes (dont la liste est non exhaustive) a été discuté dans une revue publique [10]. Il faut noter cependant que le terme de toxicophore et d'alerte structural n'est pas dédié seulement à la mutagénicité, mais cette pratique est utilisée dans les prédictions de nombreuses toxicités [11, 12].

Récemment un concours mondial a ainsi été proposé afin de trouver des modèles visant à établir un protocole de criblage virtuel dans le but de limiter l'usage du test expérimental. La modélisation a été faite sur un jeu d'entraînement de 3439 molécules dont 1920 mutagènes et un jeu de test de 2131 molécules dont l'activité n'a pas été communiquée. Les descripteurs fragmentaux ISIDA ont été générés et des modèles consensus SVM, VP, NB et SQS ont été construits, ainsi qu'un modèle consensus global regroupant les modèles consensus de chaque méthode. Les performances des modèles obtenus par des méthodes individuelles atteignent 0,80 de précision balancée pour le jeu d'entraînement en validation croisée. Le modèle global quant à lui atteint 0,81 de PB. Les résultats du challenge ont démontré que nos modèles étaient au moins aussi efficaces que le meilleur modèle obtenu dans ce concours. En effet, les performances des modèles consensus varient entre 0,75 et 0,81 de PB et le modèle global atteint 0,83 de PB sur le jeu de test ce qui est le meilleur résultat obtenu lors de ce concours.

Une étude benchmark entre les 29 modèles des 12 groupes ayant participé à ce projet international est proposée dans l'article publié [13] en octobre 2010 dans *Journal of Chemical Information and Modeling* et disponible en annexe 10.3. Le but de cet article étant de comparer les performances de plusieurs domaines d'applicabilités à l'aide de tous les modèles construits lors de ce concours.

Etant donné le petit paragraphe dédié aux méthodes utilisées dans l'article, nous détaillerons ici plus en détail les stratégies de modélisation employées ainsi que nos résultats.

6.1.1. Le test D'Ames

Ce test populaire [2, 3] est un des plus simples à mettre en œuvre dans la prédiction de la toxicité. Il permet plus particulièrement de détecter si une substance choisie est mutagène ou non. Pour ce faire, on utilise des souches bactériennes (*Salmonella typhimurium*) possédant au préalable des mutations ne leur permettant plus de synthétiser un certain acide aminé, l'histidine, essentiel à leur multiplication. Il suffit de mettre la substance que l'on souhaite tester en contact avec les différentes souches bactériennes. De nouvelles mutations peuvent alors intervenir et restaurer la fonction du gène préalablement muté. Ainsi la bactérie peut à nouveau synthétiser l'histidine. Ces bactéries sont alors capables de se développer dans un milieu dépourvu d'histidine suite aux mutations perpétrées par la substance testée.

Cependant la stratégie telle qu'elle ne suffit pas toujours à détecter les substances mutagènes chez l'homme, car certaines molécules ne sont mutagènes qu'une fois qu'elles ont été métabolisées. Pour mimer la digestion humaine, on rajoute dans les différents tests un homogénat de foie de rat (S9), ainsi que des cofacteurs (Mix). Ce mélange possède les différentes enzymes contenues dans le foie ainsi que les cofacteurs nécessaires intervenant dans la détoxification des molécules passant dans l'organisme humain. La mise en place de ce test est représentée sur la Figure 6-1.

Le test est considéré comme positif si on observe une croissance significative d'une colonie dans au moins une souche, en présence ou en absence de S9-Mix. De même un test est négatif, si aucune croissance d'aucune colonie n'est observée dans aucunes souches.

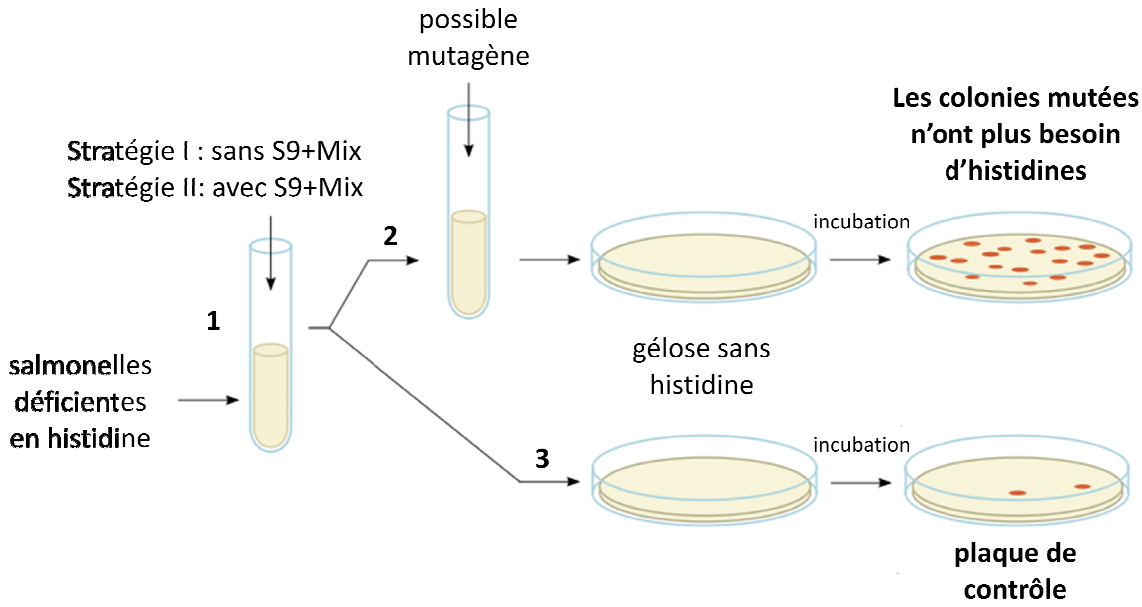


Figure 6-1 : Présentation d'un test d'Ames pour une substance mutagène. (Image adaptée de Wikipédia)

1. Une solution contenant des salmonelles déficientes en histidine est préparée. Chaque stratégie est alors explorée. Dans le premier cas on garde la solution telle quelle (stratégie I). Dans le second cas on ajoute à la solution un homogénat de foie de rat (stratégie II).
2. On ajoute à la solution le composé à tester, on transvase sur une plaque de gélose et on incube.
3. On transvase directement le contenu de la solution sur une plaque de gélose et on incube. Ce test sert de plaque de contrôle pour vérifier que la solution et l'environnement ne sont pas contaminés.

Si des colonies apparaissent sur au moins une plaque, quelque soit la stratégie employée, et que les plaques de contrôle présentent des résultats négatifs, alors le composé testé est mutagène.

6.2. Matériel et méthodes

6.2.1. Données et nettoyage

Un jeu d'apprentissage contenant 4361 molécules dont 2344 mutagènes a été envoyé par les organisateurs à chaque participant du concours. De même un jeu de test de 2131 molécules pour lesquelles la mutagénicité n'a pas été communiquée a été transmis.

Les données étant issues de la littérature, un premier nettoyage des données a été entrepris pas les organisateurs. Cependant il est apparu que le jeu de données en question n'a pas été suffisamment bien nettoyé pour satisfaire les conditions de création de modèles, et de ce fait, il comportait des entrées suspectes. Une nouvelle

correction des données a donc été entreprise pour le jeu d'entraînement ainsi que le jeu de test.

i) Filtrage automatique

Pour la 1ère étape du nettoyage, le programme Filter de la suite logicielle OpenEye [14] a été utilisé. Pour ce faire, un filtre déjà existant a été modifié afin d'éliminer :

- tous les hydrocarbures (alcanes, alcènes, benzènes), car étant insoluble en milieu aqueux, il est difficile de comprendre comment ce type de composé a pu être testé dans un environnement cellulaire.
- les molécules de faible poids moléculaire ou ne contenant pas assez d'atomes, car trop peu spécifiques. En effet un petit changement au niveau de la structure de tels composés peut entraîner un résultat différent pour la propriété donnée.
- les molécules à fort poids moléculaire ou contenant trop d'atomes, car ce sont des composés beaucoup trop spécifiques et donc une généralisation de leur comportement n'est pas possible. De plus il n'est pas aisé de générer des descripteurs pour de grosses molécules.
- les peptides, car la structure tridimensionnelle joue un rôle très important au sein de ces composés et, à priori, la structure repliée qu'adoptera la molécule en milieu aqueux n'est pas connue. Ces composés sont impossibles à traiter dans ce cas.

ii) Filtrage manuel

Ceci fait, il a fallu ensuite trier le reste des molécules à la main. Les molécules pour lesquelles les centres stéréochimiques ont été mal définis, (celles pour lesquelles les atomes n'étaient pas explicitement orientés dans l'espace), ont été éliminées (exemple Figure 6-2). Effectivement, il serait risqué de considérer que tous les stéréoisomères d'une molécule aient tous la même mutagénicité. Etant donnée cette information manquante, il est préférable d'écarter ces molécules du jeu de données. Pour ce faire, il suffit de trouver une sous structure commune pour les molécules possédant un centre stéréochimique mal défini, puis de rechercher toutes les structures semblables à l'aide d'une recherche sous-structurale dans un logiciel

qui permet d'interroger une base de données chimique, et de vérifier leur stéréochimie.

Le jeu de données a ensuite été standardisé à l'aide du logiciel Standardizer de ChemAxon et de la routine « Clear Stereo, Mesomerize, Neutralize, Remove Explicit Hydrogens ». Les doublons présents dans la base de structures moléculaires ont été répertoriés tel que :

Si deux structures identiques présentent la même mutagénicité, une des deux molécules est éliminée.

Si deux structures identiques présentent une mutagénicité différente, les deux molécules sont éliminées.

Les doublons dus à un état de protonation différent, à une forme mésomère différente, ou encore à une forme tautomère différente ont ainsi pu être éliminés. Dans le cas du jeu de test, étant donné que les activités n'étaient pas connues, seule une structure par doublon a été conservée.

Finalement 3439 et 1738 molécules ont été respectivement conservées pour le jeu qui a servi à construire les modèles de classification (jeu d'entraînement) et pour le jeu sur lequel ont été appliqués ces modèles afin de faire des prédictions (jeu de test). Le jeu d'entraînement contient 1915 molécules mutagènes et 1514 non mutagènes, et le jeu de test contient 960 mutagènes et 778 non mutagènes. Bien entendu, au moment du test aveugle la mutagénicité des molécules du test n'était pas connus ; cette information est donc rétrospective.

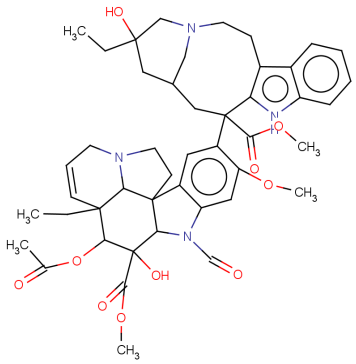


Figure 6-2. Exemple d'une molécule dont les centres stériques sont mal définis.

6.2.2. Machines d'apprentissages et descripteurs

Les descripteurs ISIDA SMF ont été sélectionnés pour coder numériquement les structures. Le jeu de données a subit au total 181 fragmentations :

- pour IA, IAB et IA-AP : 2 ≥ Nmin ≥ 10 et 2 ≥ Nmax ≥ 10
- pour IIIA, IIIAB et IIIA-AP : 2 ≥ Nmin ≥ 6 et 2 ≥ Nmax ≥ 6
- IIHy.

Plusieurs méthodes d'apprentissage ont été utilisées durant ce projet afin de différencier les composés mutagènes des composés non mutagènes : la machine à vecteurs support, le bayésien naïf et le perceptron votant.

Les modèles SVM ont été construits à l'aide du logiciel LIBSVM et d'un noyau de Tanimoto. Le coût de la SVM a été choisi de façon à ce que tous les exemples du jeu d'entraînement soient correctement classés. Pour cela, le coût minimum permettant une séparation complète des données a été déterminé pour plusieurs fragmentations et le coût minimum entre tous a été choisi.

L'apprentissage du perceptron votant s'arrête une fois la convergence atteinte. Etant donné que celui-ci ne converge pas toujours, un critère d'arrêt a été défini. A la 100^{e} époque l'apprentissage stoppe car il a été remarqué que passé ce stade les performances du perceptron n'évoluent plus.

Le logiciel ISIDA/Bayes développé au laboratoire a été utilisé pour construire les modèles de bayésien naïf.

Un modèle consensus a été construit pour chaque méthode d'apprentissage en sélectionnant les meilleurs modèles à l'aide de la distribution statistique de la précision balancée pour la SVM, VP, et NB. Les modèles se situant au-delà du pic maximum apparent de la distribution sont sélectionnés pour constituer le modèle consensus. Un modèle consensus global a aussi été construit. Il correspond à la prédiction moyenne des différents modèles consensus. Le modèle SQS construit par le Dr. Dragos Horvath a été inclus dans le consensus global.

6.2.3. Estimation des performances

Les performances prédictives des modèles construits ont été évaluées en validation croisée à 5 paquets. Cette procédure a été exécutée 5 fois pour chaque

jeu de descripteurs. Soit 25 modèles ont été générés pour chaque fragmentation et pour chaque méthode. Cela correspond au total à 4525 modèles construits pour chaque méthode d'apprentissage. La précision balancée a été choisie comme critère statistique pour la SVM et le VP, et la ROC AUC pour le NB.

Une pixel map a été dessinée pour chaque type de fragmentation et pour chaque méthode, pour permettre une vision plus détaillée des performances obtenues. Une pixel map est en fait une carte 2D où chaque case correspond à une fragmentation, et qui est colorée par une teinte allant du clair (bon score) au sombre (mauvais score) afin de rendre compte de la valeur du paramètre statistique lié à cette fragmentation. Pour la SVM et le perceptron votant, le paramètre tracé est la précision balancée, et pour le bayésien, la valeur ROC AUC.

Le domaine d'applicabilité fragment control a été utilisé afin d'estimer l'impact de son utilisation sur les performances des modèles consensus.

Les prédictions externes ont finalement été faites sur le jeu aveugle et, après envoi de nos prédictions, les labels mutagènes/non mutagènes nous ont été transmis pour le jeu de test externe.

Afin de pousser l'analyse du modèle consensus global, différents seuils d'acceptation des prédictions faites ont été définis, puis les performances ont été recalculées. Les prédictions prenant des valeurs comprises entre 0 (non mutagène) et 1 (mutagène), 0,5 étant la valeur correspondant à une indétermination, nous regarderons successivement la précision des prédictions pour les molécules possédant une probabilité $Y \leq 0,5 - X$ ou $Y \geq 0,5 + X$, X étant un nombre pouvant prendre les valeurs 0 ; 0,1 ; 0,2 ; 0,3 et 0,4.

6.3. Résultats

6.3.1. Validation croisée

6.3.1.1. Modèles individuels

Pour chaque méthode d'apprentissage 4525 modèles ont donc été construits étant donné que la procédure de validation croisée à 5 paquets a été répétée 5 fois pour chaque fragmentation. Les performances des meilleurs modèles pour chaque

méthode sont données Tableau 6-1. Les performances atteintes par les meilleurs modèles des 3 méthodes utilisées sont à peu près équivalentes avec toutefois des performances légèrement meilleures pour le perceptron votant. Les performances qualitatives de chaque méthode pour chaque type de fragmentation sont données sous la forme de pixel map.

Tableau 6-1. Performance en 5-CV pour les meilleurs modèles de chaque méthode.

méthode	descripteurs	Précision[a]	Rappel[a]	Précision[b]	Rappel[b]	PB
NB	t10l2u4	0,759	0,752	0,805	0,811	0,782
SVM	t10l2u4	0,740	0,792	0,826	0,780	0,786
VP	t7	0,775	0,780	0,825	0,821	0,800

a : non mutagène
b : mutagène

6.3.1.2. Pixels maps

L'ensemble des pixels maps est donné Figure 6-3.

La SVM et le VP possèdent tout deux un comportement similaire vis-à-vis des différents types de descripteurs utilisés pour la construction de modèles. En effet la tendance qui s'en dégage est que, pour les fragments de type IA,IA-AP et IAB, les descripteurs les plus appropriés sont les descripteurs considérant à la fois les fragments de petites tailles et de grandes tailles (2-6 à 2-10), donc les fragments qui décrivent la molécule de la façon la plus précise possible et qui laissent le moins d'ambigüité. Dès que les molécules sont prédites uniquement à l'aide de petits fragments ou de grands fragments la prédiction se dégrade.

Pour les fragments de type IIIA, IIIA-AP et IIIAB, ce sont les fragments les plus courts qui permettent les meilleurs prédictions. Ceci est en accord avec ce qui a été vu pour les descripteurs de type I puisque lorsque des fragments de type III 2-3 sont générés, les fragments obtenus peuvent déjà atteindre des longueurs de 5, 7 ou 9 atomes selon le nombre de voisin de l'atome central et le nombre de sphères de coordination considérées.

Concernant le NB, le comportement est un peu différent de celui des autres méthodes d'apprentissages. Etonnamment pour le type IA et IAB, les descripteurs permettant d'obtenir les modèles possédant les meilleurs pouvoirs prédictifs sont

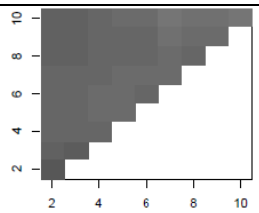
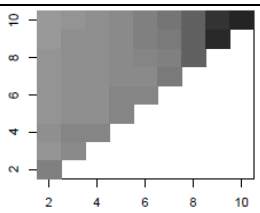
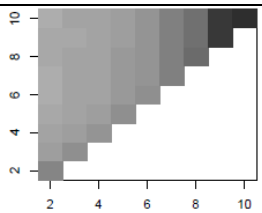
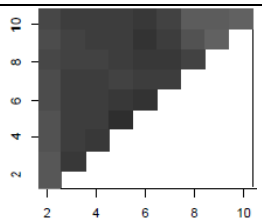
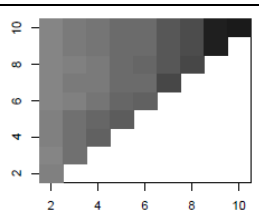
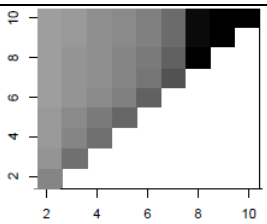
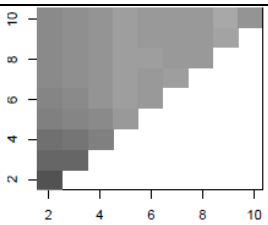
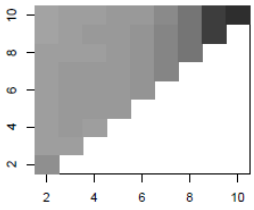
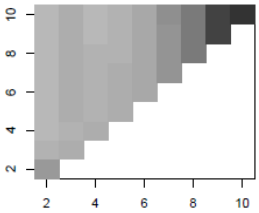
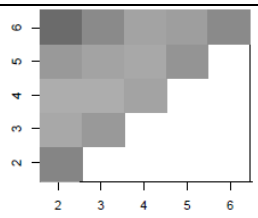
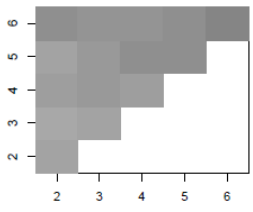
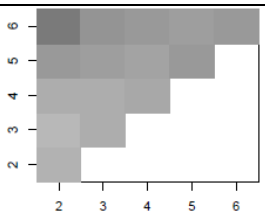
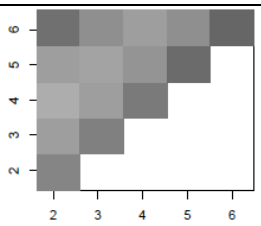
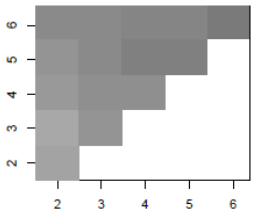
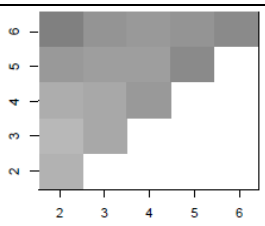
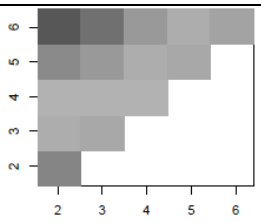
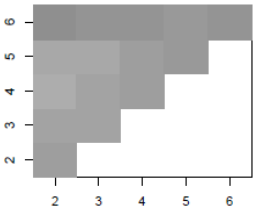
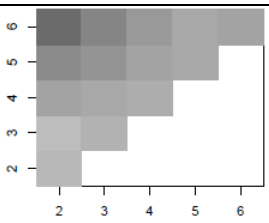
Fragmentation	NB	SVM	VP
IA			
IAP			
IAB			
IIIA			
IIIAP			
IIIAB			

Figure 6-3. Pixels maps pour chaque type de fragmentation et pour chaque méthode. Le paramètre regardé est la précision balancée. Une case sombre correspond à un mauvais score alors qu'une case claire correspond à un bon score.

respectivement ceux de longueurs 5-10 et 9-9. Il s'agit de ceux qui pourtant, n'arrivent pas à décrire les petites molécules. Il y aura beaucoup de molécules qui seront représentées par des vecteurs nuls et l'apprentissage se fera uniquement sur les composés les plus gros. Le succès de ces descripteurs tiendra alors uniquement aux fragments rares (trop spécifiques) pour lesquels les statistiques sont pauvres. Il faut donc se méfier de ces modèles car les résultats seront vraisemblablement instables. De même pour les descripteurs de type IA-AP où les fragments permettant de construire les modèles les plus prédictifs seront ceux de longueur 2-2. Ces fragments étant extrêmement courts les molécules seront décrites de manière très ambiguë et donc il faut aussi se méfier des performances liées à ce type de fragment.

Pour les descripteurs de type IIA, IIA-AP et IIAB, les fragments préférentiels seront ceux combinant les courts et les mi-courts (2-4). Par contre les fragments trop petits ou trop longs prédiront très mal le jeu de test. Ce qui est cohérent avec les résultats issus de la SVM et du VP.

6.3.1.3. Modèles consensus

Pour le consensus bayésien, les 20 meilleurs modèles ont été sélectionnés, en veillant toutefois à ôter les modèles construit à partir des descripteurs III(AB,5-10) et III(AB,9-9) car ils ne décrivent pas une grande partie des molécules (trop petites) et donc n'arriveront pas à prédire correctement les petites molécules du jeu de test. 18 modèles ont donc été sélectionnés. Pour le consensus issus de la SVM, seul les 10 meilleurs modèles ont été sélectionnés, pour le consensus issus du VP, les 15 meilleurs modèles ont été gardés.

Les performances des meilleurs modèles individuels obtenus ainsi que des consensus associés sont résumé dans la Figure 6-4.

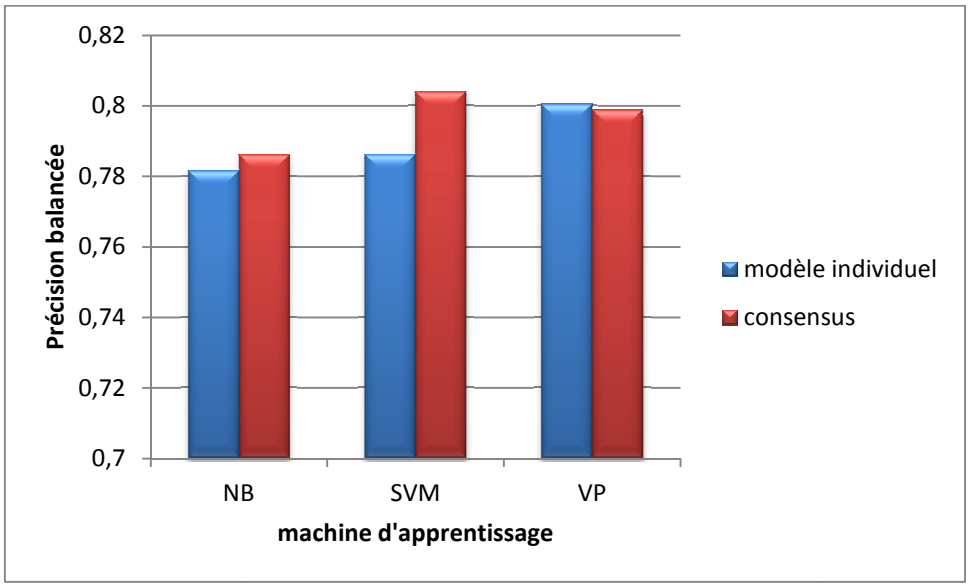


Figure 6-4 : Résultats en 5-CV des meilleurs modèles individuels pour chaque méthode d'apprentissage et des consensus associés.

On s'aperçoit que pour la machine d'apprentissage proposant le modèle individuel le plus performant, la performance du consensus est sensiblement équivalente. Au contraire, pour la SVM le consensus permet d'améliorer les performances comparé au meilleur modèle individuel. Les performances pour le NB sont en légères hausses.

Un consensus global regroupant les consensus de chaque méthode, ainsi que le consensus SQS du Dr. Dragos Horvath, a été construit. Les performances de chaque consensus, ainsi que du consensus global sont détaillées Tableau 6-2.

Tableau 6-2. Performance en 5-CV pour les modèles consensus de chaque méthode.

	Précision[a]	Rappel[a]	Précision[b]	Rappel[b]	PB
NB	0,773	0,779	0,824	0,819	0,799
SVM	0,791	0,769	0,821	0,839	0,804
VP	0,773	0,779	0,824	0,819	0,799
SQS	0,762	0,846	0,867	0,791	0,819
consensus global	0,790	0,793	0,836	0,833	0,813

a : non mutagène
b : mutagène

Le consensus global obtient en 5-CV une meilleure PB que chacun des nos 3 consensus. Cependant le consensus SQS atteint une PB légèrement meilleure que le consensus global.

6.3.2. Validation externe

6.3.2.1. Modèles individuels

Les meilleurs modèles individuels ont été validés sur un test externe de 1738 molécules dont 960 mutagènes. Les performances pour chaque méthode sont résumées dans le Tableau 6-3.

Tableau 6-3. Performance sur le jeu externe des meilleurs modèles individuels de chaque méthode.

	Précision[a]	**Rappel[a]**	**Précision[b]**	**Rappel[b]**	**PB**
NB	0,794	0,665	0,760	0,860	0,762
SVM	0,746	0,788	0,820	0,782	0,785
VP	0,771	0,788	0,825	0,810	0,799

a : non mutagène
b : mutagène

Pour une meilleure visualisation des performances, et afin de comparer les résultats entre la validation croisée et le jeu externe, un diagramme de la précision balancée est donné Figure 6-5 pour les meilleurs modèles individuels.

Pour la SVM et le VP les performances sont équivalentes entre la 5-CV et la validation externe. Pour NB c'est un peu moins bon sur le jeu de test externe, toutefois l'écart avec la 5-CV (≈0,02) est tout à fait raisonnable. On peut donc conclure que les performances obtenues sur le jeu de test sont celles qui étaient attendues.

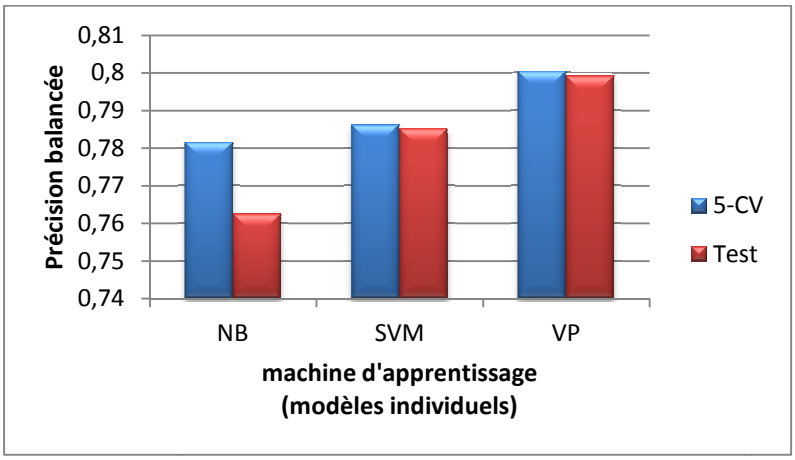


Figure 6-5 : Résultats en 5-CV et sur le test externe des meilleurs modèles individuels pour chaque méthode d'apprentissage.

6.3.2.2. Modèles consensus

Comme pour les meilleurs modèles individuels, un diagramme confrontant la précision balancée des consensus de chaque méthode pour la 5-CV et la validation externe est proposée Figure 6-6 ainsi qu'un tableau résumant les performances plus détaillées de chaque consensus sur le jeu de test Tableau 6-4.

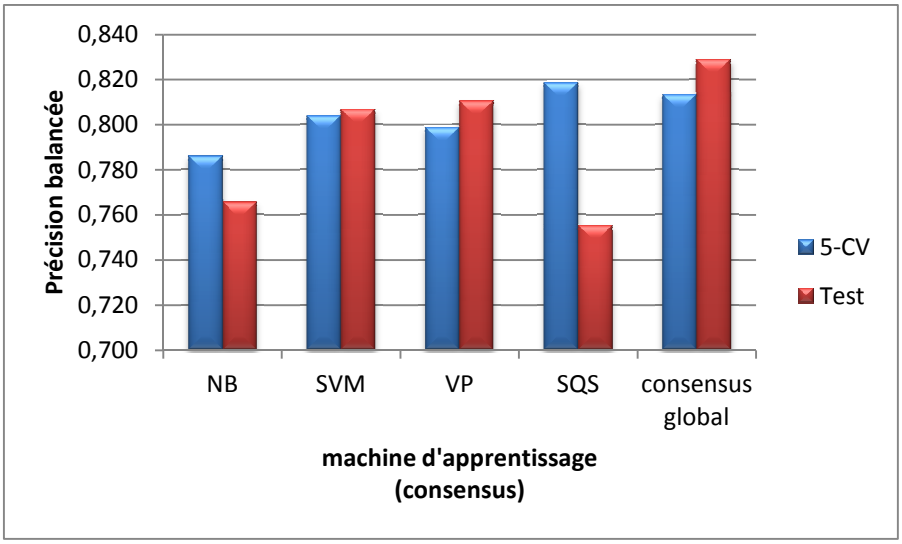


Figure 6-6: Résultats en 5-CV et sur le test externe des modèles consensus pour chaque méthode d'apprentissage.

Tableau 6-4. Performance sur le jeu externe des modèles consensus de chaque méthode et du consensus global.

	Précision[a]	Rappel[a]	Précision[b]	Rappel[b]	PB
NB	0,802	0,665	0,761	0,867	0,766
SVM	0,796	0,775	0,821	0,839	0,807
VP	0,794	0,787	0,828	0,834	0,811
SQS	0,700	0,784	0,806	0,727	0,755
consensus global	0,807	0,816	0,850	0,842	0,829

a : non mutagène
b : mutagène

Comme le laissait présager les modèles individuels, le consensus NB affiche une baisse de performance pour la validation externe vis-à-vis de la 5-CV. Le consensus SQS qui possédait la performance la plus élevée en 5-CV affiche la performance la plus faible sur le test. Ceci peut être du à un léger sur-apprentissage des modèles SQS. Au contraire les consensus SVM et VP affichent des performances légèrement meilleures sur le jeu de test qu'en 5-CV. Il n'y a donc pas eu de sur-apprentissage pour ces derniers modèles. Le consensus global atteint la performance la plus haute sur le test (PB=0,829). Mise à part la méthode SQS, les précisions balancées sont à peu près égales entre la 5-CV et les prédictions aveugles à moins de 0,02 près pour toutes les méthodes d'apprentissages. L'erreur inter-laboratoire du test d'Ames étant estimé à 10%, une précision balancée de 0,9 environ peut être obtenue au maximum. Etant donné cette information les résultats obtenus sont donc proches de leurs limites théoriques.

Il s'avère aussi que le meilleur modèle obtenu lors de ce concours atteint une précision balancée de 0,818 sur le jeu de test externe. Cependant celui-ci ne se base que sur une seule machine d'apprentissage, et prédit l'ensemble des molécules initiales sans nettoyage additionnel des données comme nous l'avons fait.

Afin de pousser l'analyse de notre modèle consensus global, nous avons défini plusieurs seuils d'acceptation des prédictions faites par celui-ci (voir matériel et méthodes). Les résultats sont résumés dans la Figure 6-7 pour différents seuils d'acceptabilité.

Ames

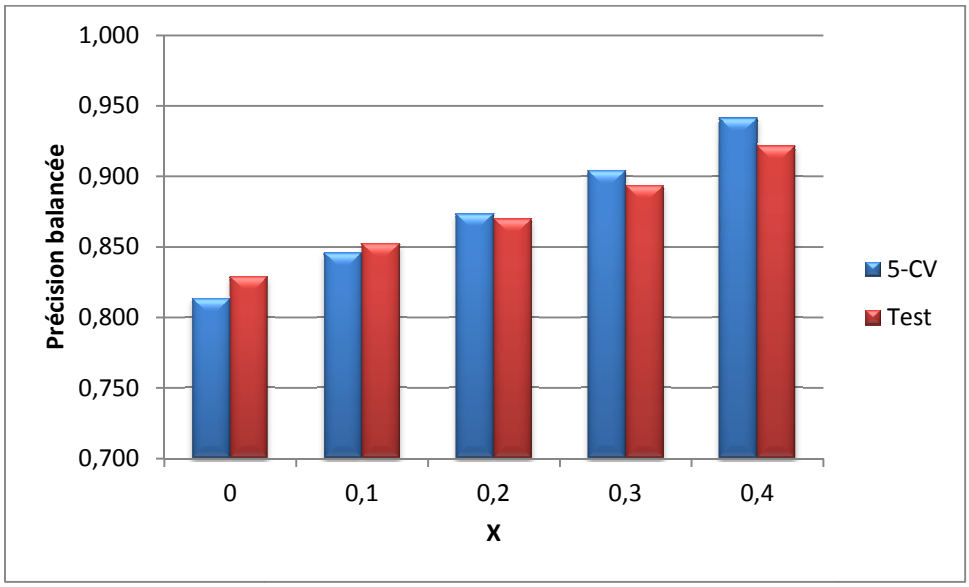
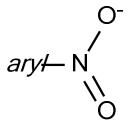
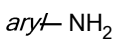
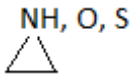
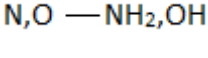
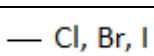
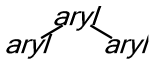


Figure 6-7: Précision balancée en 5-CV et sur le test externe des modèles consensus en faisant varier le seuil d'acceptabilité **X** de la prédiction.

On s'aperçoit que plus notre modèle consensus global fait une prédiction proche de 0 ou 1, plus la précision de celle-ci est élevée. A partir d'un seuil d'acceptabilité de 0,3 on s'aperçoit que notre modèle consensus à des performances équivalentes à celles pouvant être obtenus par le test biologique d'Ames en considérant l'erreur inter-laboratoire de 10%. Ce seuil permet donc de prédire 63% des molécules du jeu avec une erreur équivalente à l'erreur inter-laboratoire.

Notre modèle consensus global confirme les toxicophores déjà établi dans la littérature et responsable de la mutagénicité. Ainsi pour un toxicophore donné, le consensus global prédira la molécule présentant cette sous-structure comme mutagène dans la majorité des cas. Pour les 8 toxicophores généraux proposés dans la publication de Kazius & al. [7], les nombre de molécules mutagènes et non mutagènes retrouvées sont données Tableau 6-5.

Tableau 6-5. Nombre de molécules contenant un toxicophore 1) expérimentalement déterminé comme mutagènes et non mutagènes et 2) prédites comme mutagènes et non mutagènes.

toxicophore	représentation structurale	mutagène	non mutagène	mutagène bien prédit	non mutagène bien prédit
nitro aromatique		233	44	228	25
amine aromatique		148	67	137	37
hétérocycle à 3 branches		41	7	32	6
nitroso		65	2	60	1
hétéroatome non substitué lié à un hétéroatome		62	20	49	12
azo-type		52	28	49	15
halogène aliphatique		111	42	91	27
système aromatique polycyclique		237	29	231	7

On remarque néanmoins que la présence d'un toxicophore n'entrainera pas forcément une prédiction mutagène du consensus global. En effet, pour les composés possédant un groupement nitro aromatique 228/233 mutagènes sont retrouvés et parmi les 44 non mutagènes possédant cette sous-structure 25 composés sont correctement prédits. De manière générale, pour les molécules possédant un des 8 toxicophores présentés, le taux de rappel des mutagènes est très élevé. A l'inverse, le rappel des non mutagènes est bien plus faible, ce qui montre que le consensus global tend à prédire comme mutagène les composés possédant un des toxicophores. Toutefois le rappel non nul des non mutagènes

montre que la mutagénicité ne dépend pas uniquement de la présence d'une seule sous-structure au sein de la molécule, ou alors que les toxicophores présentés ne sont pas assez spécifiques. Cet effet est correctement capturé par le modèle.

6.4. Etude collective des modèles construits par les 12 groupes du concours

6.4.1. Sélection du domaine d'applicabilité

Les modèles que nous avons crée ont servi lors d'une étude visant à comparer les performances de plusieurs domaines d'applicabilités. Un consensus regroupant les prédictions de tous les modèles crées lors de ce concours a également été construit en moyennant les prédictions des 29 modèles.

Au total, 14 domaines d'applicabilités ont été testés sur chaque modèle, dont le modèle consensus. Afin de comparer les performances des différents domaines d'applicabilité, il est nécessaire d'évaluer leur capacité à séparer les prédictions ayant une faible précision des prédictions ayant une grande précision. Pour ce faire, il faut tout d'abord fixer un seuil d'acceptation des prédictions pour chaque domaine d'applicabilité. Ce seuil est fixé de façon à ce que les composés situés à l'intérieur du domaine d'applicabilité soient prédits avec une précision égale à 90%. Il suffit alors de calculer le pourcentage de composés se situant à l'intérieur du domaine d'applicabilité. Celui qui permet de prédire le plus grand nombre de composés avec une précision de 90% est considéré comme le meilleur.

En calculant les proportions prédites par chaque domaine d'applicabilité et pour chaque modèle, 3 domaines d'applicabilités se sont avérés particulièrement efficaces : CONS-STD-QUAL-PROB, CONCORDANCE et CONS-STD-PROB (voir publication en annexe). Ces domaines d'applicabilité permettent de prédire 65% du jeu de test avec une précision de 90%.

6.4.2. Benchmarking des modèles du concours

Les performances des modèles du concours sont données en fonction du seuil du domaine d'applicabilité CONS-STD-PROB sur la Figure 6-8. Plus le seuil d'acceptabilité augmente, plus le pourcentage des composés qui seront prédits est faible.

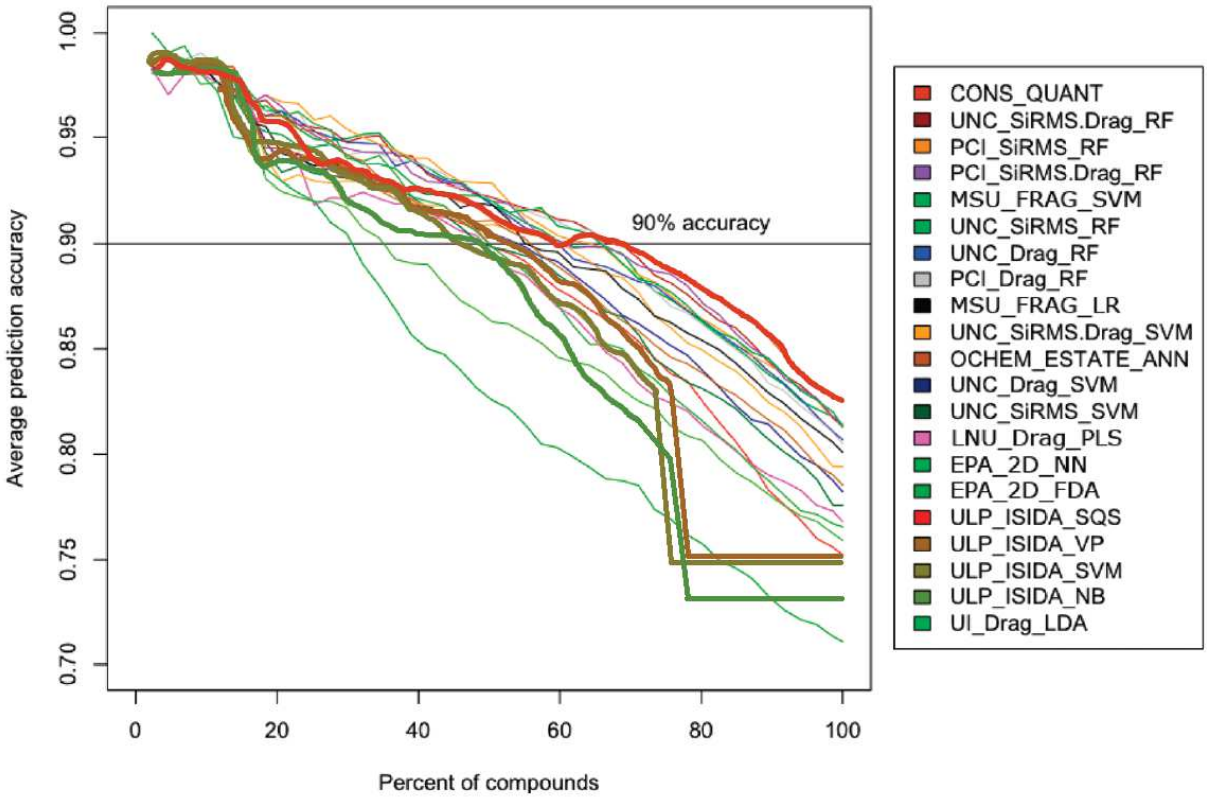


Figure 6-8. Graphique présentant la précision de prédiction des modèles sur le jeu de test en fonction du seuil d'acceptabilité pour le domaine d'applicabilité CONS-STD-PROB. Plus le seuil d'acceptabilité augmente, plus la population de composés prédit diminue. Les lignes surlignées correspondent à nos modèles ainsi qu'au modèle consensus.

Les performances de nos modèles sont honnêtes comparés aux modèles des autres groupes. On remarque que nos modèles s'effondrent brutalement lorsqu'il s'agit de prédire plus de 80% des composés. Cela s'explique par le fait que, de par notre nettoyage initial des données, nous ne prédisons pas tous les composés du jeu de test. Les prédictions pour cette partie des données ont alors été considérées incorrectes de la part des organisateurs. Si l'on regarde la partie se situant juste avant l'effondrement des performances de nos modèles, on constate que la précision

obtenue par la SVM et le VP (environ 84%) est très proches de celle du consensus lorsque celui-ci prédit tout le jeu de test. La performance du NB est la plus faible parmi nos modèles. Ceci est en accord avec les conclusions qui avaient déjà été émises plus haut.

6.5. Conclusion

Dans cette partie, des modèles prédictifs de la mutagénicité d'Ames ont été construits dans le cadre d'un concours international regroupant 12 équipes. 181 jeux de descripteurs fragmentaux ont été générés et des modèles consensus pour la SVM, le VP et le NB ont été construits. Un consensus global regroupant les consensus SVM, VP et NB ainsi que le consensus SQS du docteur Horvath a été construit.

Durant la validation 5-CV, une précision balancée PB=0,813 a été obtenue pour le modèle consensus global, ce qui est meilleur que les performances des méthodes individuelles (PB < 81%). Les performances de prédictions pour le jeu de test sont encore meilleures : PB = 0,829% (consensus global). Autrement dit, 83% des prédictions du modèle se sont révélées exactes. Ceci représente une performance satisfaisante étant donné que l'erreur inter-laboratoires concernant le test d'Ames a été estimée à 10%.

Il a aussi été montré que notre modèle consensus tend à retrouver les mêmes molécules mutagènes que les alertes structurales établies dans la littérature, en permettant toutefois une meilleure distinction des molécules non mutagènes pouvant posséder une alerte structurale.

La comparaison de nos résultats avec les autres groupes ont montrés que nos modèles étaient au moins aussi efficaces que le meilleur modèle obtenu dans ce concours (celui construit par les organisateurs de la compétition eux-mêmes). Le regroupement des modèles de tous les groupes à permis de mener une étude benchmark pour 14 domaines d'applicabilités. Cela a permis de mettre en évidence 3 domaines d'applicabilités comme étant relativement performant.

6.6. Références

1. *Guidance on Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use.* Federal Register, 2012. **77**(110): p. 33748-33769.
2. Ames, B., F. Lee, and W. Durston, *An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens.* Proc. Nat. Acad. Sci. U.S.A., 1973. **70**(3): p. 782-786.
3. Mortelmans, K. and E. Zeiger, *The Ames Salmonella/microsome mutagenicity assay.* Mutat. Res., 2000. **455**(1-2): p. 29-60.
4. Ashby, J., *Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity.* Env. Mut., 1985. **7**(6): p. 919-921.
5. Ashby, J. and R.W. Tennant, *Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP.* Mut. Res./Gen. Toxicol., 1988. **204**(1): p. 17-115.
6. Bailey, A.B., et al., *The use of structure–activity relationship analysis in the food contact notification program.* Regul. Toxicol. Pharm., 2005. **42**(2): p. 225-235.
7. Kazius, J., R. McGuire, and R. Bursi, *Derivation and Validation of Toxicophores for Mutagenicity Prediction.* J. Med. Chem., 2004. **48**(1): p. 312-320.
8. Dearden, J.C., *In silico prediction of drug toxicity.* J. comput. aided mol. des., 2003. **17**(2-4): p. 119-127.
9. Mazzatorta, P., et al., *Integration of Structure−Activity Relationship and Artificial Intelligence Systems To Improve in Silico Prediction of Ames Test Mutagenicity.* J. Chem. Inf. Model., 2006. **47**(1): p. 34-38.
10. ROSITSA, S., F.G. Mojca, and W. Andrew; *Review of QSAR Models and Software Tools for Predicting of Genotoxicity and Carcinogenicity*. 2010.
11. Lozano, S., et al., *Introduction of Jumping Fragments in Combination with QSARs for the Assessment of Classification in Ecotoxicology.* J. Chem. Inf. Model., 2010. **50**(8): p. 1330-1339.
12. Stepan, A., et al., *Structural Alert/Reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A*

Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. Chem. Res. Toxicol., 2011. **24**(9): p. 1345-1410.

13. Sushko, I., et al., *Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set.* J. Chem. Inf. Model., 2010. **50**(12): p. 2094-2111.
14. OpenEye, *Filter*, 2009: Santa Fe, New mexico.

7. Classification des médicaments DILI/non DILI chez l'homme.

7.1. Introduction

Les dommages hépatiques d'origine médicamenteuse (DILI : Drug Induced Liver Injury) sont la principale cause d'insuffisance hépatique aigue chez l'homme et l'effet secondaire le plus courant entrainant la non approbation ou le retrait du médicament sur le marché [1]. Environ 40% des nouveaux candidats médicaments échouent dans les essais cliniques à cause de leur toxicité pourtant non observée dans les études précliniques. Alors que les tests sur les animaux permettent de prédire dans 85% des cas une toxicité cardiovasculaire, 88% une toxicité gastro-intestinale, et 90% une toxicité hématologique, seul dans 50% des cas ces tests arrivent à déterminer une hépatotoxicité [2]. Ce faible taux de succès rend ce type de test inapproprié dans les campagnes de criblage de médicaments. La raison majeure pour laquelle ces tests ont de faibles performances est qu'ils tentent d'expliquer un seul *endpoint* alors que l'hépatotoxicité peut résulter de plusieurs mécanismes à la fois. Un endpoint est une maladie, un symptôme, ou un signe qui constitue une observation clinique ou une réponse à un test pour une cible donnée.

Les approches QSAR sont également largement utilisées dans le domaine de la toxicologie [3]. Cependant il n'existe que peu de méthodes in silico qui se concentrent véritablement à prédire l'hépatotoxicité. Et les modèles existants ont un pouvoir prédictif limité par rapport à ce qui peut être obtenu pour d'autres formes de toxicité. On compte tout de même quelques modèles basés sur des descripteurs à 1D ou 2D [4, 5], d'autres basés sur l'analyse des champs moléculaires [6], ou encore sur l'identification de motifs structural lié à l'hépatotoxicité comme le logiciel MCASE [7].

Depuis peu, de nouveaux modèles toxicologiques utilisant les résultats de tests biologiques et non plus uniquement la structure des molécules sont développés. Il a été démontré que de tels modèles permettent d'obtenir de bien meilleurs résultats que les modèles conventionnels. Ainsi, Low, Y. & al. [8] ont proposé une approche prédisant l'hépatotoxicité liée à la prise de médicaments chez le rat à l'aide de méthodes QSAR et de résultats de tests in vivo pouvant expliquer cette toxicité. Les résultats des tests in vivo ont été traités comme des descripteurs biologiques afin de

construire des modèles prédictifs. Il a été montré en validation croisée externe à 5 paquets que les modèles utilisant uniquement des informations liées à la structure chimique atteignent au mieux 0,6 de précision balancée, alors que les modèles utilisant les descripteurs biologiques atteignent 0,78 de PB.

Une autre étude de Liu & al. [9] propose l'utilisation de 13 marqueurs/endpoints de l'hépatotoxicité pour classer les médicaments comme DILI ou non : si une réponse positive est constatée pour un marqueur parmi les 13 possible, le médicament est considéré DILI. Ce modèle composé d'une règle atteint une précision de 91% et 74% sur 2 jeux externes. Des modèles in silico ont alors été construits pour chaque endpoint et les composés DILI ont à nouveau été prédits de la même manière : si un modèle donne une prédiction positive, le médicament est considéré comme DILI. L'application de cette stratégie pour 3 jeux externes entraine des performances de l'ordre de 60% à 70% de précision, ce qui n'est plus suffisant.

Le but du travail présent a été d'évaluer la performance prédictive des méthodes QSAR utilisant des données in vitro, générées par KaLy-Cell, en tant que descripteurs afin de prédire l'hépatotoxicité chez l'homme. Pour ce faire, une base de données contenant 424 composés de référence connus pour leur effet hépatotoxique chez l'homme a été construite. Les composés positifs (247) ont été classés selon le(s) type(s) de mécanisme(s) par le(s)quel(s) ceux-ci induisent des dommages au foie. Les données disponibles sur les mécanismes considérés comme pertinents dans la détection d'une molécule DILI ont été extraites de la littérature. Un modèle de prédiction a alors été construit en utilisant le jeu de données de référence ainsi que les effets biologiques choisis (approche binaire). Le modèle obtenu a été jugé hautement prédictif (PB = 0,865).

Un nombre limité de composés de référence (9 DILI + et 1 DILI-) a alors été testé en utilisant à la fois le modèle in silico construit et les résultats de tests in vitro censés traduire la présence ou non des mécanismes biologiques considérés important dans la détection des molécules DILI. En appliquant notre modèle nous avons pu correctement prédire 9 composés sur 10. L'originalité de ce travail repose sur le fait que contrairement aux autres études proposant des modèles basés sur des observations in vivo et/ou des modèles QSAR, nous proposons ici d'utiliser des

mesures in vitro permettant d'obtenir une stratégie applicable dès les étapes de recherche de médicaments pour prédire l'hépatotoxicité.

Le reste de ce travail a été supprimé du manuscrit d'origine car il a été considéré comme **CONFIDENTIEL**.

7.2. Références

1. Holt, M.P. and C. Ju, *Mechanisms of drug-induced liver injury.* The AAPS journal, 2006. **8**(1): p. 48-54.
2. Greaves, P., A. Williams, and M. Eve, *First dose of potential new medicines to humans: how animals help.* Nat. Rev. Drug. Discov., 2004. **3**(3): p. 226-236.
3. Dearden, J.C., *In silico prediction of drug toxicity.* J. Comput. Aided Mol. Des., 2003. **17**(2-4): p. 119-127.
4. Cheng, A. and S. Dixon, *In silico models for the prediction of dose-dependent human hepatotoxicity.* J. Comput. Aided Mol. Des., 2003. **17**(12): p. 811-823.
5. Fourches, D., et al., *Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species.* Chem. Res. Toxicol., 2010. **23**(1): p. 171-183.
6. Clark, R.D., et al., *Modelling in vitro hepatotoxicity using molecular interaction fields and SIMCA.* J. Mol. Graph. Model., 2004. **22**(6): p. 487-497.
7. Contrera, J., et al., *MCASE Prediction of Hepatotoxicity Using Post-Market Adverse Effects Data.* Hepatotoxicity Steering Committee Meeting, Rockville, MD, January 21, 2003.
8. Low, Y., et al., *Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches.* Chem. Res. Toxicol., 2011. **24**(8): p. 1251-1262.
9. Liu, Z., et al., *Translating Clinical Findings into Knowledge in Drug Safety Evaluation - Drug Induced Liver Injury Prediction System (DILIps).* PLoS Comput. Biol., 2011. **7**(12): p. e1002310.

8. Logiciels développés

8.1. GCR Designer

Le logiciel « GCR Designer » est un logiciel déjà existant et développé en Delphi par le Dr. Frank Hoonakker [1] afin de générer des graphes condensés de réactions à partir de réactions préalablement mappées. Pour rappel (voir 1.2), le mapping correspond à la procédure qui vise à assigner un même nombre pour un même atome des 2 côtés de la réaction. Cependant la procédure de « atom-to-atom » mapping ne suffit pas à générer un GCR. En effet, il faut aussi assigner manuellement pour la réaction considérée les différents centres réactionnels. Il y'a 2 types de centres réactionnels : les cassures/formations de liaisons, et les changements d'ordre de liaisons (liaison simple transformée en liaison double par exemple). Un exemple des différents types de centre réactionnels est proposé Figure 8-1.

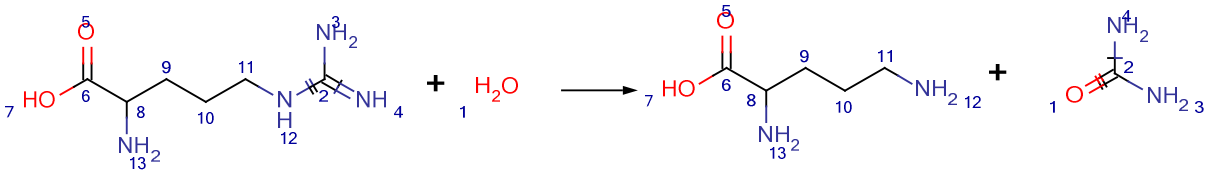


Figure 8-1. Réaction pour lequel le mapping ainsi que les centres réactionnels sont présents. Une liaison barrée d'un trait signifie que la liaison est modifiée (changement d'ordre), et une liaison doublement barrée signifie que la liaison est soit crée, soit cassée.

Un module complémentaire, exécutable en ligne de commande sous Linux et Windows et développé en FreePacal, a donc été développé durant cette thèse afin de permettre l'identification automatique des différents centre réactionnels et de modifier le fichier rxn/rdf afin d'y ajouter cette information. L'ajout des centres réactionnels se traduit par l'ajout d'un nombre dans la dernière colonne du « bond block » du fichier rxn/rdf pour la liaison considérée comme transformée: « 4 » si la liaison est cassée ou formée, « 8 » si il y'a un changement d'ordre de la liaison.

L'algorithme conceptuel de fonctionnement du programme est le suivant :

1) On sélectionne un atome du côté des réactifs.
2) On recense les atomes auquel l'atome sélectionné est lié ainsi que les types de liaisons entre eux.
3) Grace à la procédure d' « atom-to-atom mapping », on identifie l'atome sélectionné du côté des produits.
4) On répète le point 2 pour l'atome correspondant du côté des produits.
5) On compare le voisinage de l'atome des 2 côtés de la réaction :
   - Si une liaison n'existe plus dans un des 2 côtés de la réaction, alors on modifie la ligne décrivant la liaison du côté de la réaction où celle-ci existe pour y ajouter un centre réactionnel de type formée/cassée.
   - Si une liaison n'est plus du même type, alors on modifie la ligne décrivant la liaison transformée des 2 côtés de la réaction pour y ajouter un centre réactionnel de type changement d'ordre.
6) On répète l'algorithme pour un nouvel atome du côté des réactifs.

Il est maintenant possible de générer des GCRs de manière complètement automatique grâce au mapping et à l'identification des centres réactionnels.

L'algorithme de création des GCRs génère un graphe pour chaque réaction en 3 étapes :

- Les produits de la réaction sont copiés vers le nouveau graphe.
- Les liaisons des produits sur le nouveau graphe possédant un centre réactionnel sont modifiées selon la situation :
  - Si la liaison possède un centre réactionnel de type formée, on modifie alors le type de liaison entre les 2 atomes en spécifiant qu'il y'a création de liaison, et on indique le type de liaison crée (liaison simple, double liaison, etc.).
  - Si la liaison possède un centre réactionnel de type changement d'ordre, on identifie les 2 atomes formant la liaison du côté des réactifs puis on modifie le type de liaison entre les 2 atomes du nouveau graphe en spécifiant le type de la liaison dans les réactifs et le type de la liaison dans les produits.

- Les liaisons des réactifs possédant un centre réactionnel de type cassée sont énumérées. Les atomes correspondant aux liaisons cassées sont identifiés sur le nouveau graphe puis les liaisons reformées en spécifiant le type de la liaison initiale et le fait que la liaison est cassée.

Cependant l'algorithme développé initialement par le Dr. Frank Hoonakker possède plusieurs soucis et plusieurs améliorations/corrections ont du être apportées durant cette thèse dans le cadre de la modélisation correcte de réactions métaboliques entre autres :

- Lors d'une réaction, seul le produit principal (1er produit présent dans le fichier *rxn*/*rdf*) était retrouvé dans le GCR résultant, les autres produits étant considérés comme secondaires. Le programme a donc été modifié et plusieurs procédures et fonctions ont été rajoutées afin de permettre l'intégration de tous les produits dans le GCR.

- La stratégie de lecture des réactifs a aussi été revue. En effet, il y avait répétition de certains réactifs lorsque la réaction comportait plus de deux réactifs. Par exemple pour une réaction contenant trois réactifs, les deux premiers étaient correctement lus et le 3e était un doublon du 2^{e} réactif, ce qui donnait au final l'écriture du 1er réactif puis celle du 2e réactif par deux fois.

- Pour les réactions faisant intervenir des protons H+ du côté des produits, la liaison rompue entre l'hydrogène et l'atome auquel il était attaché du côté des réactifs était absente du GCR ce qui laissait apparaître un proton isolé sur le graphe. En même temps que la résolution de ce problème, des procédures ont été ajoutées afin d'éliminer tous les hydrogènes explicites du graphe final, c'est-à-dire ceux qui ne sont pas directement impliqués dans une transformation de liaison, et ainsi permettre une visualisation plus aérée du graphe final.

- De même pour les réactions comprenant plus de 99 atomes, les atomes possédant un numéro de mapping supérieur à 99 ne sont pas traités. Il a fallut modifier la forme de lecture des données pour corriger ce problème.

- Lorsqu'une double ou une triple liaison était entièrement rompue, ce changement apparaissait dans le GCR comme la rupture d'une simple liaison. Par défaut le type de liaison rompu était simple. Il a donc fallut modifier la procédure pour que le programme prenne en compte le type de liaison effectivement rompu.

8.2. DILIpredictor

8.2.1. Présentation

Le logiciel « DILIpredictor » permet d'appliquer un modèle de classification séparant les composés DILI des non DILI en introduisant le résultat des tests vitro. Deux modèles sont disponibles dans l'interface actuelle : un modèle (modèle 1) ne faisant intervenir que les endpoints biologiques expérimentalement mesurables à l'aide de tests vitro (voir **Erreur ! Source du renvoi introuvable.**), et un modèle (modèle 2) utilisant ces mêmes tests vitro avec en supplément les prédictions QSAR pour les endpoints RM et DILI (voir **Erreur ! Source du renvoi introuvable.**). Le programme est compilé pour une utilisation sous Linux et sous Windows.

L'interface graphique se présente comme tel :

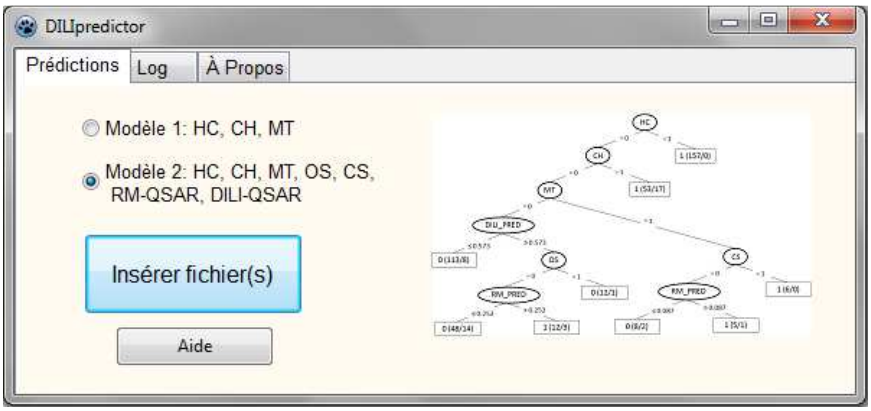


Figure 8-2. Interface graphique du logiciel DILIpredictor.

8.2.1. Mise en route rapide

Une fois le modèle sélectionné on clique sur « Insérer les fichier(s) ». On distingue 2 types de fichiers d'entrées :

- Les fichiers .csv qui contiennent les résultats des tests vitro, voir des prédictions QSAR si celles-ci sont déjà disponibles.
- Les fichiers .sdf qui contiennent les structures chimiques des composés à prédire.

  NB : les fichiers .sdf ne sont requis que lorsque le modèle 2 est sélectionné.

On indique finalement un nom pour le fichier de sortie qui sera au format .csv. Dans ce dernier fichier la dernière colonne contiendra la prédiction DILI.

8.2.2. Fonctionnement

Quel que soit le modèle sélectionné, le fichier .csv d'entrée doit contenir un entête spécifiant à quel endpoint fait référence la colonne lue. Selon le modèle, seules les colonnes d'intérêt à l'application de celui-ci sont identifiées. Par exemple la colonne contenant les résultats des tests vitro relevant d'une apoptose/nécrose sera déterminée par *Hepatocellular*, *HC*, *necrosis* ou encore *apoptosis*.

Dans le cas du modèle 1, une fois que les colonnes correspondant aux endpoints sont identifiées, l'algorithme de prédiction peut être directement utilisé.

Dans le cas du second modèle il y'a 2 options. La première prévoit que le fichier .csv d'entrée contient déjà les prédictions des modèles QSAR pour les endpoints RM et DILI. Dans ce cas, la même procédure que pour le modèle 1 est appliquée. Si les prédictions QSAR ne sont pas disponibles, la seconde option est envisagée. Cette seconde option prévoit de fournir, en plus du fichier .csv contenant les résultats des tests vitro, un fichier .sdf contenant les structures chimiques des composés à prédire. Chaque structure est alors fragmentée en descripteurs ISIDA via le logiciel ISIDA/Fragmentor, et un modèle SVM préalablement construit est appliqué avec le logiciel LIBSVM. Les prédictions sont alors récupérées, ajoutées au fichier d'entrée .csv, et l'algorithme de prédiction de l'arbre peut alors être exécuté.

Le choix des fragmentations à effectuer et des modèles à appliquer pour sur les structures chimiques est décidé grâce au fichier .xml montré ci-dessous : *Frg_pred.xml*.

```
<?xml version="1.0"?>
- <DILI_Predictor>
    <Title>Fragment SDF File</Title>
    <Author>C. Muller</Author>
    <Comment>PhD Thesis, 2012</Comment>
  - <Fragments>
      <Fragmentation AP="False" Max="2" Min="2" type="3">RM_t3l2u2</Fragmentation>
      <Fragmentation AP="True" Max="5" Min="2" type="11">DILI_t11l2u5AP</Fragmentation>
    </Fragments>
  - <Models>
      <SVM_Model>RM_t3l2u2</SVM_Model>
      <SVM_Model>DILI_t11l2u5AP</SVM_Model>
    </Models>
</DILI_Predictor>
```

Figure 8-3. Fichier XML utilisé par le DILIpredictor.

Il y'a deux balises d'intérêt dans ce fichier : Fragments et Models. Les noeuds enfants de *Fragments* contiennent le nom du fichier header de référence utilisé par le fragmenteur (ex : RM_t3l2u2.hdr), ainsi que le type de fragmentation et la longueur minimale et maximale des fragments générés. En effet, les modèles prédisant la propriété RM et DILI ne sont pas basés sur les mêmes fragments. Les noeuds enfants de *Models* contiennent le nom du fichier modèle préalablement construit et requis par le logiciel LIBSVM (ex : RM_t3l2u2.model).

A noter qu'il possible de modifier ce fichier afin d'ajouter de nouveaux modèles pour d'autres endpoints ou de remplacer ceux existant. Il suffit alors de placer dans le répertoire contenant l'exécutable du fragmenteur et de LIBSVM les fichiers .hdr et .model correspondant.

8.3. Références

1. Hoonakker, F., *Graphes condenses de réactions, applications à la recherché par similarité, la classification et la modélisation.*, 2008, Université de Strasbourg. p. 252.

Conclusion générale

Le devenir d'un composé dans l'organisme humain est un enjeu stratégique pour le développement de nouvelles molécules à visé thérapeutiques ou comme outils de diagnostique. Les connaissances actuelles publiques sont parcellaires et rares. Par ailleurs, l'acquisition de nouvelles données est difficile pour des raisons économiques et éthiques. Cela justifie les travaux de rationalisation des connaissances, au travers notamment d'approches chémoinformatique, qui sont entrepris dans la communauté scientifique. Les résultats proposés dans cette thèse s'inscrivent dans ce cadre.

Au cours de cette thèse nous avons proposé l'utilisation de graphes condensés de réactions afin de représenter les réactions métaboliques :

Cette représentation nous a permis de développer une approche nouvelle dans le domaine visant à détecter les mappings atomique incorrect obtenus automatiquement. L'approche se base sur le fait que pour une même réaction, un mapping incorrect conduit à l'obtention d'un GCR différent de celui obtenu à l'aide d'un mapping correct. Une très forte précision balancée (PB=0,95) a pu être obtenu dans l'identification de mappings correct et incorrect pour les réactions métaboliques de la KEGG des classes 1, 2 et 3. Cette étude a d'ailleurs fait l'objet d'une publication acceptée dans le *Journal of Chemical Information and Modeling*.

L'utilisation de la technologie des GCRs permet de générer des descripteurs impliquant des liaisons dynamiques traduisant les modifications effectuées sur les réactifs lors de la réaction. Il est alors possible de générer des descripteurs reflétant exclusivement le cœur réactionnel ainsi que son environnement proche. L'efficacité des descripteurs issus de cette technologie a été démontrée lors de la classification des réactions métaboliques et de la prédiction de la régiosélectivité des biotransformations chez l'homme :

- Une carte de Kohonen représentant des réactions métaboliques de 3 classes métaboliques distinctes a été générée. Pour ce faire des descripteurs tenant compte uniquement des liaisons transformées autour du cœur réactionnel ont été générés. L'observation de la carte obtenue a démontré une nette séparation des 3 classes.

- Les sites d'hydroxylations aromatiques pour les substrats du CYP1A2 de l'homme ont été prédits avec une précision balancée de 0,78 de moyenne sur les jeux de test. Comparé à MetaSite, qui est la référence dans le domaine, nos modèles ont atteint de meilleures performances que ce dernier.

- Les sites d'oxydation pour les substrats du CYP3A4 de l'homme ont été prédits avec une précision balancée de 0,78 sur le jeu de test. Il a été montré que le meilleur modèle obtenu atteint des performances comparables à celles de SMARTCyp qui est un modèle fondé sur le mécanisme réactionnel.

La seconde partie de cette thèse traite de la toxicité et des diverses approches mis en jeu pour parvenir à prédire cette propriété :

La mutagénicité d'Ames a été prédite dans le cadre d'un concours international regroupant 29 modèles construits par 12 équipes. Ce challenge a permis de comparer plusieurs domaines d'applicabilité et de proposer une méthode afin d'identifier les meilleurs. Cette étude a d'ailleurs fait l'objet d'une publication dans le *Journal of Chemical Information and Modeling*. D'un point de vue personnel, les modèles que nous avons obtenus se sont révélés performants. Une précision balancée de 0,83 a pu être obtenue pour le modèle consensus. Les principaux toxicophores présents dans la littérature ont aussi été confirmés par nos modèles.

Enfin, des descripteurs biologiques ainsi qu'hybrides (biologiques + QSAR) ont permis de construire des modèles bien plus performants que les modèles basés uniquement sur les descripteurs QSAR. Grace à ces descripteurs, un très bon modèle (PB=0,865) permettant d'identifier les molécules hépatotoxiques pour l'homme a été obtenu. Des tests in vitro ont aussi été mis en place afin d'extrapoler les observations in vivo utilisées dans la construction du modèle. Les résultats des tests in vitro ont permis de classer correctement 9 des 10 composés externes.

L'ensemble du travail effectué lors de cette thèse ouvre de nombreuses perspectives:

- Il serait intéressant de classer les 6 classes enzymatiques de la KEGG selon le type de réactions qu'elles catalysent (1er nombre du chiffre EC) et pour un

plus grand jeu de données afin de tenter de les séparer plus précisément en déterminant les 2^{e}, 3^{e} et 4^{e} chiffres du nombre E.C.

- Dans certains cas, les descripteurs focalisés sur le centre réactionnel et son environnement proche ne suffisent pas, notamment pour la prédiction des sites d'oxydation où l'orientation du substrat dans le complexe rentre en compte. Pour surmonter cette difficulté, on pourrait par exemple ajouter des descripteurs qui prennent en compte la taille de la molécule ou encore sa forme. Une étape de docking pourrait aussi être mise en place pour limiter l'application de nos modèles aux zones accessibles par la porphyrine lors de l'oxydation.

- Pour confirmer l'extrapolation vivo/vitro qui a été faite lors de la prédiction de l'hépatotoxicité chez l'homme il faudrait tester notre modèle sur un pleu grand jeu de test. L'obtention de nouveaux endpoints ainsi que la prédiction de paramètres pharmacocinétiques comme la concentration circulante pourrait permettre de tester et d'identifier avec plus de précisions les molécules DILI.

Ces résultats sont une contribution aux efforts colossaux encore nécessaires pour parvenir à maîtriser les interactions entre un organisme humain vivant et son environnement chimique. Cette question dépasse de loin, le seul cadre de l'industrie pharmaceutique et touche à des débats de sociétés beaucoup plus généraux.

9. Communications

9.1. Publications

Sushko I., Novotarskyi S., Körner R., Pandey A.K., Cherkasov A., Li J., Gramatica P., Hansen K., Schroeter T., Müller K.R., Xi L., Liu H., Yao X., Öberg T., Hormozdiari F., Dao P., Sahinalp C., Todeschini R., Polishchuk P., Artemenko A., Kuz'min V., Martin T.M., Young D.M., Fourches D., Muratov E., Tropsha A., Baskin I., Horvath D., Marcou G., Muller C., Varnek A., Prokopenko V.V., Tetko I.V., *Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set.* J. Chem. Inf. Model., 2010. **50**(12): p. 2094-2111.

Muller C., Marcou G., Aires-de-Sousa J., Varnek A., *Identification of incorrect atomic mapping of reactions from automatic software using condensed graph of reactions. J*. Chem. Inf. Model., 2012, **52**(12), p. 3116–3122

Muller C., Pekthong D., Desbans C., Alexandre E., Marcou G., Richert L. and Varnek A., *Prediction of Drug-Induced Liver Injury in human.* (en préparation)

9.2. Communications par affiche

Muller C., Marcou G., Horvath D., Varnek A., *Blind predictions of the AMES mutagenicity using SVM, Naïve Bayes, Voting Perceptron and SQS models.* 2009, Montpellier (France), Journées nationales de la Chémoinformatique.

Muller C., Marcou G., Varnek A., *Predictive models to detect incorrect atom-atom mapping of reactions using condensed graph of reactions.* 2010, Obernai (France), 2nd Strasbourg Summer School on Chemoinformatics.

Muller C., Marcou G., Varnek A., *Prediction of Aromatic Hydroxylation Sites for CYP1A2 substrates.* 2010, Lviv (Ukraine), 3rd International Summer School “Supramolecular Systems in Chemistry and Biology”.

Muller C., Marcou G., Varnek A., *Predictive models to detect incorrect atom-atom mapping of reactions using condensed graph of reactions.* 2011, Budapest (Hongrie), ChemAxon European User Group Meeting.

Muller C., Marcou G., Varnek A., *Prediction of Oxidation Sites for CYP3A4 substrates.* 2011, Cabourg (France), Journées nationales de la Chémoinformatique.

Muller C., Marcou G., Varnek A., *Prediction of Oxidation Sites for human CYP3A4 substrates using Condensed Graph of Reactions.* 2012, Strasbourg (France), 3nd Strasbourg Summer School on Chemoinformatics.

10. Annexes

10.1. Algorithme de Ripper pour l'apprentissage de règles et interprétation des symboles. (extrait de Witten, I.; Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann: 2005.)

```
Initialize E to the instance set
For each class C, from smallest to largest
        BUILD:
                Split E into Growing and Pruning sets in the ratio 2:1
                Repeat until (a) there are no more uncovered examples of C; or (b) the
                        description length (DL) of ruleset and examples is 64 bits greater
                        than the smallest DL found so far, or (c) the error rate exceeds 50%:
                GROW phase: Grow a rule by greedily adding conditions until the rule
                        is 100% accurate by testing every possible value of each attribute
                        and selecting the condition with greatest information gain G
                PRUNE phase: Prune conditions in last-to-first order. Continue as long
                        as the worth W of the rule increases
        OPTIMIZE:
                GENERATE VARIANTS:
                For each rule R for class C,
                        Split E afresh into Growing and Pruning sets
                        Remove all instances from the Pruning set that are covered by other
                        rules for C
                        Use GROW and PRUNE to generate and prune two competing rules from
                the newly-split data:
                                R1 is a new rule, rebuilt from scratch;
                                R2 is generated by greedily adding antecedents to R.
                        Prune using the metric A (instead of W) on this reduced data
                SELECT REPRESENTATIVE:
                Replace R by whichever of R, R1 and R2 has the smallest DL.
        MOP UP:
                If there are residual uncovered instances of class C, return to the
                        BUILD stage to generate more rules based on these instances.
        CLEAN UP:
                Calculate DL for the whole ruleset and for the ruleset with each rule in
                turn omitted; delete any rule that increases the DL
                Remove instances covered by the rules just generated
Continue
```

DL: The description length DL is a complex formula that takes into account the number of bits needed to send a set of examples with respect to a set of rules, the number of bits required to send a rule with k conditions, and the number of bits needed to send the integer k—times an arbitrary factor of 50% to compensate for possible redundancy in the attributes.

$G = p[\log(p/t) - \log(P/T)]$

$W = \frac{p+1}{t+2}$

$A = \frac{p+n'}{T}$; accuracy for this rule

p = number of positive examples covered by this rule (true positives)

n = number of negative examples covered by this rule (false negatives)

t = p + n; total number of examples covered by this rule

n′ = N – n; number of negative examples not covered by this rule (true negatives)

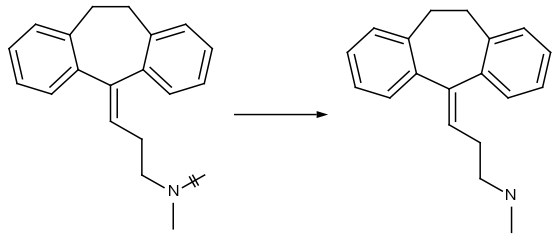
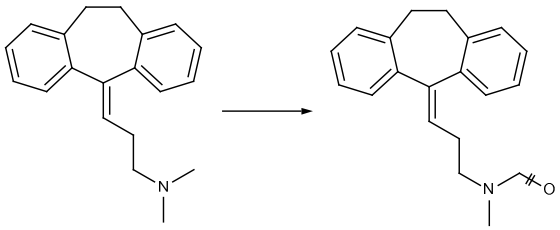
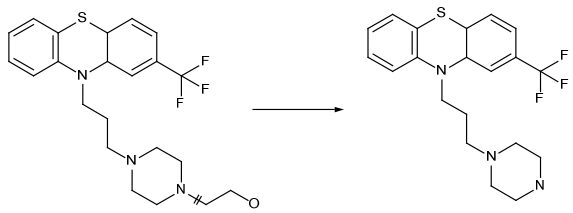
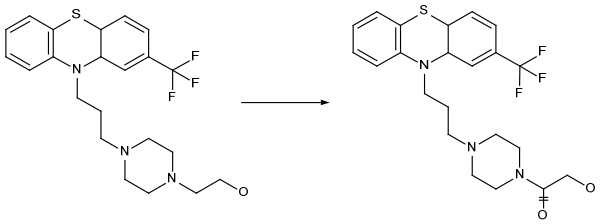
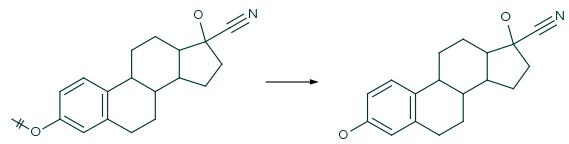
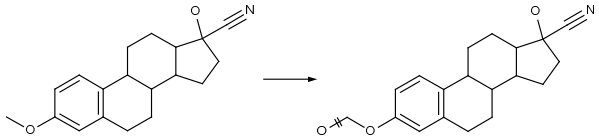
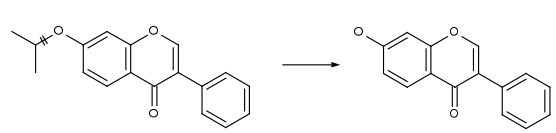
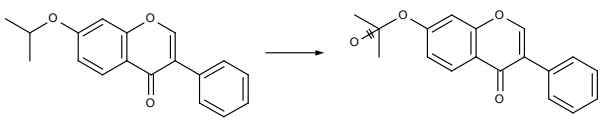
P = number of positive examples of this class

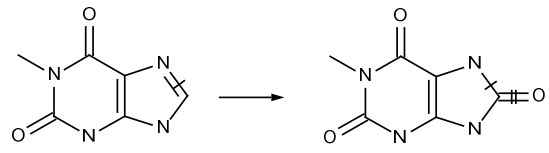
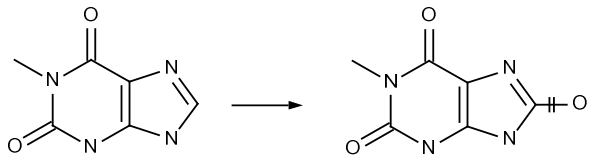
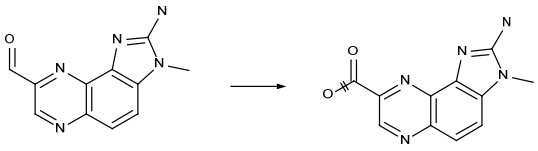
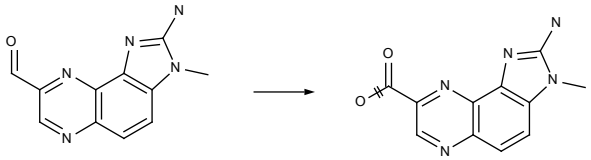
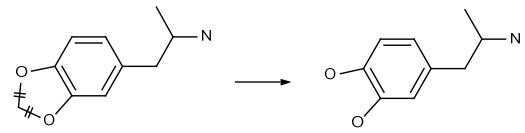
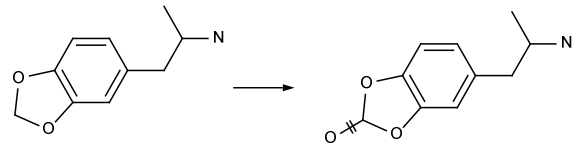
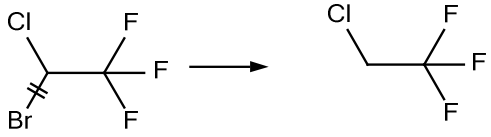
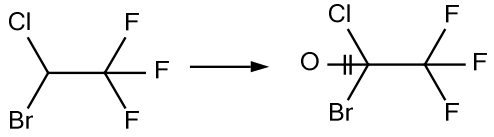
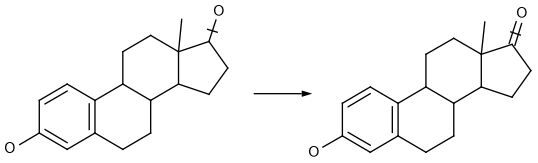
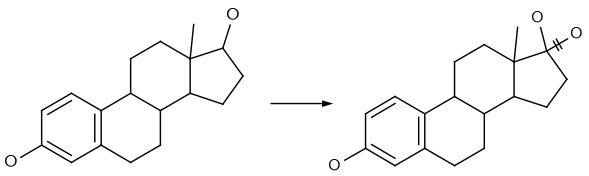
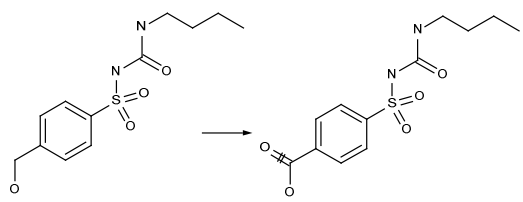
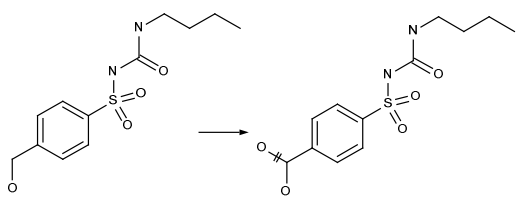
N = number of negative examples of this class

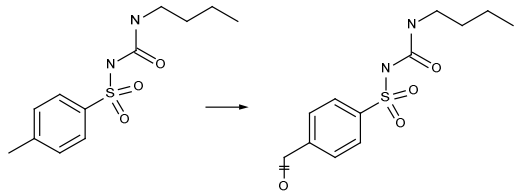
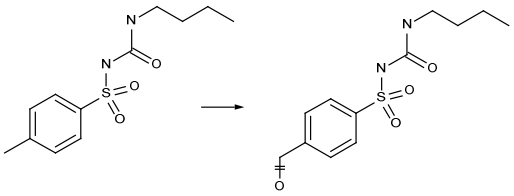
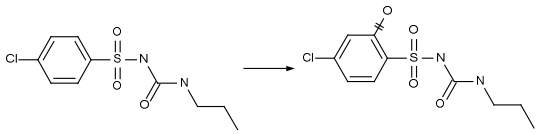
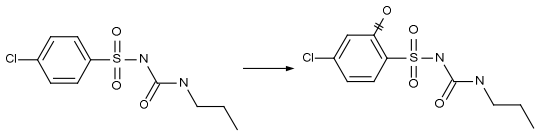
T = P + N; total number of examples of this class

10.2. Exemples de sites d'oxydation marqués avec la méthode P450 CARBOX.

Pour chaque type de biotransformations la partie à gauche représente la transformation initiale, la partie à droite représente la transformation après application de la méthode P450CARBOX.

N-deméthylation	
	
N-dealkylation	
	
O-deméthylation	
	
O-dealkylation	
	

Attachement d'un groupe oxo sur un carbone	
	
Transformation d'aldéhyde en acide carboxylique	
	
Ouverture d'un hétérocycle aliphatique avec détachement d'un atome de carbone	
	
Dehalogénation	
	
Transformation d'un hydroxyle en cétone	
	
Transformation d'un hydroxyle en acide carboxylique	
	

Hydroxylation aliphatique	
	
Hydroxylation aromatique	
	

10.3. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set.

Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set

Iurii Sushko,[†] Sergii Novotarskyi,[†] Robert Körner,[†] Anil Kumar Pandey,[†] Artem Cherkasov,[‡] Jiazhong Li,[§] Paola Gramatica,[§] Katja Hansen,[||] Timon Schroeter,[||,⊥] Klaus-Robert Müller,[||] Lili Xi,[#] Huanxiang Liu,[▽] Xiaojun Yao,[#] Tomas Öberg,[○] Farhad Hormozdiari,[◆], Phuong Dao[◆] Cenk Sahinalp,[◆] Roberto Todeschini,[¶] Pavel Polishchuk,[+] Anatoliy Artemenko,[+] Victor Kuz'min,[+] Todd M. Martin,[%] Douglas M. Young,[%] Denis Fourches,[□] Eugene Muratov,[+,□] Alexander Tropsha,[□] Igor Baskin,[■] Dragos Horvath,[●] Gilles Marcou,[●] Christophe Muller,[●] Alexander Varnek,[●] Volodymyr V. Prokopenko,[△] and Igor V. Tetko*[,†]

Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen—German Research Center for Environmental Health (GmbH), Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany, University of British Columbia, Vancouver Prostate Centre, 2660 Oak str., Vancouver, BC, V6H 3Z6, Canada, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Via Dunant 3, Varese 21100, Italy, Machine Learning Department, Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, Germany, Bayer Schering Pharma AG, Nonclinical Drug Safety, Müllerstr. 178, 13353 Berlin, Germany, Department of Chemistry, Lanzhou University, Tianshui South Road 222, Lanzhou 730000, China, School of Pharmacy, Lanzhou University, Lanzhou 730000, China, School of Natural Sciences, Linnæus University, 391 82 Kalmar, Sweden, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada, Milano Chemometrics & QSAR Research Group, Dept. Environmental Sciences, University of Milano—Bicocca, 20126 Milan, Italy, A.V. Bogatsky Physico-Chemical Institute of National Academy of Science of Ukraine, Lustdorfskaya doroga 86, Odessa 65080, Ukraine, Clean Processes Branch, National Risk Management Research Laboratory, United States Environmental Protection Agency, Cincinnati, Ohio 45268, United States, Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States, Department of Chemistry, Moscow State University, 119991, Moscow, Russia, Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg, 4 rue B. Pascal, Strasbourg 67000, France, and Institute of Bioorganic & Petrochemistry, Ukrainian Academy of Sciences, Murmanskaya 1, 02660 Kyiv-94, Ukraine

Received July 12, 2010

The estimation of accuracy and applicability of QSAR and QSPR models for biological and physicochemical properties represents a critical problem. The developed parameter of "distance to model" (DM) is defined as a metric of similarity between the training and test set compounds that have been subjected to QSAR/QSPR modeling. In our previous work, we demonstrated the utility and optimal performance of DM metrics that have been based on the standard deviation within an ensemble of QSAR models. The current study applies such analysis to 30 QSAR models for the Ames mutagenicity data set that were previously reported within the 2009 QSAR challenge. We demonstrate that the DMs based on an ensemble (consensus) model provide systematically better performance than other DMs. The presented approach identifies 30−60% of compounds having an accuracy of prediction similar to the interlaboratory accuracy of the Ames test, which is estimated to be 90%. Thus, the *in silico* predictions can be used to halve the cost of experimental measurements by providing a similar prediction accuracy. The developed model has been made publicly available at http://ochem.eu/models/1.

INTRODUCTION

Any QSAR/QSPR prediction of biological and/or physicochemical properties has limited value without an estimated applicability domain of a model. Researchers cannot make much use of a prediction for a particular compound if there is no information available on whether this prediction is reliable or not, in other words, whether the given model is applicable. Currently, this problem is being addressed by ongoing studies of applicability domain (AD) assessment.

The conventional methods for estimating model performance are the root-mean-square error (RMSE) and the Pearson correlation coefficient (R^2) of cross-validation. These measures are easily computable and interpretable. However, in general, some groups of chemical compounds can be predicted well, whereas others allow only low prediction accuracy. Thus, depending on the data composition, one can

* Corresponding author tel./fax: +49-89-3187-3575, e-mail: itetko@vcclab.org.
[†] Helmholtz Zentrum Muenchen—German Research Center for Environmental Health (GmbH).
[‡] University of British Columbia.
[§] University of Insubria.
[||] Technical University of Berlin.
[⊥] Bayer Schering Pharma AG.
[#] Department of Chemistry, Lanzhou University.
[▽] School of Pharmacy, Lanzhou University.
[○] Linnæus University.
[◆] Simon Fraser University.
[¶] University of Milano—Bicocca.
[+] A.V. Bogatsky Physico-Chemical Institute of National Academy of Science of Ukraine.
[%] United States Environmental Protection Agency.
[□] University of North Carolina at Chapel Hill.
[■] Moscow State University.
[●] Université de Strasbourg.
[△] Ukrainian Academy of Sciences.

10.1021/ci100253r
Published on Web 10/29/2010

observe significant differences in the estimated and observed statistical parameters.

In particular, when assessing QSAR model performance, one should not only ensure that the predicted accuracies for the training and testing sets are comparable and high but also that the distribution of descriptors' values is uniform within the sets. Under this assumption, the statistical parameters for the new data should indeed be similar to the estimated average values. However, average values can provide biased results if external data distributed differently compared to the modeling set.

Moreover, the number of experimentally available observation points is usually in the range of hundreds (complex biological properties, such as ADMETox data) to hundreds of thousands of measurements (physicochemical properties or HTS data). These numbers are dramatically smaller than the number of compounds for which estimation of properties is needed, e.g. 2×10^7 commercially available molecules or 10^{20} to 10^{24} synthetically accessible molecules[1] or even 10^{80} to 10^{100} theoretically existing chemical structures. Thus, the scenario when QSAR/QSPR predictive models are intended for chemical structures that are different from the training/testing set molecules is a rule rather than an exception.

Thus, the goal of the AD approaches is to estimate the prediction accuracy for each modeled compound individually. Using this information, one can estimate the accuracy of prediction for an arbitrary data set regardless of its similarity to the set used to validate the model.

QSAR studies can assess the accuracy of predictions in different ways. The simple ones try to distinguish reliable vs. nonreliable predictions. They usually assume that the accuracy of prediction of molecules, which are inside a space of descriptors covered by the training set, is similar to the estimated accuracy of the model. These methods include:

• Descriptor boxes: consider compounds with descriptors, lying in predefined parallelepipeds in multidimensional descriptor space, as being inside of the applicability domain of the model.[2–4]

• Leverage-based: all compounds, whose leverage (known as the Mahalanobis distance) with the training set exceeds some predefined limit, are considered to be outside the applicability domain.[5,6]

Approaches that are more sophisticated directly assess the accuracy prediction of each compound, instead of "inside AD/outside AD" information:

• approaches that evaluate the probability distribution of predictions rather than giving point estimates[3,7]

• empirical approaches based on the "distance to model" concept.

The latter approaches are most commonly used in QSAR modeling[8] and represent the subject of the current study. The "distance to a model" (DM) stands for a numerical measure, which monotonically increases as the accuracy of the model decreases.[8] The AD can be defined on the basis of DM; namely, all compounds that have DM values less than a predefined threshold are considered to be inside the AD. The threshold for the DM is chosen to ensure necessary prediction accuracy for compounds within the AD. For predefined prediction accuracy, DMs covering large numbers of molecules are preferred. Leverage, mentioned before, can be used as a distance to the model. In our analysis, we did not fix a "warning leverage" threshold but, rather, investigated the prediction accuracy for all leverage values.

The accuracy of a model can be specified in terms of RMSE, MAE, classification rate, etc. among others. It is worthwhile to distinguish DMs, based solely on descriptor values, from those that use models' predictions, so-called DMs in the property space. To some extent, this terminology may be confusing, since both types of measures solely rely on the structural information. The DMs in the property space can be, of course, applied to new molecules for which experimental values are not known. Both these measures explore disagreements between models developed with different subsets of the initial training data set. To some extent, the DMs in the property space use descriptors that are normalized according to the target property, while the descriptor space DMs ignore this information (e.g., all descriptors are normalized and contribute equally in the LEVERAGE measure). Thus, if some descriptors are more relevant for a given property, they will have higher impact on the DMs in the property space and *vice versa*. As it has been shown,[2] the DMs in the property space yield higher quality AD assessments compared to the DMs in the descriptor space. We confirmed this observation in our previous study[8] and demonstrated that DMs based on the standard deviation of predictions of the model ensemble outperformed descriptor-based DMs such as leverage.

This and several other studies were used for the analysis of classification models which differ from regression-based modeling by the discrete nature of the target (output) labels, which are commonly selected as "−1" (inactive) and "+1" (active; sometimes "0" and "1" are used instead). Interestingly, most machine learning methods, such as neural networks, KNN, or linear regressions, yield continuous predictions. These quantitative values are frequently used to assess classification accuracy, with values close to "−1" and "+1" considered as more reliable predictions than those that are near 0.[9,10] For example, Manallack et al.[9] showed that the classification accuracy of molecules on soluble and insoluble compounds dramatically increased when only molecules with values close to "−1" and "+1" were considered.

In our previous study we introduced a new DM, STD-PROB, which combined measures used by Manallack et al.[9] with the standard deviation of predictions. The latter measure was the best DM criterion for quantitative models.[8]

In the current study, we extend our benchmarking analysis to 30 classification models developed within the 2009 Ames mutagenicity challenge.[11]

METHODS

Data Sets. *The Data Set of the Ames Test Measurements.* The Ames mutagenicity data set[12] described in our previous article[11] was used in the current benchmarking study. The Ames test relies on the determination of the mutagenic effect of a given compound on histidine-dependent strains of *Salmonella typhimurium*. Thus, the measurable mutagenic ability of a compound may signal its potential carcinogenicity.[13] The Ames test can be used with different bacteria strains and can be performed with or without metabolic activation using liver cells. For this study, all such diverse data were pooled together as described in ref 12. According

Table 1. Summary of the Analyzed QSAR Models[a]

model name	descriptors used	training method	numeric predictions	DM provided
CONS			+	
EPA_2D_FDA	2D		+	
EPA_2D_NN	2D	NN	+	
LNU_Drag_PLS	Dragon	PLS	+	
MSU_FRAG_LR	Fragments	Linear regression	+	
MSU_FRAG_SVM	Fragments	SVM	+	SVM1 AD
OCHEM_ESTATE_ANN	E-State indices	ASNN	+	
PCI_Drag_RF	Dragon	Random forest	+	
PCI_SiRMS.Drag_RF	SiRMS+Dragon	Random forest	+	
PCI_SiRMS_RF	SiRMS	Random forest	+	
TUB_3DDrag_RF	Dragon	Random forest		DA Index
TUB_3DDrag_SVM	Dragon	SVM		DA Index
UBC_ID_IWNN	Inductive descriptors	IWNN		
UBC_ID_NN	Inductive descriptors	NN		
UI_Drag_KNN	Dragon	KNN		
UI_Drag_LDA	Dragon	LDA		
ULP_ISIDA_NB	ISIDA Fragments	Naïve Bayes	+	Trust level
ULP_ISIDA_SQS	ISIDA Fragments	Stochastic QSAR sampler	+	Trust level
ULP_ISIDA_SVM	ISIDA Fragments	SVM	+	Trust level
ULP_ISIDA_VP	ISIDA Fragments	Voted Perceptron	+	Trust level
ULZ_3DDrag_KNN	Dragon	KNN		
ULZ_3DDrag_SVM	Dragon	SVM		
UMB_Drag_DT	Dragon	Decision Tree		
UNC_Drag_KNN	Dragon	KNN		
UNC_Drag_RF	Dragon	Random forest	+	
UNC_Drag_SVM	Dragon	SVM	+	AD Mean
UNC_SiRMS.Drag_RF	SiRMS+Dragon	Random Forest	+	
UNC_SiRMS.Drag_SVM	SiRMS+Dragon	SVM	+	AD Mean
UNC_SiRMS_RF	SiRMS	Random forest	+	
UNC_SiRMS_SVM	SiRMS	SVM	+	AD Mean

[a] There were 30 models including the consensus model. The continuous numeric prediction values were available for 20 models.

to that approach, a molecule can be considered as active if it demonstrates mutagenic activity for at least one strain. Thus, considering that the benchmark set molecules were tested with different strains, there may be a significant variance in results. Moreover, different authors used different thresholds to decide whether a given molecule is active or not. As shown in the Results and Discussion section, we estimated the intra- and interlaboratory accuracies of measurements in the Ames mutagenicity data set to be 94% and 90%, respectively.

The initial data set was randomly divided into training and external test sets. The training set contained 4361 compounds, including 2344 (54%) mutagens and 2017 (46%) nonmutagens. The external test set contained 2181 compounds (1/3 of initial set) including 1172 (54%) mutagens and 1009 (46%) nonmutagens. These data sets were used for the 2009 Ames mutagenicity challenge, where the external test set was given to the participants for "blind predictions".[11]

The Data Sets of Chemical Compounds. To investigate the performance of the QSAR models on the Ames test, we have estimated the prediction accuracy for three external data sets: ENAMINE, EINECS, and HPV. The ENAMINE data set contains over *287 000* drug-like chemicals synthesized in 2009 by the Enamine company (http://www.enamine.net). The HPV (high production volume) data set contains chemicals produced or imported into the United States in quantities over 1 million pounds per year. After filtering out composite substances, stereoisomers, and metals from the HPV data set, 2356 compounds were used for analysis. The EINECS (European Chemical Substances Information System) data set was downloaded from http://ecb.jrc.it/qsar/information-sources and contained 68 779 compounds.

Analyzed Models. Twelve international teams submitted 29 models to the 2009 Ames mutagenicity challenge (the models are summarized in Table 1). All of the models were evaluated according to a 5-fold cross-validation procedure as described in the work by Tetko et al.[8] Additionally, each group developed their models using the whole training set, and these models were "blindly" applied to predict test compounds. The resulting consensus model (CONS) was calculated by averaging the predictions of all 29 individual models. The complete information on descriptors, methods, and specific details about each approach can be found elsewhere,[11] while below we will briefly describe the utilized methodologies.

University of Insubria (UI). Linear discriminant analysis (LDA) was used to develop the UI_Drag_LDA model. The LDA calculates a hyperplane, which subdivides the *n*-dimensional descriptor space into two regions corresponding to analyzed classes of compounds. The model was based on 454 Dragon descriptors, which were selected from a total pool of 2032 descriptors after removing constant and highly correlated ($r > 0.9$) descriptors.

Technical University of Berlin (TUB). The Random Forest model (TUB_3DDrag_RF) was a collection of 50 decision trees where each tree depended on a set of randomly selected descriptors.[14] In comparison to the original work of Breimann,[14] all samples were used to build trees (no bagging). The TUB_3DDrag_SVM model was developed using the libsvm[15] implementation with the radial basis kernel. Both of the models were based on 957 3D Dragon descriptors, which were reduced to 872 by removing the descriptors with constant and missing values.

Lanzhou University (LZU). All of the molecules were converted to 3D structures and optimized using MM+ molecular mechanics with semiempirical PM3 partial charges implemented in the HyperChem program (HyperChem for Windows—Molecular Modeling System, Hypercube, Inc., Gainesville, Florida). The Dragon software[16] was used to calculate 1664 molecular descriptors for each molecule. After deleting the descriptors with constant or highly correlated ($r > 0.95$) values, 716 descriptors remained. Support vector machine—recursive feature elimination (LZU_3DDrag_SVM model)[17] was employed to select calculated descriptors and perform classification of the new molecules as described elsewhere.[11] Another model was calculated using the k nearest neighbors method (LZU_3DDrag_KNN).

Linnæus University (LNU). Partial least-squares discriminant analysis (PLS-DA) was used, which is an extension of PLS regression for classification.[18,19] The initial set of descriptors contained 929 2D Dragon descriptors. After removal of 103 constant variables, 826 remained. Nonsignificant descriptors were further removed using a jack-knife method for significance testing of the PLS procedure.[20] Finally, 82 descriptors were used to develop the LNU_Drag_PLS model.

Helmholtz Zentrum Muenchen, Online CHEmical Modeling Environment (OCHEM). The associative neural network method[21,22] was applied using an ensemble of 50 neural networks. Each neural network had three hidden neurons. Both atom- and bond-type 2D E-state indices[23] (362 descriptors) were used for the structure representation. The filtering of highly correlated $r > 0.95$ indices and singletons (found only in a single molecule) left 233 descriptors, which were used to develop the OCHEM_ESTATE_ANN model.

University of British Columbia (UBC). "Inductive" descriptor IND_I[24–26] and MOE QSAR parameters (http://www.chemcomp.com) were used to quantify the structures of the studied compounds. The *in house* SVL scripts were used to calculate IND_I descriptors from 3D structures of molecules optimized with the MOE MMFF molecular force field. All correlated descriptors ($r > 0.9$) were eliminated, and the most relevant QSAR descriptors (15 and 3 descriptors for the IWNN model and NN models, respectively) were selected according to the Information Gain criteria[27] using Weka software (v. 3.5.8).[28] The weighted nearest neighbor (UBC_ID_NN) and iterative weighted nearest neighbor (UBC_ID_IWNN) models were created as described elsewhere.[11]

Laboratory of Chemoinformatics, Institute of Chemistry, Louis Pasteur University, Strasbourg, France (ULP). Two classes of substructural molecular fragments, "sequences" (I) and "augmented atoms" (II), were used.[29] The ULP_ISIDA_NB model was developed using naive Bayesian approach. The ISIDA/VotedPerceptron ULP_ISIDA_VP model implemented a simple perceptron algorithm re-expressed in terms of the Tanimoto kernel. For nonlinearly separable cases, all perceptrons were combined in a voting pool. The weighted vote was done according to the accuracy of perceptrons for the training set. The ULP_ISIDA_SVM model used libSVM with the Tanimoto similarity coefficient as a kernel. ULP_ISIDA_SQS was created using the stochastic QSAR sampler (SQS) algorithm, which is a genetic algorithm-driven regression tool supporting nonlinear descriptor transformations.[30]

Moscow State University (MSU). The ν-modification support vector machines method[31] and regularized logistic regression implemented in the package LIBLINEAR[32] were used to develop the MSU_FRAG_SVM and MSU_FRAG_LR models, respectively. Optimal values of algorithm parameters were found using the grid search and the cross-validation procedure. Both of the models used the same set of 19 603 fragmental descriptors[33,34] with the size of each fragment up to five non-hydrogen atoms, which were computed using the NASAWIN software.[34] No descriptor selection procedures have been applied.

Physico-Chemical Institute of NAS of Ukraine (PCI). The Simplex representation of molecular structure (SiRMS)[35] was used to calculate 21 378 2D Simplex descriptors (number of tetra-atomic fragments with fixed composition and topological structure).[35,36] In addition, 2D Dragon descriptors (943) were used separately and in combination with Simplex descriptors. The Random Forest (RF)[14] method was employed for obtaining models.[37] The final models have been selected by the highest out-of-bag statistic values. The PCI group contributed three models, based on SiRMS descriptors (PCI_SiRMS_RF), 2D Dragon descriptors (PCI_Drag_RF), and a combination of both (PCI_SiRMS.Drag_RF).

University Milano—Bicocca (UMB). A total of 2489 molecular descriptors[16] were calculated using the Dragon software.[38] Constant and nearly constant descriptors were removed, leading to a final number of 1601 retained descriptors. The CART (Classification and Regression Trees) algorithm, a binary tree classification method,[39] was used to develop the decision trees UMB_Drag_DT model. The final classification tree included 29 descriptors.

University of North Carolina (UNC). The WinSVM program implementing the open-source libSVM package[40] was employed to build and select mutagenicity models. An ensemble of 467 Dragon descriptors calculated for two-dimensional hydrogen-depleted structures, 609 two-dimensional SiRMS descriptors, and a combined set of Dragon/SiRMS descriptors were used as inputs to build the UNC_DRAG_SVM, UNC_SiRMS_SVM, and UNC_SiRMS.DRAG_SVM models respectively.

United States Environmental Protection Agency (EPA). A total of 790 2D descriptors[41] were used. The EPA_2D_FDA model (FDA, Food Drug Administration) was built according to a methodology developed by Contrera et al.[42] For each test chemical, 30—75 of the most similar chemicals from the training set in terms of the cosine similarity coefficient were selected. Then, a local linear regression model was built to predict the test compound. In the EPA_2D_NN model, the three closest chemicals in the training set in terms of the cosine similarity coefficient were selected for the test compound. The predicted mutagenicity was simply the class, which dominated for the three chemicals.

Additional details of models and a detailed description of models and results can be found elsewhere.[11]

Preprocessing of Results. Some models provided prediction as {0,1} while other models provided it as {−1,+1} for mutagenic and nonmutagenic compounds, respectively. In order to be consistent, we converted all predictions to the {−1,+1} values. After this processing, for some models, i.e., neural networks or SVM, there were predictions outside of the [−1,+1] interval. We did not normalize or round these

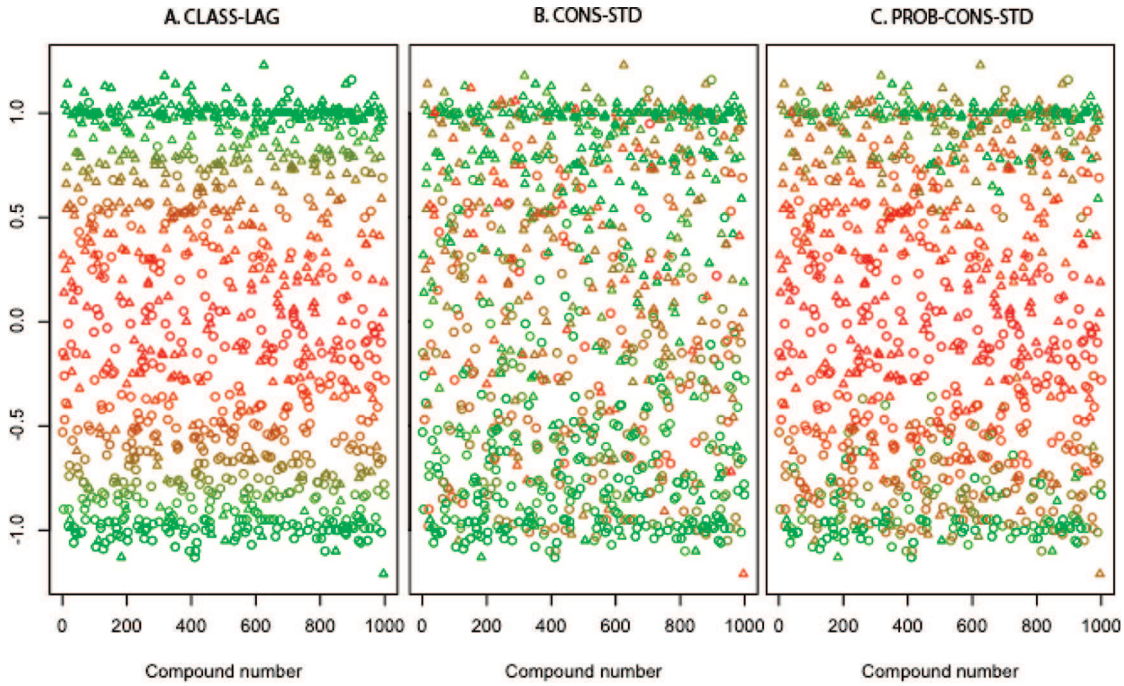


Figure 1. Test set predictions of the OCHEM_ESTATE_ANN model. Three DMs (CLASS-LAG, CONS-STD, and PROB-CONS-STD) are encoded by color. Green represents low values of the corresponding DM; red represents high values. Triangles are mutagens, and circles are nonmutagens according to the Ames mutagenicity test. Values outside the [−1,+1] interval appear due to a specific normalization for neural network training (value −1 corresponded to 0.1, and +1 corresponded to 0.9).

values to [−1,+1]; instead, we used the original values for the calculation of DM as described below.

Numeric prediction values were available only for 20 models (including the consensus model). As some of the investigated DMs require numeric prediction values, only these 20 QSAR models were used in the current study. Several DMs that could be used only with qualitative predictions were applied to all 30 models.

Distance to Model and Applicability Domain. Let us designate any numeric measure calculated solely on the basis of chemical structures or prediction values and which increases with a decrease in the reliability of classification as "distance to model" (DM). Then, on the basis of a model performance, we can identify a threshold for the DM that provides a predefined accuracy of classification. All data set entries with DM values below the threshold form a model's "applicability domain" (AD). Criteria for the performance of distances to the model are suggested in the section below.

Most DMs investigated in this article are developed on the basis of those used previously for regression problems[2,8,43] and were introduced in our preliminary study.

Let us introduce notation to represent predictive modeling entities: J, a compound to be predicted; $y(J)$, a continuous prediction value, calculated by the model; $c(J)$, the predicted class for the given compound J, identified by:

$$c(J) = \begin{cases} 1, & y(J) > 0 \\ -1, & y(J) \le 0 \end{cases} \quad (1)$$

We will designate DM for a compound J as $d(J)$.

CLASS-LAG. For the binary classification problem, labels for the predictive model are discrete and are selected in our study as −1 and +1. However, most machine learning methods give a quantitative number as a result of prediction. The absolute value of the difference between the prediction value and the nearest of the labels can be used as a measure of prediction uncertainty. This measure, referred to as CLASS-LAG, is calculated according to

$$d_{\mathrm{CLASS-LAG}}(J) = \min\{|-1 - y(J)|, |1 - y(J)|\} \quad (2)$$

CLASS-LAG can be interpreted as the amount of rounding to the nearest class label; the more rounding that is required, the less reliable the prediction is expected to be. Thus, the measure punishes deviations from target class values {−1,+1}, both positive and negative deviations (i.e., both 1.2 and 0.8 predicted values have the same DM). Obviously, punishing negative deviations applies only to models that have prediction values outside of the [−1,+1] interval; there were only three models with such predictions: EPA_2D_FDA, LNU_Drag_PLS and OCHEM_ESTATE_ANN.

Figure 1A illustrates the simplicity of this idea: green dots, which are closer to the edge of the class, are predicted to have better prediction accuracy than red dots, located in the "uncertainty area" between the classes, near a value of 0. In this figure, triangles are positive (mutagens) and circles are negative (nonmutagens) predictions. The classes are more mixed together near zero line. The continuous values of predictions may not always be available: some machine learning methods provide only discrete {−1,+1} outputs. In this case, CLASS-LAG is always equal to zero and obviously cannot be used. This DM is the most obvious one, and it was used, e.g., by Mannalack et al.[9]

STD. The standard deviation of the predictions, obtained from an ensemble of models, can be used as an estimator of model uncertainty for a given compound. The general idea is that if different models yield significantly different predictions for a particular compound, then the prediction for this compound is more likely to be unreliable. The sample standard deviation can be used as an estimator of model uncertainty.

Let us assume that $Y(J) = \{y_i(J), i = 1-N\}$ is a set of predictions for a compound J given by a set of N trained models. The corresponding distance to model (STD) is calculated by

$$d_{\text{ASNN-STD}}(J) = \text{stdev}(Y(J)) = \sqrt{\frac{\sum (y_i(J) - \bar{y})^2}{N - 1}} \quad (3)$$

This DM has been proven to provide excellent results for the discrimination of highly accurate predictions in the case of regression models.[2,8,9] In the given study, we investigate two variations of the STD measure that differ in the contents of the used models: (i) ASNN-STD, based on predictions of a neural network ensemble of OCHEM_ESTATE_ANN, and (ii) CONS-STD, based on predictions of several models that were built using different machine learning methods and different parameters (and including OCHEM_ESTATE_ANN as one of the models).[2] Although it is possible to calculate STD for virtually any model, i.e., by replicating multiple models of the same method using the bagging technique[44] and computing the standard deviation of predictions, in this study, STD values were available only for OCHEM_ESTATE_ANN.

In our study, we used two variations of this measure: CONS-STD uses quantitative values of predictions to calculate standard deviation, and CONS-STD-QUAL uses qualitative (discretized) values. The rationale for using CONS-STD-QUAL lies in the unavailability of quantitative values for some machine learning methods.

Applicability of the standard deviation to classification tasks follows from the property of the Bernoulli-distribution, which is used for characterizing the distribution of random binary values. The standard deviation of the Bernoulli distribution rises as the probability for each class approaches 0.5, which corresponds to the most uncertain prediction. Hence, both the normal (used in regression tasks) and Bernoulli distributions (used in classification tasks) follow the same law—the prediction uncertainty rises as the standard deviation rises.

In Figure 1B, built on the basis of the OCHEM_ESTATE_ANN model, the green dots denote the highest level of agreement between 20 individual models, used for CONS-STD. These points correspond to low values of standard deviation. The red points, on the contrary, show that the individual models yielded quite a wide range of predictions; so the standard deviation for these points is relatively high. In this figure, we observe that red and green points are mixed, which means the STD measure does not depend on the value of prediction and can provide information that is complementary to CLASS-LAG.

STD-PROB. This DM, suggested in a recent study,[45] combines the two previously mentioned measures into a single value to improve the estimation of prediction accuracy. Having obtained a prediction $p(x)$ for a given compound, we replace this point prediction with a distribution of probabilities. In other words, instead of giving a point prediction, we provide a probabilistic one. We assume the mentioned distribution is Gaussian with a mean $p(x)$ and standard deviation that correspond to its STD value. The suggested distance to model is

$$d_{\text{STD-PROB}}(J) = \min\begin{cases}\text{probability}(c > 0 | N(y(J), d_{\text{STD}}(J)))\\ \text{probability}(c < 0 | N(y(J), d_{\text{STD}}(J)))\end{cases} \quad (4)$$

namely,

$$d_{\text{STD-PROB}}(J) = \min \begin{matrix}\int_0^{+\infty} N(x, y(J), d_{\text{STD}}(J))\, dx\\ \int_{-\infty}^{0} N(x, y(J), d_{\text{STD}}(J))\, dx\end{matrix} \quad (5)$$

where $N(x,y(J),d_{\text{STD}}(J))$ is the normal distribution density function with mean $y(J)$ and standard deviation $d_{\text{STD}}(J)$. Here, $y(J)$ is an actual prediction of the analyzed model for a compound J and $d_{\text{STD}}(J)$ is an STD-based distance to model (ASNN-STD or CONS-STD), calculated according to eq 3.

This measure can be graphically illustrated as the square of the area under the curve of the normal distribution density function.

Four examples are given in Figure 2, where the rounded prediction value is always fixed to "+1"; however, the quantitative prediction values and STD values are different. It is obvious that shifting the curve away from the center (decreasing CLASS-LAG) results in a decrease of the filled area. The same effect appears when we make the curve less flat, i.e., decrease the STD value. Thus, STD-PROB combines information about uncertainty from both measures: CLASS-LAG and STD.

STD-PROB has an easy interpretation: values close to 0.5 indicate an equal probability of finding the given compound in either class; i.e., the model cannot provide reliable prediction. On the contrary, values close to 0 indicate a high probability of finding the compound in one of the classes.

We analyze two variations of STD-PROB, ASNN-STD-PROB and CONS-STD-PROB, which correspond to the ASNN-STD and CONS-STD measures, respectively.

Similarly to the depiction of previously introduced measures, green dots in Figure 1C denote compounds whose CONS-STD-PROB value, i.e., minimal area under the probability density chart on the intervals $(-\infty; 0]$ and $[0; +\infty)$, is relatively high. The square of this area is computed using eq 4.

Importantly, the STD-PROB measure is an empirical one. This approach proved to work successfully in our previous study using ensembles of neural networks,[45] and it is applied here to analyze the results produced by other machine learning methods.

CONCORDANCE. This measure shows whether a prediction of an individual model is concordant with predictions of other models within the ensemble. More accurately, CONCORDANCE is the number of models that give the same prediction that the current model does:

$$\text{CONCORDANCE}(J) = \sum_{i=1}^{N} \text{eq}(y(J), y_i(J)) \quad (6)$$

where $y(J)$ and $y_i(J)$ are predictions of compound J, given by the target model and the members of the ensemble, N is the size of the ensemble, and eq is equality indicator (equal to 1 if the arguments are equal and to 0 otherwise).

CORREL. This measure is based on the correlation of vectors of the ensemble's predictions for the target compound and compounds from the training set. More precisely, the

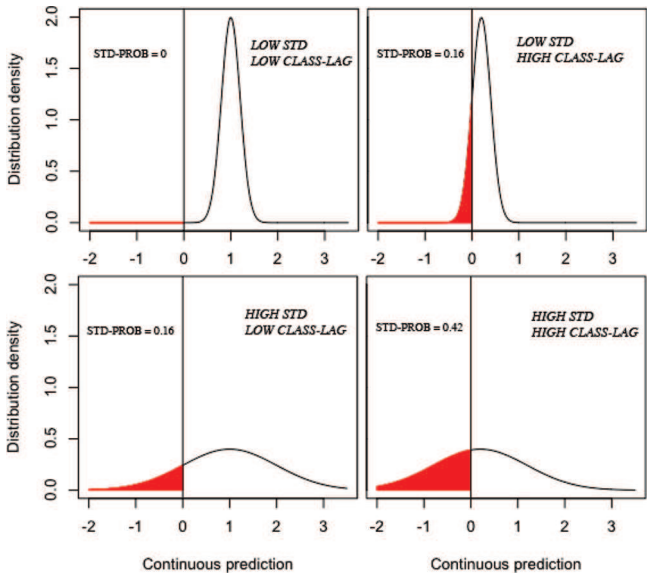


Figure 2. STD-PROB is the square of the filled area on each of the four charts. The charts show how CLASS-LAG and STD affect STD-PROB. STD corresponds to the flatness of the curve, and CLASS-LAG corresponds to the shift of the curve from the center. Larger values of STD correspond to flatter curves and larger STD-PROB values. As CLASS-LAG decreases, the curve shifts more from the center and the STD-PROB value decreases.

CORREL measure for the target compound J is calculated according to the following expression:

$$\mathrm{CORREL}(J) = 1 - \max_{i=1-M} |\mathrm{corr}(\vec{y}(T_i), \vec{y}(J))| \qquad (7)$$

where $\vec{y}(T_i)$ and $\vec{y}(J)$ are vectors of the ensemble's predictions for the training set compound T_i and the target compound J, and corr designates the Spearman rank correlation coefficient between the two vectors, and M is the number of compounds in the training set. The low value of CORREL (i.e., high Spearman correlation coefficient) indicates that for target compound J there is a compound T_k from the training set for which predictions of the ensemble of models are strongly correlated. Indeed, if a compound T_k has the same descriptors as J, then the predictions of the models will be identical for both molecules, and thus CORREL(J) = 0. The performance of this measure for regression models is discussed elsewhere.[8,46]

LEVERAGE. Leverage is a descriptor-based DM; i.e., it is based only on model input but not on output, in contrast to CLASS-LAG, STD, and STD-PROB. LEVERAGE is a special case of Mahalanobis distance, calculated according to expression 8:

$$\mathrm{LEVERAGE}(J) = x(X^T X)^{-1} x^T \qquad (8)$$

where x is a vector of descriptors for compound J and X is the matrix of descriptors for the training set. The LEVERAGE values were available only for the OCHEM_ESTATE_ANN model and were based on E-State indices.

DA-Index. The applicability domain employed by the TUB group is based on the κ, γ, and δ indices introduced by Harmeling et al.[47] The first two indices are heuristics that have been previously used in the chemoinformatics community: κ is the distance (here in this section and below, Euclidian distance calculated using descriptors is assumed) to the k-nearest neighbor, and γ is the mean distance to the k nearest neighbors. The last index, δ, corresponds to the length of the mean vector (i.e., a mean of vectors) to the k nearest neighbors. Since κ and γ are only based on distances, they do not explicitly indicate whether interpolation or extrapolation is expected for prediction. δ allows making this distinction and indicates the degree of extrapolation. Input descriptors for all indexes were weighted following the development of the Gaussian process classification model.[48] The arithmetic mean values of γ and δ indices were used to estimate prediction confidence. A threshold value determined using the training set was used to decide whether a test compound was inside or outside the AD. The output of this decision process was called DA-Index.

AD_MEAN. AD_MEAN values were provided by the UNC group for SVM models that were developed using three sets of descriptors (SiRMS, Dragon, and combined). AD_MEAN corresponds to the average Euclidean distances between a compound and its three nearest neighbors in the training set. All distances are calculated using the entire pool of descriptors. AD_MEAN was available for two models, UNC_SiRMS_SVM and UNC_Drag_RF; therefore, we investigated two respective measures, AD_MEAN1 and AD_MEAN2.

ELLIPS. ELLIPS values were calculated using the EPA_2D_FDA model. A prediction is within the applicability domain of the model if the test chemical is within the multidimensional ellipsoid defined by the ranges of descriptor values for the chemicals in the cluster (for the descriptors appearing in the cluster model). The model ellipsoid constraint is satisfied if the leverage of the test compound (h_{00}) is less than the maximum leverage value ($h_{\max}$) for all of the compounds used in the model.[49] The ratio $h_{00}/h_{\max}$ was used as a distance to the model, referred to as ELLIPS.

SCAvg (Average Similarity Coefficient). The cosine similarity coefficient to the three nearest neighbors used in the EPA_2D_NN method was used as the SCAvg DM.

Two groups classified predicted molecules in several classes with different qualities of prediction as described below.

Trust Level. The applicability domain for the models, provided by the ULP group, is based on a measure, referred to as the *trust score*. This measure has values in the range of {1,2, ...5}, where the "5" corresponds to the highest trust level ("optimal") and 1 is the lowest trust level ("none"). The trust score for a particular compound is based on three factors: (i) the number of models having the compound in their local applicability domain, MINDIFF-OK, as described in ref 50, (ii) the number of dissident predictors in the set (i.e., models that gave predictions, different from the mean prediction), and (iii) the average prediction value, where values close to 0.5 are considered less reliable. Further details on the calculation of the trust score are shown in Figure SF1 (Supporting Information).

SVM1 AD. The applicability domain for the MSU models was computed using the one-class classification approach (novelty detection) based on 1-SVM.[51] The parameters of the 1-SVM method were chosen as follows: the RBF-kernel parameter γ was taken from the same value used for building classification SVM models, while the value of ν was fixed at 0.01.

The SVM1 AD procedure associates the applicability domain of QSAR/QSPR models with the area in the input

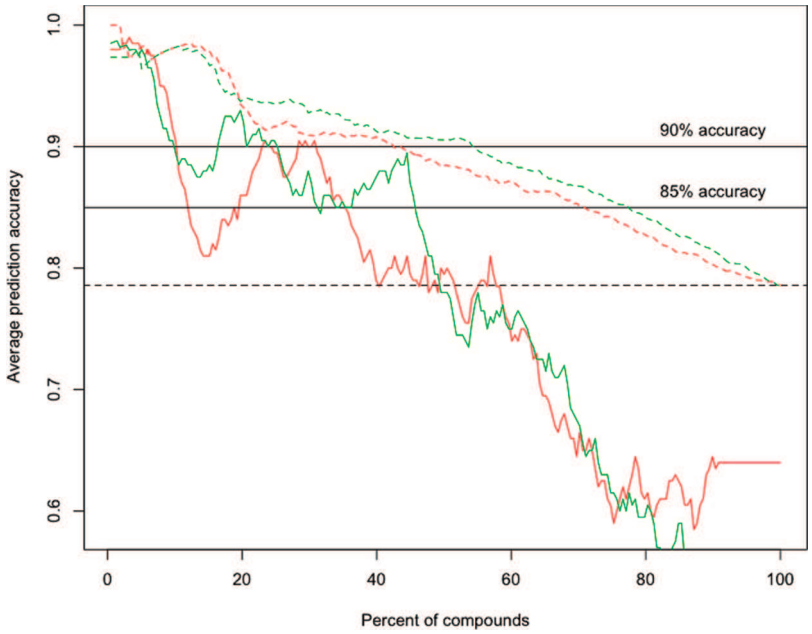


Figure 3. Prediction accuracy of the consensus model as a function of CONS-STD and CONS-STD-PROB. The solid lines (bin-based averaging) show the averaged accuracy on a moving window with a size of 200 compounds. Although there is a trend that the accuracy of prediction decreases with both DMs, the dependency is not smooth, and there are significant fluctuations. The dashed lines (cumulative averaging) indicate the average prediction accuracy for a variable percentage of compounds. Cumulative averaging smooths the variations, which makes it more suitable for the threshold-based comparison of DMs.

descriptor space where the density of training data points exceeds a certain threshold. The main assumption of this procedure is that the predictive performance of the models tends to be higher for the test compounds inside the high density areas than for those that are outside. This could take place since outside the high density area all test objects are located far from training objects, which makes interpolation of the properties from the training to test objects unreliable. Instead of searching a decision surface separating high and low density areas in the input space, the one-class classification 1-SVM approach looks for a hyperplane in the feature space associated with the RBF-kernel.

The ability of novelty detection models to be used as the AD of machine learning models was earlier demonstrated by Bishop.[52] The use of a one-class SVM novelty detection method to assess the applicability domain of models based on structured graph kernels has recently been suggested by Fechner et al.[53]

Benchmarking Criteria. To compare the performances of different DMs, it is necessary to assess their ability to separate predictions with low and high accuracy. Our approach is to determine the percentage of compounds in the training and test sets that are predicted with a DM-defined accuracy. For a particular DM, there are two possible ways to separate compounds, predicted with a given accuracy:

Bin-Based Accuracy Averaging (BBA). BBA groups the compounds, sorted by a particular DM, into bins having an equal number of compounds, averages the accuracy in the bins, and selects a DM threshold, which provides predefined model accuracy for every bin within this threshold. However, this criterion has some drawbacks. First, it does not take into account the actual prediction accuracy as long as it is higher than the threshold. Second, the detection of a DM threshold in practice can be a subjective task and will depend on the size of the bin. For example, when predictions for different models were sorted according to DMs, and their accuracies were averaged using a sliding window of, e.g., 200 molecules, we could observe a significant variation in predictions as a function of the DMs when using one defined threshold (see solid lines in Figure 3).

Integral Accuracy Averaging (IA). Instead of bin-based averaging, one can use the average accuracy of a model for molecules with a DM less than a predefined threshold value. The plots of average predictions of models for a DM less than the predefined threshold are smoother and easier to interpret: i.e., this threshold defines the average (accumulative) accuracy of the model. Moreover, this criterion directly corresponds to, e.g., the average accuracy of inter- or intralaboratory measurements. Therefore, for all further analyses, we used the integral criterion and compared the DMs with respect to their accumulative average accuracy. A threshold of 90% was used. More precisely, we did the following steps to estimate the performance of the investigated DMs:

- For the training and test sets, sort all of the compounds according to DM.
- For each model, identify the largest DM value for which the accumulative accuracy of compounds from the analyzed set (training or test) is ≥ 90% (DM_90%).
- For each model, calculate the percentage of compounds with a DM less than the *respective* DM_90% threshold for the training set (referred to as $C_{\text{TRAIN-90\%}}$, C stands for coverage) and the test set ($C_{\text{TEST-90\%}}$). Notice that thresholds are selected separately for the training and test sets.

Values $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ are used to estimate performances of the DMs for each analyzed model. Indeed, for a given model, the larger $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ values correspond to DMs with larger numbers of reliable predictions. Similar to our previous study,[8] we ranked DMs according to their $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ values (i.e., the DM with the highest $C_{\text{TRAIN-90\%}}$ or $C_{\text{TEST-90}}$ receives a rank

of "1" and so on) and averaged the ranks over all models. These averaged ranks were used to compare different DMs.

Under prediction accuracy, we understand

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total number of compounds}} \times 100 \quad (9)$$

where true positives, true negatives, and the total number of compounds are within a DM threshold. In addition to the prediction accuracy, the sensitivity and the specificity are frequently used in machine learning methods. These measures are particularly useful for nonbalanced data sets. The Ames data set has a very small imbalance of active and nonactive compounds; therefore, specificity and sensitivity are to a large extent redundant and were not analyzed in this study.

To verify whether there are significant differences between analyzed DMs, we used the Wilcoxon signed-rank test[54] applied to $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ values. This test is used for two-sample designs involving repeated measures, matched pairs, which is the case in our study.

As a graphical illustration of the DM performance, we used cumulative accuracy−coverage plots (see, e.g., dashed lines in Figure 3). On these charts, we plotted prediction accuracy for a group of compounds, having a DM less than some threshold (y axis), against a percentage of this group of compounds in the whole set (x axis). The threshold for DM is not directly present in the chart but is implicitly represented by the x axis.

Additionally, we intended to confirm whether a particular DM can not only separate high and low accuracy predictions but also estimate the external accuracy of prediction. For this purpose, we compare prediction accuracies for compounds within the *same* DM threshold on training and test sets.

There are two drawbacks to the aforementioned accuracy coverage ($C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$) as an estimator of the DM performance. First, the coverage depends on the accuracy threshold, and different thresholds could possibly result in different rankings of the analyzed DMs. Second, the accuracy coverage depends not only on the ability of DM to separate highly accurate predictions but also on the performance of the analyzed model. Indeed, the models having higher prediction accuracies will probably have higher accuracy coverages.

The AUC (Area under the Curve) Criterion. Another criterion for DM performance that does not have the aforementioned drawbacks is the area under the curve (AUC) parameter, calculated as the area of the square between the bin-based averaging curve and the line of the average model performance. In Figure 3, this is the area of the square between one of the solid lines and the dashed horizontal line. The AUC is higher for the DMs that provide better separation of compounds with higher and lower accuracies compared to the average accuracy of models. Similarly to the accuracy coverage, the weighted accuracy spread can be calculated for both the training set ($\text{AUC}_{\text{TRAIN}}$) and the test set ($\text{AUC}_{\text{TEST}}$).

To rank the investigated DMs, we used both criteria: the accuracy coverage ($C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$) and the area under the curve ($\text{AUC}_{\text{TRAIN}}$ and AUC_{TEST}).

Table 2. Average Ranks of the DMs Ranked by the Percentages of Compounds with 90% Accuracy[a]

distance to model	average rank (cTRAIN 90%)	average rank (cTEST 90%)
CONS-STD-QUAL-PROB	2.15	1.83
CONCORDANCE	1.65	2.15
CONS-STD-PROB	3.38	2.95
CONS-STD-QUAL	3.7	4.95
ASNN-STD-PROB	6.4	5.48
CONS-STD	4.88	5.75
CLASS-LAG	7.5	6.68
ASNN-STD	8.4	7.78
ELLIPS	9.15	8.98
AD_MEAN1	12.43	10.18
CORREL	10.35	11.65
SCAvg	11.08	11.85
AD_MEAN2	11.3	12.33
LEVERAGE	12.65	12.48

[a] The ranks for both the training and validation sets are shown.

Comparison of Models. The most commonly used measure of model performance is its accuracy on the test set. This measure, however, does not reveal what is the maximum possible performance of a particular model. For this reason, a percentage of compounds that are predicted with a fixed accuracy level (90% in our example) can be identified and used for model ranking.

RESULTS AND DISCUSSION

Comparison of Distances to Model. The calculated $c_{\text{TRAIN-90\%}}$ and $c_{\text{TEST-90\%}}$ values are summarized in Table 2, where DMs are sorted accordingly to their rank on the basis of $c_{\text{TEST-90\%}}$ values (see Table S1 of the Supporting Information for more details). The data demonstrate that the CONS-STD-QUAL-PROB measure appeared to be the best one, considering averaged ranks over all models on the test set. Details for the calculation of averaged ranks can be found in the Supporting Information in Table S1 (part B). According to the Wilcoxon test,[54] the top three models (CONS-STD-QUAL-PROB, CONDCORDANCE, and CONS-STD-PROB) were not significantly different from each other, with $p > 0.05$ for both analyzed sets, but were significantly better ($p < 0.05$) than other investigated measures. The LEVERAGE distance could not separate 90% accuracy predictions for any model ($c_{\text{TEST-90\%}} = c_{\text{TRAIN-90\%}} = 0$); therefore it was not analyzed further.

The rankings based on the accuracy coverage (Table 2) are not significantly different from those based on the AUC (Table 3). Namely, the rankings changed for the four last DMs (LEVERAGE, SCAvg, CORREL, and AD_MEAN2), which were however not significantly different from each other. One difference of the AUC rankings from the accuracy coverage rankings is that, according to the AUC criterion, CLASS-LAG outperformed the ASNN-STD-PROB. For all further analysis, we used the accuracy coverage ($c_{\text{TRAIN-90\%}}$ and $c_{\text{TEST-90\%}}$) because of its simpler and more intuitive interpretation.

According to the PCA plot in Figure 4, some of the models were quite similar, since they were based on the same descriptors and machine-learning methods, e.g., UNC_Drag_RF and PCI_Drag_RF, PCI_SiRMS_RF, and UNC_SiRMS RF. Combining these four models into two did not

Table 3. Averaged Rankings of the DMs Ranked by the AUC Criterion

distance to model	average rank (AUC, training set)	average rank (AUC, test set)
CONS-STD-QUAL-PROB	2.15	1.95
CONCORDANCE	1.4	2.1
CONS-STD-PROB	3.4	2.75
CONS-STD-QUAL	3.8	4.9
CLASS-LAG	6	4.95
ASNN-STD-PROB	6.4	5.65
CONS-STD	5.3	6.1
ASNN-STD	8.05	7.9
ELLIPS	12.1	9.6
AD_MEAN1	10.9	11.25
LEVERAGE	12.85	11.3
SCAvg	11.6	11.7
CORREL	9.95	11.85
AD_MEAN2	11.1	13

affect the sorting of compounds according to the DMs. Therefore, the rankings of the DMs, given in Tables 2 and 3, were not affected.

The dependency of the model performances for the CONS-STD-PROB DM is shown in the cumulative accuracy−coverage plot (Figure 5). The plot indicates that 25−70% of all compounds (depending on the model) are predicted with 90% accuracy. The same kind of plot for the CLASS-LAG DM (Figure 6) reveals poorer performance of the latter measure when it is not used in combination with the STD measure. The difference is visually apparent: for some of the models, CLASS-LAG was not able to separate predictions with 90% accuracy; in Figure 6, these models correspond to curves under the 90% line.

The performance of the CLASS-LAG DM appeared to be very dependent on the model, as can be observed in Figure 6 and in Table S1 (Supporting Information). This can be explained by different distributions of quantitative values of predictions, given by different models. Two histograms in Figure 7 reveal that the prediction values of UNC_SiRMS_SVM are similar to discrete values {−1,+1}; therefore, they contain less information than the predictions by PCI_SiRMS.Drag_RF, which are distributed more uniformly. Indeed, the CLASS-LAG DM failed for the first model, $c_{\text{Test-90\%}} = 0\%$ coverage, and yielded excellent results for the second one, $c_{\text{Test-90\%}} = 62\%$ coverage.

As described in the Methods section, CLASS-LAG punishes both negative and positive deviations from the class labels {−1;+1}. Thus, predictions outside of the [−1,+1] interval (referred to as "outer predictions") are considered less reliable than the exact −1 or +1. There were three models with outer predictions: EPA_2D_FDA, LNU_Drag_PLS, and OCHEM_ESTATE_ANN. When we rounded the outer predictions to {−1;+1} labels, their performance for CLASS-LAG did not change significantly from those for LNU_Drag_PLS and OCHEM_ESTATE_ANN; however, the performance significantly dropped for the EPA_2D_FDA model (see Figure SF2 in the Supporting Information).

The percentage of active (mutagenic) compounds within the range of 90% prediction accuracy is 51−55% and is not significantly different from the percentage of active compounds in the whole test set (53%). Therefore, mutagenic compounds are neither over-represented nor under-represented in the applicability domain of the models. Moreover, the prediction accuracy, sensitivity, and specificity of all of the models were not significantly different within the area of 90% prediction accuracy. Thus, the analysis of specificity and sensitivity is redundant; therefore, we used only prediction accuracy, calculated according to eq 9.

Several distances to the model were investigated in our study. A recently introduced probability-based measure of distance to a binary classification model, CONS-STD-PROB,[45] as well as its qualitative analog, CONS-STD-QUAL-PROB, provided a significantly better separation ($p < 0.05$ using the Wilxocon test) of predictions with low and high accuracy. Therefore, the quality of applicability domain estimation, using these methods, is significantly better than

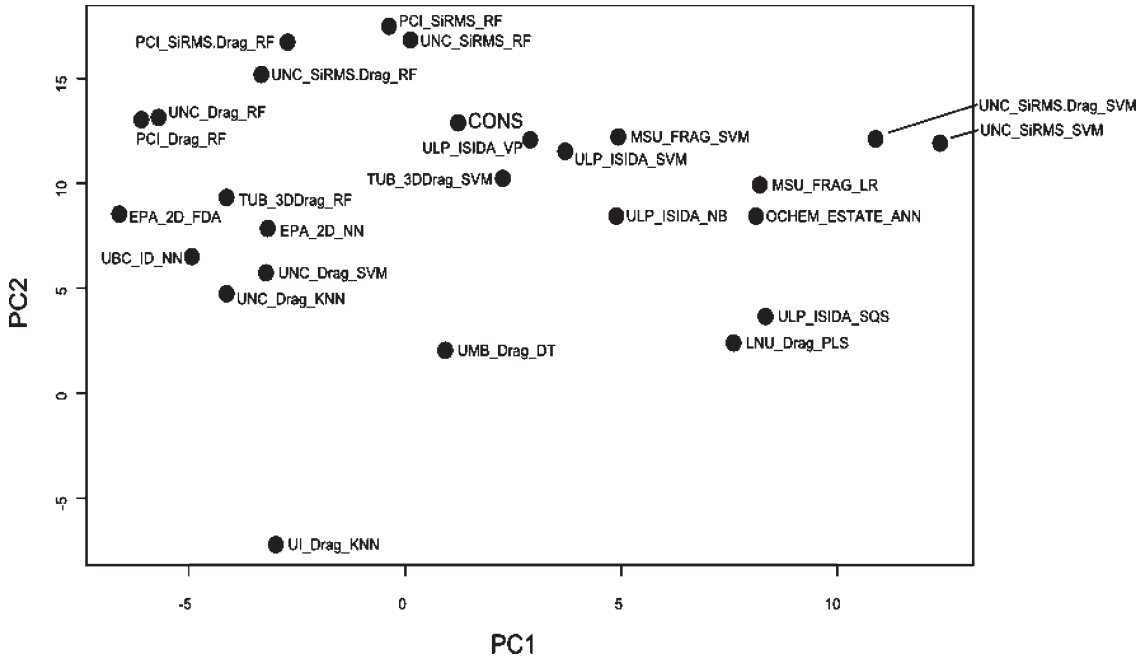


Figure 4. PCA plot of the Ames challenge models, based on the space of predictions for the test set. Four models (UI_Drag_LDA, UBC_ID_IWNN, ULZ_3DDrag_SVM, and ULZ_3DDrag_KNN) are not shown, since they were outliers of this graph.

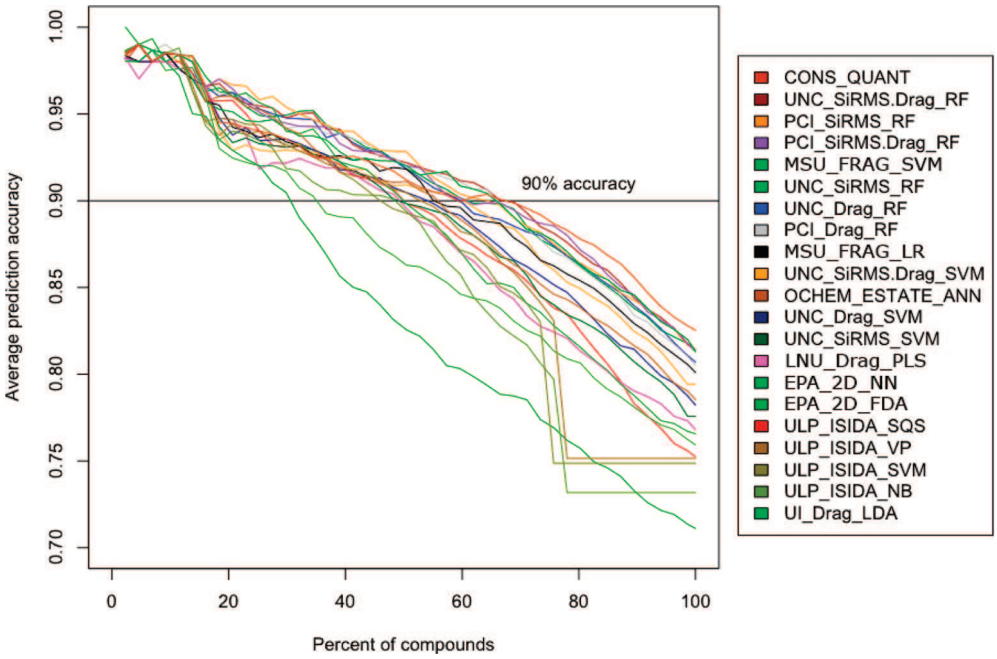


Figure 5. Cumulative accuracy−coverage plot for the CONS-STD-PROB DM based on the test set predictions. Only those 20 models are shown which had numeric prediction values available. The curves show the accumulative accuracy for a particular (variable) percentage of compounds. The curves clearly show that CONS-STD-PROB is highly correlated with the prediction accuracy. The models are ordered according to their overall performance for the test set.

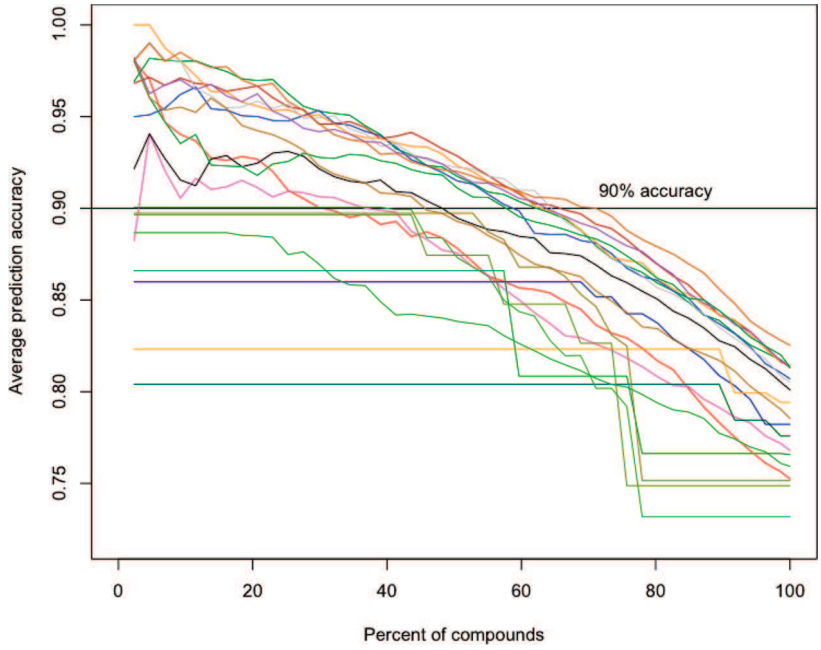


Figure 6. Cumulative accuracy−coverage plot for the CLASS-LAG DM. The plot is based on the test set predictions. The colors of the models are the same as in Figure 5.

that of the traditionally used CLASS-LAG method. It is interesting that CONCORDANCE, i.e., the measure of an agreement of predictions of a considered individual model with other members of the ensemble, was also amid the top three models and provided the best results for the training set. Therefore, it may be reasonable to use this simple measure along with the STD-PROB DMs.

The distances to the model, based on the space of descriptors (LEVERAGE, DA Index, ELLIPS, SCAvg, and AD_MEAN) identified only very small percentages of molecules with >90% accuracy (see Table 2 and Table S1, Supporting Information) and thus performed worse compared to other DMs considered in this study. The measures on which DA Index was based (namely, δ index and γ index) did not outperform DA Index when used as stand-alone DMs; therefore, they were not analyzed. Figure 8 (solid lines) demonstrates AD_MEAN results, which are worse compared to those of CONS-STD-PROB (dashed lines), which identified more than 40% of compounds as having this prediction accuracy for analyzed models.

The PCA plot of the DMs (Figure 9) calculated using the DM-based rankings of Ames challenge compounds reveals high similarity of the five DMs, which are based on the global consensus model. Indeed, these models explore

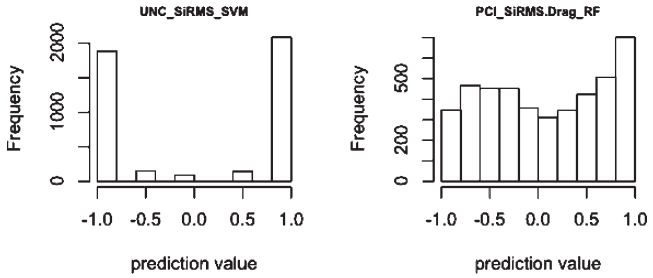


Figure 7. Distribution of prediction values for the two selected models. The prediction values of the model on the left chart resemble rounded discretized "−1" and "+1" values, whereas the values on the right chart have a continuous distribution and therefore provide more information for the estimation of uncertainty. This fact is confirmed in practice: CLASS-LAG of UNC_SiRMS_SVM (left chart) has poor performance (0% coverage of 90% accuracy) in contrast to PCI_SiRMS.Drag_RF (right chart), which separates 63% of compounds with 90% prediction accuracy.

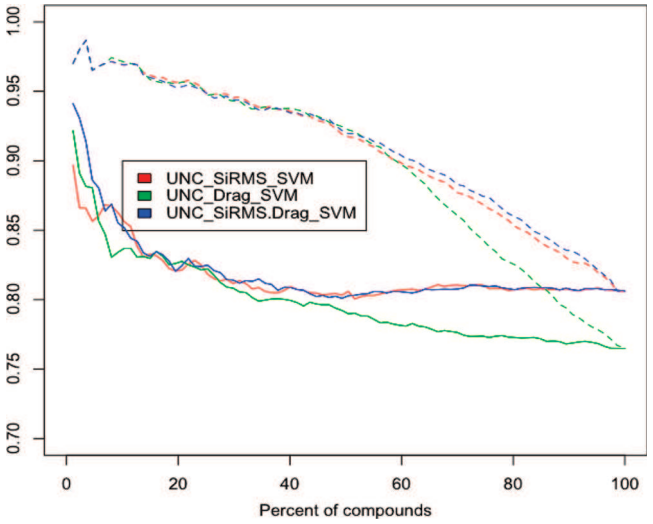


Figure 8. Comparison of AD_MEAN distance to the model (solid line) with CONS-STD-PROB (dashed line).

slightly different aspects of the same data and are strongly intercorrelated (see Table S5 in the Supporting Information). The CONS-STD, CONS-STD-QUAL, and CONCORDANCE DMs form one cluster within which the CONCORDANCE DM provided the best discrimination of the highly accurate predictions (Tables 2 and 3).

Analysis of the Qualitative Distances to Models. As mentioned in the Methods section, several groups provided qualitative AD measures for their respective models. The performance of CONST-STD-PROB for these models binned on several intervals is shown in Table 4 and is compared to the aforementioned models in this section.

Trust Level. This AD-related information, provided by the ULP group, is a generic estimation of the degree of trust for the prediction of a particular compound, ranging from optimal to poor, depending on how concordant individual models were in the prediction of this compound and how many of them had the compound in the applicability domain. We grouped all compounds by trust level and computed defacto prediction accuracy within each group. Results are summarized in Table 4 for the test set.

Prediction accuracy apparently drops with a decrease in trust level, excluding the poor trust level that has only 33 compounds in the corresponding group, which may not be sufficient for an evaluation of prediction accuracy. This measure provides worse results than the CONST-STD-PROB measure, as demonstrated in Table 5.

The 681 molecules with the largest CONS-STD-PROB values have an accuracy of about 52% only (Table 4), i.e., the same as the random guess. Of course, one should not use predicted results for these molecules but rather experimentally measure them. Once measured, such molecules will be important in extending the applicability domain of models and will allow for reliable predictions of new molecules, which are similar to them.

One-Class Classification AD (SVM1 AD). This measure was provided by the MSU group, and it distinguishes compounds inside and outside of AD. Accuracies, grouped by this flag, are summarized in Table 6.

A majority of compounds from the training and test sets were predicted to be inside the applicability domain using SVM1. The prediction accuracy for these compounds was on average 5% higher than those outside of AD. The CONS-STD-PROB method provided a much better separation of molecules; it achieves differences up to 40% for reliable and nonreliable predictions (Table 4).

DA Index. In addition to quantitative values analyzed in the previous section, the TUB group provided qualitative values for their DA_Index, summarized in Table 7.

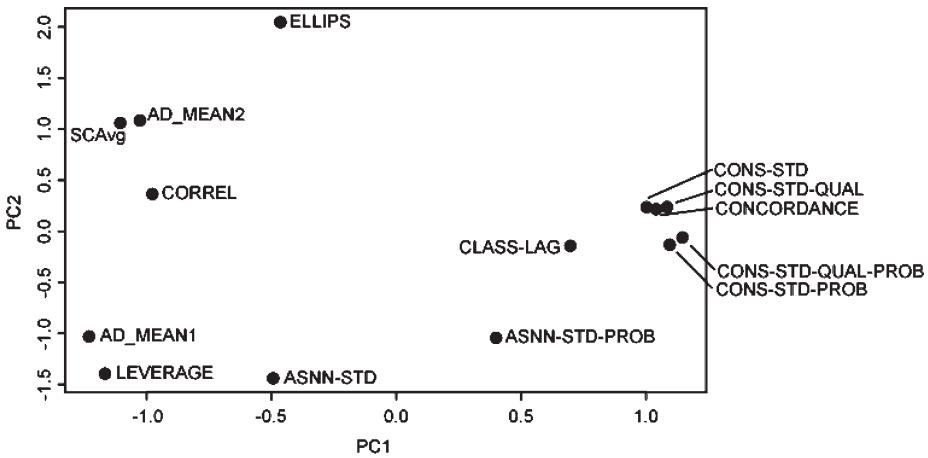


Figure 9. Principal component plot for the analyzed DMs. The PCA was based on the rankings that the DMs gave to the compounds from the training and test sets. Apparently, the five consensus-based DMs form two clusters: CONS-STD, CONS-STD-QUAL, and CONCORDANCE in the first cluster and CONS-STQ-QUAL-PROB and CONS-STD-PROB in the second one.

Table 4. Accuracy of Predictions According to CONS-STD-PROB[a]

	observed prediction accuracy				
number of compounds	ULP_ISIDA_SQS	TUB_3DDrag_SVM	TUB_3DDrag_RF	MSU_FRAG_LR	MSU_FRAG_SVM
500	96%	93%	93%	94%	95%
500	86%	89%	90%	89%	90%
500	76%	79%	81%	80%	83%
500	53%	64%	65%	66%	68%
181	48%	61%	55%	54%	54%
2181	75%	80%	80%	80%	81%

[a] For the first 500 compounds, it achieved an accuracy of 93−96%. This accuracy was higher than other qualitative ADs summarized in Tables 3−5.

Table 5. De Facto Performance of ULP_ISIDA_SQS Model for the Test Set with Regard to Trust Level and CONS-STD-PROB

	number of	observed prediction accuracy	
trust level	compounds	trust level	CONS-STD-PROB
optimal	1221	81%	89%
good	512	79%	69%
medium	415	53%	46%
poor (or less)	33	70%	45%
overall test set	2181	75%	

Most compounds (1819, or 83% of the test set) had a DA-Index value of 0, which corresponds to the highest expected accuracy. However, the increase in accuracy of 2−6% was not significant for both TUB models, TUB_3DDrag_SVM and TUB_3DDrag_RF, as shown in Table 7. For the same models, the 500 most accurately predicted compounds identified using CONS-STD-PROB had 93% classification accuracy for both models, as shown in Table 4.

Ability to Estimate Accuracies of Predictions. So far, we investigated the abilities of DMs to separate accurate and inaccurate predictions. The main criterion for such performance was the percentage of compounds that were predicted with 90% accuracy for the training and test sets ($c_{\text{TRAIN-90\%}}$ and $c_{\text{TEST-90\%}}$) with regard to a particular DM. However, it is also important to estimate the expected accuracy of predictions for new molecules. Under the assumption that a model is correctly cross-validated and the investigated DM is consistent, the prediction accuracy for compounds within the same DM threshold should be not significantly different for both 5-CV results and the test set. Thus, the DM selected using 5-CV should cover the same percentage of molecules having about the same accuracy of prediction for the test set. In order to check this assumption, we selected a DM threshold that provides 90% accuracy using 5-CV and calculated accuracies of predictions for compounds within the same threshold on the *test set*.

We have compared accuracies for the training and test sets on compounds, having a DM within the threshold that provides 90% accuracy for the 5-CV result. The comparison was performed for all of the models in combination with all of the investigated DMs. There are 20 models tested against 12 DMs; therefore, there are $20 \times 12 = 240$ comparison cases. We found that the accuracies of predictions for 5-CV and test sets are consistent with significance level $p = 0.01$. With significance level $p = 0.05$, the estimated and observed accuracies were significantly different for two cases (Table S3, part C, Supporting Information), which does not exceed the statistically expected number of failures (for 240 comparison cases, 12 failures at the 0.05 level of signifi-

Table 6. Performance of MSU_FRAG_LR and MSU_FRAG_SVM Models Depending on the SVM1 AD Factor and CONS-STD PROB (For the Same Numbers of Compounds)

		observed prediction accuracy			
		MSU_FRAG_LR		MSU_FRAG_SVM	
SVM1 AD	number of compounds	SVM1 AD	CONS-STD-PROB	SVM1 AD	CONS-STD-PROB
		training set			
inside (= 1)	4194	79%	80%	80%	81%
outside (= −1)	167	75%	59%	79%	53%
overall training set	4361	79%		80%	
		test set			
inside (= 1)	2046	81%	82%	81%	83%
outside (= −1)	135	73%	53%	79%	55%
overall test set	2181	80%		81%	

Table 7. Performance of TUB Models for the Test Set Depending on DA Index and CONS-STD-PROB

		observed prediction accuracy			
		TUB_3DDrag_SVM		TUB_3DDrag_RF	
DA Index	number of compounds	DA Index	CONS-STD-PROB	DA Index	CONS-STD-PROB
0	1819	81%	83%	80%	84%
between 0 and 1	183	75%	62%	78%	61%
1	179	75%	60%	80%	60%
overall test set	2181	80%		80%	

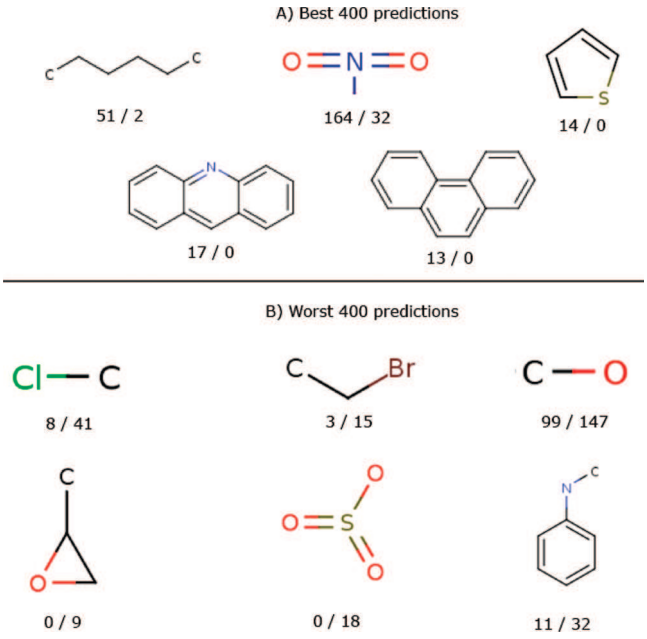


Figure 10. Molecular fragments, presented in the reliably and nonreliably predicted compounds. Shown are the fragments, significantly over-represented in the molecules with the highest accuracy (A) and the lowest accuracy (B) according to CONS-STD-PROB DM. Below the fragments are the numbers of relevant molecules with accurate (left of the slash) and inaccurate predictions (right of the slash).

cance). Thus, the accuracies estimated *a priori* using the training set are in agreement with observed accuracies for the test set.

Substructural Analysis of the Applicability Domain. To determine which types of molecules are predicted accurately and which are not, we have analyzed molecular subfragments for 400 predictions with the highest and lowest accuracies according to the CONS-STD-PROB DM, respectively. We will refer to these sets as "worst-400" and "best-400". We enumerated all of the fragments presented in these molecules and counted the number of molecules containing each fragment in each set.

If a fragment is equally distributed, the number of molecules from the "best-400" (or the "worst-400") containing this fragment should be distributed binomially with p equal to 0.5 and N equal to the total number of the molecules containing this fragment. If this assumption was invalidated with at least a $p < 0.05$ level of significance, we then considered the fragment as over-represented in one of the sets.

An overview of several significant fragments is presented in Figure 10. Apparently, the molecules containing long carbon chains, nitro groups, and thiophene groups were over-represented in the "best-400" predictions. We found out that long carbon chains were mostly presented in nonmutagenic compounds, whereas nitro and thiophene groups are mostly in mutagenic compounds. For the prediction of such compounds, there was a high level of agreement between the models. In contrast, the compounds containing chlorine, bromine, sulfonate, and epoxide groups are not reliably predicted by the models investigated in this study. We plan to provide a more detailed analysis of these fragments to detect "toxicophores", i.e., structural elements responsible for the mutagenicity of analyzed compounds.

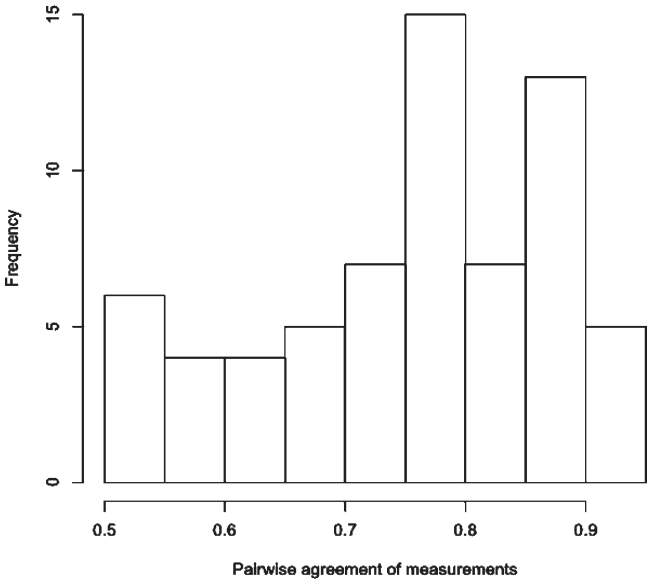


Figure 11. Distribution of the pairwise agreements of the Ames test measurements carried out by 12 laboratories. The data for the plot were taken from a study by Benigni and Giuliani.[55]

Data Variability Analysis. Several studies analyzed the variability of the Ames test experiments. Let us critically review them for a better understanding of the results of our modeling.

The first study by Benigni and Giuliani[55] assessed the Ames tests conducted for 42 compounds by 12 different experimental laboratories. Using the same data, for every pair of laboratories, we calculated the level of agreement as the number of the concordant measurements divided by the total number of measurements. The distribution of agreements of 66 lab pairs is shown in Figure 11. The average pairwise agreement is only 75%. At the same time, Figure 11 reveals that the agreement of results between some laboratories can be sometimes higher than 90%. This result was observed for 4 out of 66 pairs of laboratories (7% of all data). However, it is possible to expect a higher agreement if the data are measured within the same laboratory.

In the study by Piegorsch and Zeiger,[56] the experimental concordance between different laboratories was reported in the range of 70−87%. Each molecule in this set was measured in several experiments either in different laboratories or in the same lab but at different times. The outcomes of experiments were positive (+), weak positive (+W), negative, (−), and questionable (?). Let us consider, similar to how it was done in the analysis by the original authors, positive and weak positive as Ames mutagens and ignore nondecisive experiments, which, of course, are usually expected to be remeasured. Similar to the previous section, let us define the accuracy of one compound as the maximum number of positive or negative tests divided by the total number of decisive experiments. Such accuracy could be expected for our analysis, if we assume that molecules were tested on average just once. The average accuracy of the Ames test was 93% and 90% if we considered molecules with at least two (209 molecules) or three (49 molecules) decisive measurements, respectively.

We further explored this result using the variability of measurements used in our study. For this analysis, we used the Ames test data collected and publicly available at the

OCHEM website (http://ochem.eu). The database contains results for 3205 of the 6542 Ames challenge compounds. We used the same definition of accuracy as above and calculated an average accuracy of 94% for compounds, which had at least three measurements (1680 compounds selected from 189 articles). The variation of the minimal number of measurements from four to seven did not change this number more than ±0.3%. The 94% agreement is conformable with the achievable prediction accuracies of the models investigated in this study.

In this analysis, we mainly considered intralaboratory variations, as compared to the interlaboratory and mixture of the inter- and intralaboratory variations estimated in works of Benigni and Giuliani[55] and Piegorsch and Zeiger,[56] respectively. Unfortunately, it was impossible to carry out interlaboratory analysis in our study as there was an overlap in molecules reported in different articles. Moreover, in some cases, several authors, in particular Errol Zeiger, have contributed to the majority of articles, thus invalidating the goal of the interlaboratory comparison. Therefore, for the comparison of the DMs, we selected the accuracy of 90% obtained in work of Piegorsch and Zeiger[56] as a conservative threshold for interlaboratory comparison.

Confidence of Predictions vs Variability of Experimental Measurements. Different subsets of molecules may behave differently in experiments: some of them may have easily reproducible results (either mutagenic or nonmutagenic), while the other molecules may show higher variability, e.g., because of difficulties in experimental measurements such as metabolic stability, low solubility, etc. It would be interesting to know whether the methods described in the article can differentiate such chemicals.

We have analyzed the variability of measurements for molecules from the Piegorsch and Zeiger data set.[56] The total set contained 239 molecules, but three of them did not have structures defined and were excluded from our analysis. We developed a new ASNN model using all of the Ames challenge molecules with an exception of these 236 molecules, which formed the test set. The confidence of predictions was determined using the ASNN-STD-PROB DM. Amid 50 compounds with the highest and the lowest calculated confidences, we selected molecules that had at least three decisive measurements. There were 14 and 9 such molecules for the top and lower ranges with an agreement of experimental measurements of 96% and 89%, respectively. Moreover, there were also 13 and 21 compounds with questionable measurements within the same intervals.

We applied a similar analysis to the 1680 Ames challenge compounds having at least three measurements. We found that 150 molecules with the highest and the lowest confidence of predictions had an agreement of experimental measurements of 97% and 91%, respectively. Thus, the confidence of predictions determined by the DM correlated with the variability of experimental measurements: the molecules with a higher confidence of predictions have better agreements of experimental measurements and vice versa.

The Prediction Accuracy for EINECS, ENAMINE, and HPV Data Sets. In order to estimate the applicability of the QSAR Ames models to diverse chemical compounds, the OCHEM_ESTATE_ANN model was applied to the ENAMINE, EINECS, and HPV databases, described in the Methods section. The prediction accuracy for these data sets

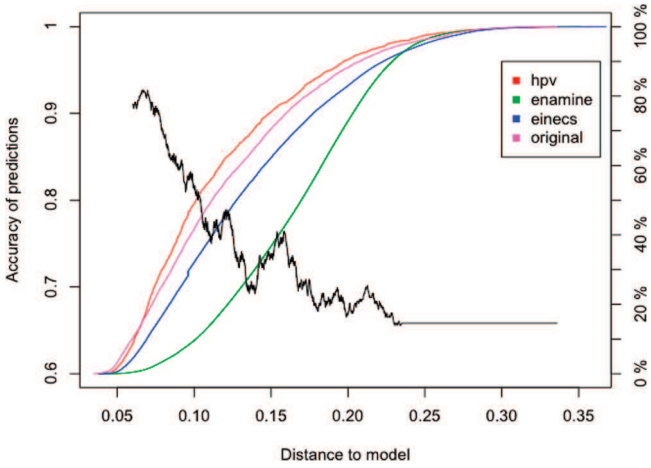


Figure 12. Estimated prediction accuracy for the original Ames challenge data set and the HPV, EINECS, and ENAMINE data sets. The black curve, based on bin-based averaging, plots the prediction accuracy (left *y* axis) against the ASNN-STD-PROB DM. Colored curves show percentages of compounds (right *y* axis) from the four data sets, having ASNN-STD-PROB not more than a particular threshold (*x* axis).

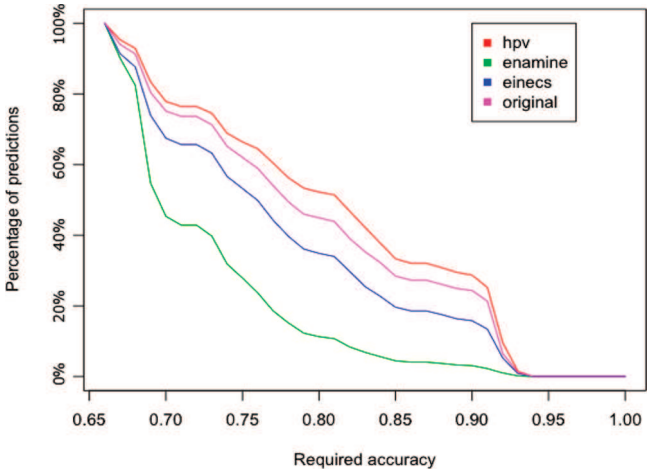


Figure 13. Percentages of compounds (*y* axis) having a required prediction accuracy (*x* axis). This plot is built for four data sets and uses the same data as Figure 12.

was estimated using bin-based accuracy averaging based on the ASNN-STD-PROB DM.

In Figure 12, the black curve corresponds to the average prediction accuracy as a function of ASNN-STD-PROB, while four colored curves illustrate the percentages of compounds from the four data sets having DM values less than corresponding thresholds. The plot in Figure 13 shows the percentages of compounds from the four data sets depending on the required prediction accuracy. Apparently, for the HPV and EINECS data sets, the percentages of reliable predictions (with at least 90% estimated prediction accuracy) were 30% and 16%, respectively, which is close to the percentage in the original data set, used for training and validation (25%). However, the percentage of reliable predictions in the ENAMINE data set was only 4%, probably due to a higher chemical diversity of compounds in comparison to the training set.

CONCLUSIONS

In this study, we have analyzed the AD problem for binary classification models. We investigated the relevance of

classical approaches to AD estimation for predictions of quantitative properties. The analysis was based on the Ames mutagenicity data set and involved 30 independent classification models.[11] The model developed by the HMGU group has been made publicly available in OCHEM, Online Chemical Modeling Environment,[57] at http://ochem.eu/models/1.

The analysis in this study was based on abstract measures of prediction uncertainty, referred to as "distances to models" (DMs). While the fact that measures such as CLASS-LAG, which can be used to discriminate accurate and inaccurate predictions, have been known for years, not many researches utilize them to assess the performance of their QSAR methods (frequently, only average model characteristics are reported).

The important message of this study was to demonstrate practical advantages of using DM and AD approaches. The most reliable predictions of the Ames test achieved experimental accuracy (ca. 90%), while unreliable predictions had an accuracy of random guessing (50%). The predictions of the later compounds are useless; one should measure such compounds experimentally rather than rely on predictions.

Several DMs were investigated and benchmarked. The DMs, based on the global consensus model provided significantly better separation ($p < 0.05$ using the Wilxocon test) of low and high accuracy predictions. The top-ranked DMs included a recently introduced probability-based measure of distance to a binary classification model, CONS-STD-PROB,[45] its qualitative analog CONS-STD-QUAL-PROB, as well as another very simple measure, CONCORDANCE, i.e., the agreement of a model's predictions with the global consensus model. Moreover, as shown in Figure 9, these three DMs were strongly correlated. The quality of the AD estimation using these methods was significantly better than tthat of the traditionally used CLASS-LAG method. Nonetheless, while CLASS-LAG did not work for the majority of the analyzed individual models, its performance for the global consensus model was not significantly different from the three aforementioned top-ranked DMs (see Table 2). It is important to mention that all three measures (CONS-STD-QUAL-PROB, CONS-STD-PROB, and CONCORDANCE) implicitly use the predictions given by the consensus model. As the consensus model is the best of all 31 models, these DMs may have performed best because they incorporate information from the best (consensus) model. If we do not consider the consensus-based DMs, the best measures were CLASS-LAG and ASNN-STD-PROB. Importantly, the DMs based on the output of the models outperformed the DMs solely on the basis of molecular structures (e.g., LEVERAGE and AD_MEAN).

Similar to our previous analysis of quantitative QSAR models,[8] we found that the best separation of the reliable and nonreliable predictions was provided by the same DMs. In other words, the compounds having the best prediction accuracy were the same for all of the models, regardless of the descriptors or the machine-learning technique used to develop them. This conclusion is in agreement with the work of Sheridan et al.[58] as well as with our own conclusions that the performance of models is dominated by the size and quality of the training set rather than by the method or the descriptors.[59]

Another important result of this study is the discovery of a correlation between the prediction uncertainty and the variability of experimental measurements of molecules. Namely, we have demonstrated that molecules with more accurate predictions had a higher agreement of experimental measurements and, vice versa, molecules with less accurate predictions showed higher disagreement with experimental measurements. Indeed, molecules from the first group contributed cleaner training sets and thus allowed models to achieve a higher accuracy of predictions for their analogs.

The discrimination of accurate and nonaccurate predictions is important from the practical point of view. If a compound is predicted with the accuracy, which is close to the accuracy of experimental measurements, one can use *in silico* values instead of measuring the activity for this compound. We have shown that the developed models predicted Ames mutagenicity for 35−65% of Ames challenge molecules with an accuracy similar to that of interlaboratory variation. Similar results were also achieved for quantitative models: the octanol/water partition coefficient (log P) was calculated for more than 60% of molecules with experimental accuracy.[60]

An accuracy of 90% was achieved for 35% and 20% of the HPV and EINECS databases of compounds using the ASNN model. However, for a larger and more diverse Enamine data set, only 6% of the compounds were predicted with such accuracy, presumably because of the higher chemical diversity of the Enamine collection. Thus, to increase the accuracy of predictions for such compounds, new experimental measurements are required.

In summary, the differentiation of reliable and nonreliable predictions of *in silico* approaches can decrease experimental costs by delivering accurate predictions for up to 2/3 of the molecules. At the same time, those compounds, which cannot be reliably predicted, should be measured. This additional data will extend the applicability domain of models and will allow reliable prediction of an even larger number of molecules. The combination of *in silico* approaches and experimental measurements can help to avoid redundant measurements, to screen large amounts of molecules even before they are synthesized, and, thereby, to provide significant savings of time and cost for the industry.

ACKNOWLEDGMENT

This study was partially supported with GO-Bio BMBF grant 0313883, FP7 project CADASTER 212668, and Germany−Ukraine collaboration project UKR 08/006, and by the NIH grants R01GM66940 and R21GM076059. We would like to thank all participants of the Ames challenge, who contributed to the development of models used in this study as well as the reviewers for their constructive remarks.

Note Added after ASAP Publication. This paper was published ASAP on October 29, 2010 with an error in the presentation of the names of the authors. The corrected version was published ASAP on November 8, 2010.

Supporting Information Available: Detailed tables with the DM scores and the values of DM comparison criteria. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

(2) Tetko, I. V.; Bruneau, P.; Mewes, H.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions. *Drug Discovery Today* **2006**, *11*, 700–707.

(3) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155–173.

(4) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.

(5) Hemmateenejad, B.; Yazdani, M. QSPR models for half-wave reduction potential of steroids: A comparative study between feature selection and feature extraction from subsets of or entire set of descriptors. *Anal. Chim. Acta* **2009**, *634*, 27–35.

(6) Tropsha, A.; Gramatica, P.; Gombar, V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.

(7) Aires, F.; Prigent, C.; Rossow, W. B. Neural Network Uncertainty Assessment Using Bayesian Statistics: A Remote Sensing Application. *Neural Comput.* **2004**, *16*, 2415–2458.

(8) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(9) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674–679.

(10) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjogren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137–142.

(11) Muratov, E. *Summer School on Chemoinforamtics*; Obernai: France, 2010; poster no 13.

(12) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K. Benchmark data set for *in silico* prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.

(13) Ames, B. N.; Lee, F. D.; Durston, W. E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci. U.S.A* **1973**, *70*, 782–786.

(14) Breiman, L. Random Forests. *Mach. Learning* **2001**, *45*, 5–32.

(15) Chang, C.; Lin, C. LIBSVM - A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed Oct 1, 2010).

(16) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: New York, 2009.

(17) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learning* **2002**, *46*, 389–422.

(18) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab.* **2001**, *58*, 109–130.

(19) Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.

(20) Martens, H.; Martens, M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Qual. Prefer.* **2000**, *11*, 5–16.

(21) Tetko, I. V. Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.

(22) Tetko, I. V. Associative neural network. *Methods Mol. Biol.* **2008**, *458*, 185–202.

(23) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074–1080.

(24) Karakoc, E.; Sahinalp, S. C.; Cherkasov, A. Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2167–2182.

(25) Cherkasov, A.; Ban, F.; Li, Y.; Fallahi, M.; Hammond, G. L. Progressive docking: a hybrid QSAR/docking approach for accelerating *in silico* high throughput screening. *J. Med. Chem.* **2006**, *49*, 7466–7478.

(26) Cherkasov, A. Can 'Bacterial-Metabolite-Likeness' Model Improve Odds of '*in silico*' Antibiotic Discovery. *J. Chem. Inf. Model.* **2006**, *46*, 1214–1222.

(27) Cover, T.; Thomas, A. J. *Elements of information theory*; Wiley: New York, 1991; pp 1−543.

(28) Witten, I. H.; Frank, E. *Data mining: practical machine learning tools and techniques with Java implementations*; Morgan Kaufmann: San Francisco, CA, 1999; pp 1−374.

(29) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191–198.

(30) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship GenerationHow Much Effort May the Mining for Successful QSAR Models Take. *J. Chem. Inf. Model.* **2007**, *47*, 927–939.

(31) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998; pp 1−736.

(32) Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; Lin, C. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.

(33) Artemenko, N. V.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks within the Substructure Approach. *Dokl. Chem.* **2001**, *381*, 317–320.

(34) Baskin, I. I. ; Halberstam, N. M. ; Artemenko, N. V. ; Palyulin, V. A. ; Zefirov, N. S. In *Euroqsar 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*; Ford, M. , Livingstone, D., Dearden, J., Van de Waterbeemd, H., Eds.; Blackwell Science Inc: Bournemouth, 2003; pp 260−263.

(35) Kuz'min, V.; Artemenko, A.; Muratov, E. Hierarchical QSAR technology based on the Simplex representation of molecular structure. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 403–421.

(36) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Volineckaya, I. L.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. Quantitative structure-activity relationship studies of [(biphenyloxy)-propyl]isoxazole derivatives. Inhibitors of human rhinovirus 2 replication. *J. Med. Chem.* **2007**, *50*, 4205–4213.

(37) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.

(38) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON software: an easy approach to molecular descriptor calculations. *MATCH* **2006**, *56*, 237–248.

(39) Breiman, L.; Friedman, H.; Olshen, R. A.; Stone, C. J. *Classification and regression trees*; Wadsworth International Group: Belmont, CA, 1984; pp 1−359.

(40) Fan, R.; Chen, P.; Lin, C. Working Set Selection Using Second Order Information for Training Support Vector Machines. *J. Mach. Learn. Res.* **2005**, *6*, 1889–1918.

(41) Martin, T. M.; Harten, P.; Venkatapathy, R.; Das, S.; Young, D. M. A Hierarchical Clustering Methodology for the Estimation of Toxicity. *Toxicol. Mech. Methods* **2008**, *18*, 251–266.

(42) Contrera, J. F.; Matthews, E. J.; Daniel Benz, R. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharmacol.* **2003**, *38*, 243–259.

(43) Tetko, I. V.; Poda, G.; Ostermann, C.; Mannhold, R. Accurate *in silico* log P Predictions: One Can't Embrace the Unembraceable. *QSAR Comb. Sci.* **2009**, *28*, 845–849.

(44) Breiman, L. Bagging predictors. *Mach. Learning* **1996**, *24*, 123–140.

(45) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for *in silico* models to achieve accuracy of experimental measurements. *J. Chemom.* **2010**, *24*, 202–208.

(46) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J. Pharm. Sci.* **2009**, *98*, 861–893.

(47) Harmeling, S.; Dornhege, G.; Tax, D.; Meinecke, F.; Müller, K. From outliers to prototypes: Ordering data. *Neurocomputing* **2006**, *69*, 1608–1618.

(48) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; Ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K. A probabilistic

approach to classifying metabolic stability. *J. Chem. Inf. Model.* **2008**, *48*, 785–796.
(49) Montgomery, D.; Peck, E. A.; Vining, G. G. *Introduction to linear regression analysis*; Wiley: New York, 2006; pp 1−639.
(50) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **2009**, *49*, 1762–1776.
(51) Schölkopf, B.; Smola, A. J. *Learning with kernels*; MIT Press: Cambridge, U.K., 2002; pp 1−644.
(52) Bishop, C. M. Novelty Detection and Neural Network Validation. *IEEE Proc.: Vis. Imag. Sign. Proc.* **1994**, *141*, 217–222.
(53) Fechner, N.; Jahn, A.; Hinselmann, G.; Zell, A. Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *J. Cheminf.* **2010**, *2*, P2.
(54) Wilcoxon, F. Individual Comparisons by Ranking Methods. *Bio. Bull.* **1945**, *1*, 80–83.
(55) Benigni, R.; Giuliani, A. Computer-assisted analysis of interlaboratory Ames test variability. *J. Toxicol. Environ. Health* **1988**, *25*, 135–148.
(56) Piegorsch, W.; Zeiger, E. Measuring intra-assay agreement for the Ames Salmonella assay. *Lect. Notes Med. Inf.* **1991**, *43*, 35–41.
(57) Novotarskyi, S.; Sushko, I.; Körner, R.; Kumar, A.; Rupp, M.; Prokopenko, V.; Tetko, I. OCHEM - on-line CHEmical database & modeling environment. *J. Cheminf.* **2010**, *2*, P5.
(58) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
(59) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.
(60) Tetko, I. V.; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chem. Biodivers.* **2009**, *6*, 1837–1844.

CI100253R

UNIVERSITÉ DE STRASBOURG

Christophe MULLER

Relations Structure-Activité pour le métabolisme et la toxicité.

Résumé

Prédire à l'avance quels composés seront toxiques chez l'homme ou non représente un réel challenge dans le monde pharmaceutique. En effet, les mécanismes à l'origine de la toxicité ne sont pas toujours bien connus, et à cela s'ajoute le fait qu'un composé peut devenir néfaste seulement après qu'il ait été métabolisé.

Nous proposons ici une approche originale utilisant les graphes condensés de réactions afin de modéliser les réactions métaboliques et prédire le devenir des xénobiotiques dans l'organisme humain. Différentes formes de toxicité sont aussi prédites : la mutagénicité et l'hépatotoxicité. Pour cette seconde toxicité, l'approche utilisée est la première à notre connaissance à prédire avec succès les molécules toxiques décrites par des données autres que résultant d'observations in vivo.

Mots clefs : chémoinformatique, Graphe Condensé de Réaction, QSAR, métabolisme, toxicité.

Abstract

Predict in advance which compounds will be toxic in humans or not is a real challenge in the pharmaceutical world. Indeed, the mechanisms responsible for toxicity are not always well known, and in some case a compound become toxic only after it has been metabolized.

We propose here a novel approach using condensed graphs of reactions to model and predict the metabolic fate of xenobiotics in the human body. Various forms of toxicity are also predicted: mutagenicity and hepatotoxicity. For this second toxicity, the approach proposed is the first to our knowledge to successfully predict the toxic molecules described by data other than resulting from observations in vivo.

Keywords: chemoinformatics, Condensed Graph of Reaction, QSAR, metabolism toxicity.