Property-enriched fragment descriptors for adaptive QSAR

Fiorella Ruggiu

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

[UMR 7140]

THÈSE

présentée par :

[Fiorella RUGGIU]

soutenue le : **22 septembre 2014**

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

Discipline/ Spécialité : Chimie/Chémoinformatique

Property-enriched fragment descriptors for adaptive QSAR

THÈSE dirigée par :

[M. VARNEK Alexandre]	Professeur, Université de Strasbourg
[M. HORVATH Dragos]	Directeur de Recherche, CNRS, UMR 7140

RAPPORTEURS :

[M. AIRES DE SOUSA João]	Professeur, Université de Lisbonne
[M. TABOUREAU Olivier]	Professeur, Université Paris Diderot

AUTRES MEMBRES DU JURY :

[M. ERTL Peter]	Docteur, HDR, société Novartis
[Mme. KELLENBERGER Esther]	Docteur, HDR, Université de Strasbourg

Fiorella RUGGIU

Property-enriched fragment descriptors for adaptive QSAR

Résumé

Les descripteurs ISIDA enrichis par propriété ont été introduit pour encoder les structures moléculaires en chémoinformatique en tant que nombre d'occurrence de sous-graphes moléculaires spécifiques dont les sommets représentant les atomes sont colorés par des propriétés locales tel que les pharmacophores dépendant du pH, les identifiants de champs de force, les charges partielles, les incréments LogP ou les propriétés extraites d'un modèle QSAR. Ces descripteurs, par leurs large choix d'option, permettent à l'utilisateur de les adapter au problème à modéliser. Ils ont été utilisés avec succès dans une étude de criblage virtuel sur des inhibiteurs de protéases et plusieurs modèles QSAR sur le coefficient de partage octanol-eau, l'index d'hydrophobicité chromatographique, l'inhibition du canal hERG, la constante de dissociation acide, la force des accepteurs de liaison hydrogène et l'affinité de liaison des GPCR.

Résumé en anglais

ISIDA property-enriched fragment descriptors were introduced as a general framework to numerically encode molecular structures in chemoinformatics, as counts of specific subgraphs in which atom vertices are coloured with respect to a local property - notably pH-dependent pharmacophore, force field, partial charges, logP increments and QSAR model extracted properties. The descriptors leave the user a vast choice in terms of the level of resolution at which chemical information is extracted into the descriptors to adapt them to the problem. They were successfully tested in neighbourhood behaviour and QSAR modelling challenges, with very promising results. They showed excellent results in similarity-based virtual screening for analogue protease inhibitors, and generated highly predictive octanol-water partition coefficient, chromatographic hydrophobicity index, hERG channel inhibition, acidic dissociation constant, hydrogen-bond acceptor strength and GPCR binding affinity models.

Acknowledgements

First and foremost, I would like to express my gratitude to my advisers, Pr. Alexandre Varnek and Dr. Dragos Horvath for their kindness, patience and advice. Besides my advisers, I am particularly grateful to Dr. Gilles Marcou for his advice in my research and for his kindness as well. Without them, this thesis would have never come into existence.

I would like to thank all the members of the jury, Pr. João Aires de Sousa, Pr. Olivier Taboureau, Dr. Peter Ertl, Dr. Esther Kellenberger and Dr. Jean-Luc Galzi as an invited member, for accepting to judge and revise my work.

I am thankful to all my collaborators on the different projects, Dr. Vitaly Solov'ev, Dr. Patrick Gizzi, Dr. J.B. Brown, Dr. Jean-Luc Galzi, Dr. Jérôme Graton et Pr. Jean-Yves Le Questel for their help and advice.

A particular thanks to Pr. Okuno and to the Japan Society for the Promotion of Science for enabling me to do research in Japan.

I am thankful to my colleagues, Dr. Fanny Bonachera, Dr. Ioana Oprisiu, Dr. Olga Klimchuk, Grace Delouis, Héléna Gaspar, Pr. Igor Baskin, Dr. Tetiana Khristova, Dr. Christophe Muller, Dr. Laurent Hoffer and Aurélie De Luca for their advice and support. I would like to thank my students, Jacques Ehret, Guillaume Charbonnier, Julien Denos and Jiangyue He, for working together. Thank to Sandrine Garcin, Danièle Ludwig and Soumia Hnini for their help on administrative tasks.

Résumé en français :
Descripteurs fragmentaux enrichis par propriété pour QSAR adaptif

Cette thèse présente des développements méthodologiques dans le domaine de la relation structure-activité et leurs applications à la modélisation de certaines propriétés chimiques et activités biologiques. L'objectif principal de cette thèse était l'extension des descripteurs développés au Laboratoire de Chémoinformatique (UMR 7140, Chimie de la Matière Complexe, Université de Strasbourg, France), les descripteurs fragmentaux ISIDA. ISIDA est le nom de la suite de logiciels du laboratoire et est l'abréviation de « In Silico Design and data Analysis » (Conception et analyse de données in silico). Comme présenté dans la Figure 1, les Triplets Pharmacophoriques Flous (TPF) et les Fragments Moléculaires Sous-structuraux (FMS) avaient été développés précédemment. Ces types différents de descripteurs ont été unifiés et étendus dans ce travail. Trois différents types de descripteurs sont proposés dans ce travail (en bleu sur la Figure 1) : les fragments avec projection de propriétés, les fragments locaux et les fragments avec projection d'incréments QSAR.

Le manuscrit est divisé selon les sections suivantes:

- Une première section introductive au sujet et une partie bibliographique pour la méthodologie,
- une deuxième section de présentation des nouveaux descripteurs développés dans cette thèse,
- une troisième section présentant les différentes études d'applications des descripteurs effectuées,
- une section présentant les développements logiciels,
- et finalement une section pour la conclusion et les perspectives de ce travail.

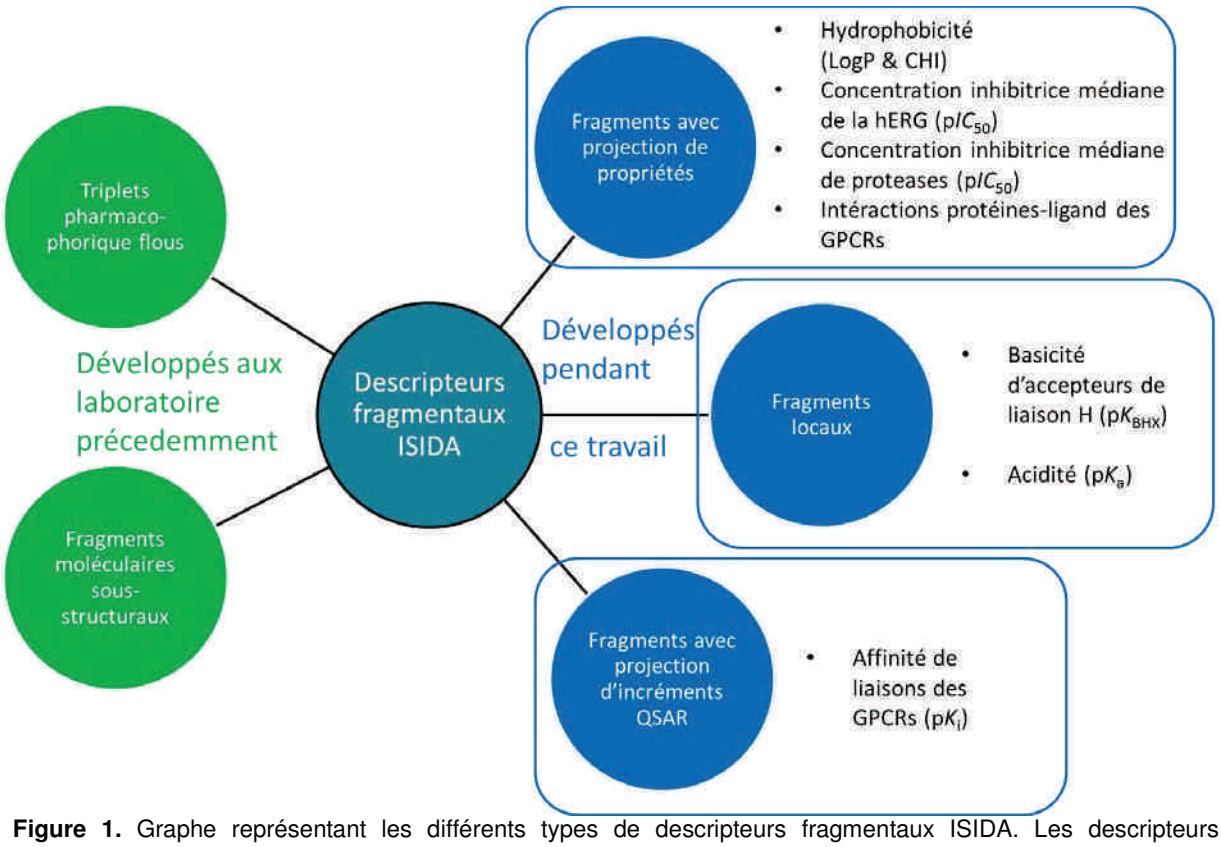


Figure 1. Graphe représentant les différents types de descripteurs fragmentaux ISIDA. Les descripteurs existants avant la thèse sont en vert (à gauche) et ceux développés durant la thèse sont en bleu (à droite). Les cadres bleus incluent les propriétés sur lesquels les différents descripteurs ont été testés.

1. Descripteurs fragmentaux : état de l'art

La chémoinformatique utilise des outils informatiques afin de gérer, interpréter et extraire des connaissances de l'information chimique. Celle-ci fait abondamment appel à la notion de graphe moléculaire (par exemple des structures de Lewis. Or les graphes sont des structures de données compliquées ce qui rend mal aisée leur utilisation direct pour analyser et extraire des connaissances par des moyens informatiques. Une représentation sous forme de descripteurs moléculaires leurs est préférée. Les descripteurs moléculaires consistent à extraire l'information du graphe moléculaire et de l'encoder dans des vecteurs, D, où chaque composant i représente une caractéristique de la molécule (D_i). Ces vecteurs de descripteurs sont bien adaptés pour l'analyse mathématique des relations structure-activité, en particulier pour

(a) le Criblage Virtuel par Similarité (CVS) reposant sur le principe de similarité classique : « des molécules similaires présenteront probablement des propriétés similaires ». D'un point de vue mathématique, une mesure de similarité est définie en fonction des vecteurs de descripteurs Ainsi, le calcul de similarité entre un vecteur référence décrivant une molécule d'intérêt, par exemple active en se liant à une protéine, et des vecteurs d'autres molécules d'une base de données permet de sélectionner les molécules les plus probablement intéressantes.

(b) les Relations Quantitative Structure-Activité ou propriété (Quantitative Structure-Activity Relationship - QSAR), où les techniques d'apprentissage automatique sont employées pour chercher des équations empiriques qui expriment une propriété moléculaire Y comme une fonction des composants D_i des vecteurs de descripteurs. Si une telle fonction est établie et qu'elle retourne des approximations proches de la valeur Y de molécules connues, elle peut être utilisée pour estimer cette valeur pour des composés virtuels n'ayant pas été synthétisés ou testés. En effet, les termes Di peuvent être calculés préalablement à la synthèse et ainsi des composés avec des valeurs de Y intéressantes pourront être privilégiées pour la synthèse et les tests.

L'art de concevoir des descripteurs moléculaires consiste à décider sur quels types de caractéristiques de la molécule doivent se concentrer les vecteurs. Evidemment, le CVS et le QSAR ne fonctionne qu'à condition que des informations pertinentes pour la propriété soient inclus dans le vecteur D. Par exemple, le laboratoire de Chémoinformatique de l'université de Strasbourg, a développé des descripteurs fragmentaux, Fragments Moléculaires Sous-structuraux (FMS) d'ISIDA, qui sont composés de motifs sous-structuraux linéaires ou branchés ayant pour valeurs de descripteurs leurs occurrences dans la structure. L'interprétation des propriétés moléculaires en identifiant et en analysant des sous-structures du graphe moléculaire est proche du raisonnement des chimistes.

Les FMS fonctionnent bien en modélisation car ils semblent bien saisir l'information sur la connectivité de la molécule. Ceci peut être considéré comme un aspect fort de ces descripteurs.

Cependant, la connectivité n'est pas le seul aspect contrôlant les propriétés moléculaires. Par exemple, la liaison d'un ligand à un récepteur biologique est contrôlée par des points d'ancrage spécifiques définissant le « pharmacophore ». Si deux molécules ont des motifs de pharmacophores similaires, *cad*. des arrangements similaires de groupes équivalents d'un point de vue physico-chimique (comme des caractéristiques hydrophobes, accepteurs/donneurs de liaison hydrogène, charges, *etc*.), elles peuvent possiblement se lier à la même cible même si elles n'ont pas le même châssis moléculaire. Les descripteurs pharmacophoriques comptabilisent les occurrences des combinaisons (paires, triplets, *etc*.) de groupements à caractère pharmacophorique dans les molécules. Ces derniers peuvent ainsi capturer un aspect de la similarité moléculaire alternatif à celui des fragments qui n'est

pas évident à percevoir par l'esprit humain. Les Triplets Pharmacophoriques Flous (TPF) font partis de ce type de descripteurs. Ce sont des triplets d'atomes auxquels une caractéristique pharmacophorique a été attribuée avec les distances séparant les atomes deux à deux.

Dans ce contexte, le premier développement effectué dans cette thèse était d'unifier et de généraliser les deux points de vue : pharmacophore et fragment. Le logiciel permet d'utiliser n'importe quelle projection définie par l'utilisateur. Différentes propriétés projetées sur le graphe moléculaire, incluant le symbole atomique, les caractéristiques pharmacophoriques, les identifiants d'un champ de force, les charges partielles, les potentiels électrostatiques topologiques et les incréments d'hydrophobicité selon Ghose-Crippen ont été testées. Ce développement a fait l'objet d'une publication dans *Molecular Informatics* en 2010. Afin de projeter les propriétés, l'histogramme des distributions des incréments atomiques est utilisé. Un test a également été effectué en utilisant des incréments tirés d'un modèle QSAR de l'affinité de liaison des GPCRs à partir de l'analyse des descripteurs fragmentaux ISIDA et de la fonction régissant le modèle. Les premiers résultats de ce dernier ne sont néanmoins pas concluants pour le moment.

Le second développement majeur en termes de descripteurs moléculaires a été l'introduction des fragments locaux en « marquant l'atome » concernée par une propriété dite « locale » telles que l'acidité ou la force d'une liaison hydrogène. En effet, ce type de propriété dépend du groupe fonctionnel et de ses alentours plutôt que de la molécule dans sa globalité. De plus, les molécules peuvent être polyfonctionnelles, ce qui n'est pas représentable explicitement avec la plupart des descripteurs classiques. En indiquant l'atome ou les atomes concernés d'une fonction particulière, on pourra obtenir un vecteur de descripteurs pertinent pour représenter les fonctions chimiques de la molécule. Ce second développement a également fait l'objet d'une publication dans *Molecular Informatics* en 2014.

2. Nouveaux descripteurs ISIDA

Les descripteurs ISIDA, après les développements de cette thèse, sont composés d'une multitude d'options contrôlables par l'utilisateur. D'une part, une propriété peut être projetée sur le graphe moléculaire, d'autre part, ce dernier peut être fragmenté par différentes topologies et finalement l'occurrence des fragments peut être contrôlée pour prendre en compte ou non le pH.

Plusieurs propriétés pouvant être associées à un atome ont été testées au cours de cette thèse (la nomenclature est donnée entre parenthèses):

- Symboles atomiques (A)
- Caractéristiques pharmacophoriques (Ph)
- Identifiants atomiques de champs de force (Ff)
- Charges partielles (Pc)
- Potentiel électrostatique topologique (Ep)
- Incréments d'hydrophobicité selon Ghose-Crippen (Lp)

Trois classes de topologies principales sont distinguées :

- Séquences (I)
- Fragments centrés sur un atome (II)
- Triplets (III)

En plus de cela, une option pour faire des paires d'atome (indiquée par un P dans la nomenclature) peut être utilisée en conjonction avec les séquences et les fragments centrés sur un atome. Cela permet de créer des paires d'atomes où la distance topologique

les séparant est indiquée dans le cas des séquences (topologie I) et dans le cas des fragments centrés (topologie II), seules les extrémités sont représentées avec la distance topologique à l'atome central. Il faut également définir une distance topologique minimum et maximum pour limiter l'énumération des sous-structures possibles. Celles-ci sont données en nombre d'atomes inclus dans le fragment.

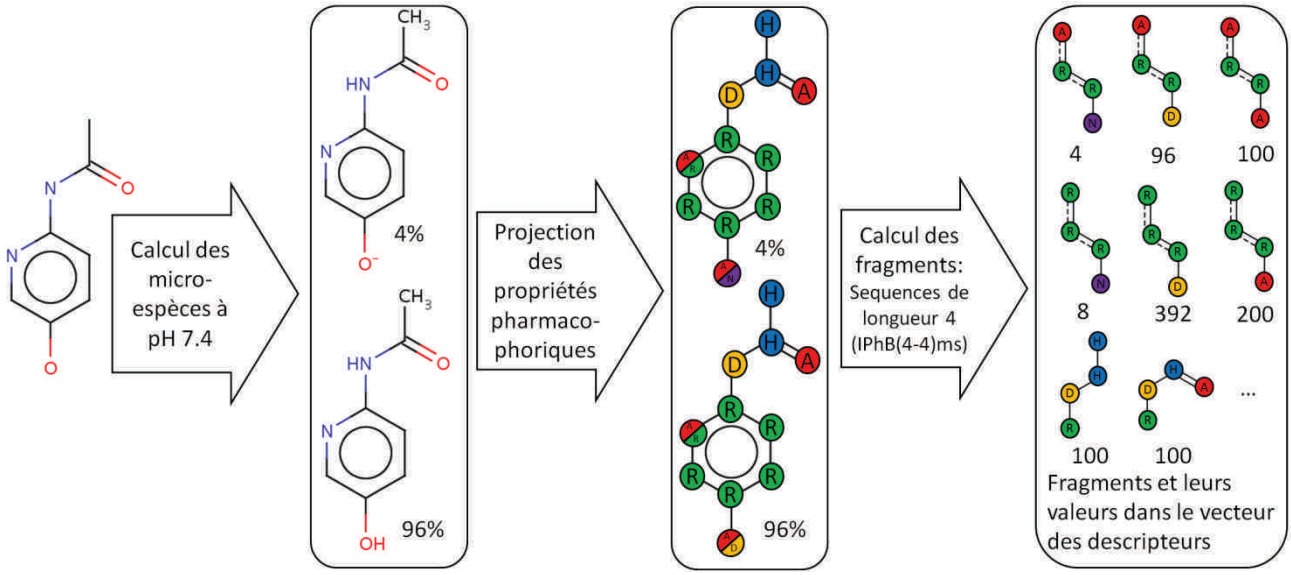


Figure 2. Processus du calcul de descripteurs ISIDA avec une projection des caractéristiques pharmacophoriques sur le graphe en prenant en compte le pH. La valeur d'un fragment dans le vecteur de descripteurs correspond à la somme des occurrences multiplié par la population en % de chaque micro-espèce. Des exemples de la valeur du vecteur pour une topologie de type séquence d'une longueur 4 avec l'information des liaisons sont présentés dans la dernière case.

La représentation de la dépendance au pH se fait par le compte des fragments, la fragmentation est effectuée sur les différentes micro-espèces à un pH donné (par défaut 7.4). La valeur d'un fragment dans le vecteur correspond alors à la somme sur toutes les micro-espèces de son occurrence dans la micro-espèces multiplié par la population de la micro-espèce en pourcentage. De plus, seule la partie entière de cette valeur est conservée pour garder une représentation en nombre entier des valeurs de descripteurs (voir équation 1 ci-dessous).

$$valeur_{desc1} = \left(\sum_{ms} occurence_{desc1,ms} \times population_{ms} \times 100\right)_{int} \quad (1)$$

où desc1 est un descripteur du vecteur et ms est la micro-espèce.

Ceci est illustrée dans la Figure 2, par exemple, le fragment D-R*R*R (fragment central dans le cadre des fragments à droite de la Figure 2) est retrouvé 2 fois dans la micro-espèce à 4% et 4 fois dans la micro-espèce à 96%, on obtient donc un compte de 2*4+4*96=392. Il est à noter que certains atomes peuvent être pourvu de plusieurs caractéristiques comme l'azote du cycle aromatique qui est considéré aromatique (R) et accepteur de liaison hydrogène (A). Dans ce cas, tous les fragments sont générés systématiquement avec les deux caractéristiques.

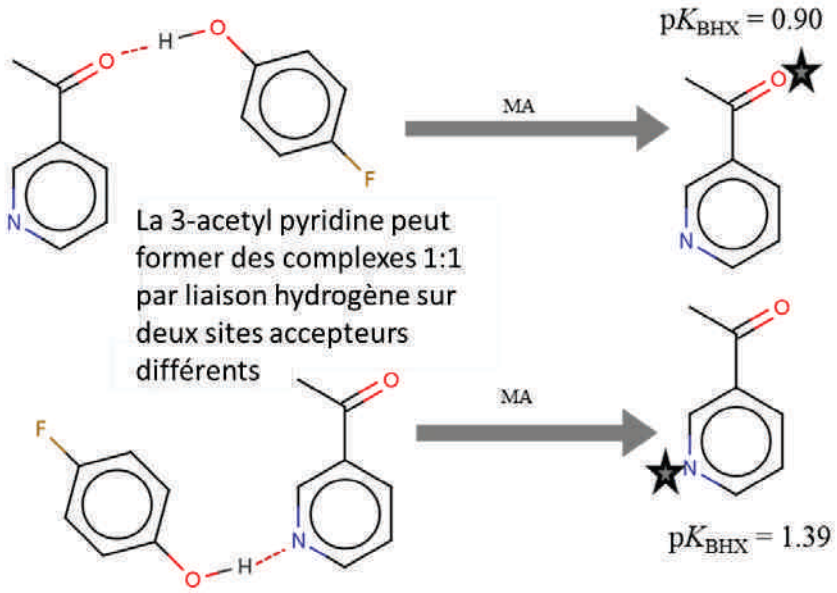


Figure 3. Exemple de marquage d'atome (MA) pour indiquer où se forme la liaison hydrogène avec la 3-acetyle pyridine.

Une des autres options concerne la possibilité d'indiquer la localité d'une propriété avec le « marquage d'atomes » (voir Figure 3). Une fois l'atome marqué sur le graphe moléculaire, les fragments sont calculés selon différentes restrictions ou ajout :

1. Uniquement les fragments commençant par un atome marqué sont générés (MA1).

2. Uniquement les fragments contenant l'atome marqué sont générés (MA2).

3. Tous les fragments sont générés mais l'atome marqué est indiqué par une marque particulière (MA3).

Une nomenclature des descripteurs est proposée pour refléter les options utilisées entre autres, les codes pour la topologie choisie (I, II ou III), ensuite la propriété projetée (A, Ph, Pc, Ep, Lp), l'inclusion de l'information sur les liaisons (avec un B pour « bond »), la longueur minimum et maximum des fragments entre parenthèse, le type de compte d'occurrence utilisées (rien si juste l'occurrence, ms si dépendent du pH) et ensuite les options particulières sont indiqués (P, MA*X*) sont explicités.

Les descripteurs ISIDA sont ainsi devenus très versatiles. Ils sont adaptables pour un problème donné et une grande liberté est laissée à l'utilisateur.

3. Applications des descripteurs ISIDA

Plusieurs études ont été effectuées en utilisant les descripteurs ISIDA : une étude d'étalonnage sur le CVS sur des protéases et plusieurs modèles de QSAR. Les études de QSAR qui présentent des similarités méthodologiques à chaque étude ont été résumées dans le Tableau 1.

3.1 CVS des protéases

Les espaces de descripteurs ISIDA ont été rétrospectivement ajoutés à une étude de CVS qui impliquait déjà 50 autres espaces de descripteurs. L'étude est basée sur un ensemble de données de 2500 composés dont les concentrations inhibitrices médianes (pIC_{50}) ont été mesurées vis-à-vis de 5 protéases à sérine différentes.

Chaque actif connu de chaque protéase est considéré comme composé de référence pour le CVS à tour de rôle et est comparé au reste de la base de données, les 2499 autres molécules. Les espaces de descripteurs sont classés sur chaque recherche selon le nombre d'autres molécules actives retrouvées comme similaires. Leur classement moyen sur toutes les requêtes effectuées et la déviation standard correspondante sont utilisés pour évaluer leur performance globalement. Les « bons » espaces de descripteurs devraient préférablement avoir un rang global petit et une petite déviation standard. La Figure 4 résume les résultats pour les 31 premiers descripteurs du classement. Les descripteurs ISIDA y sont prédominants et ont généralement une petite déviation standard.

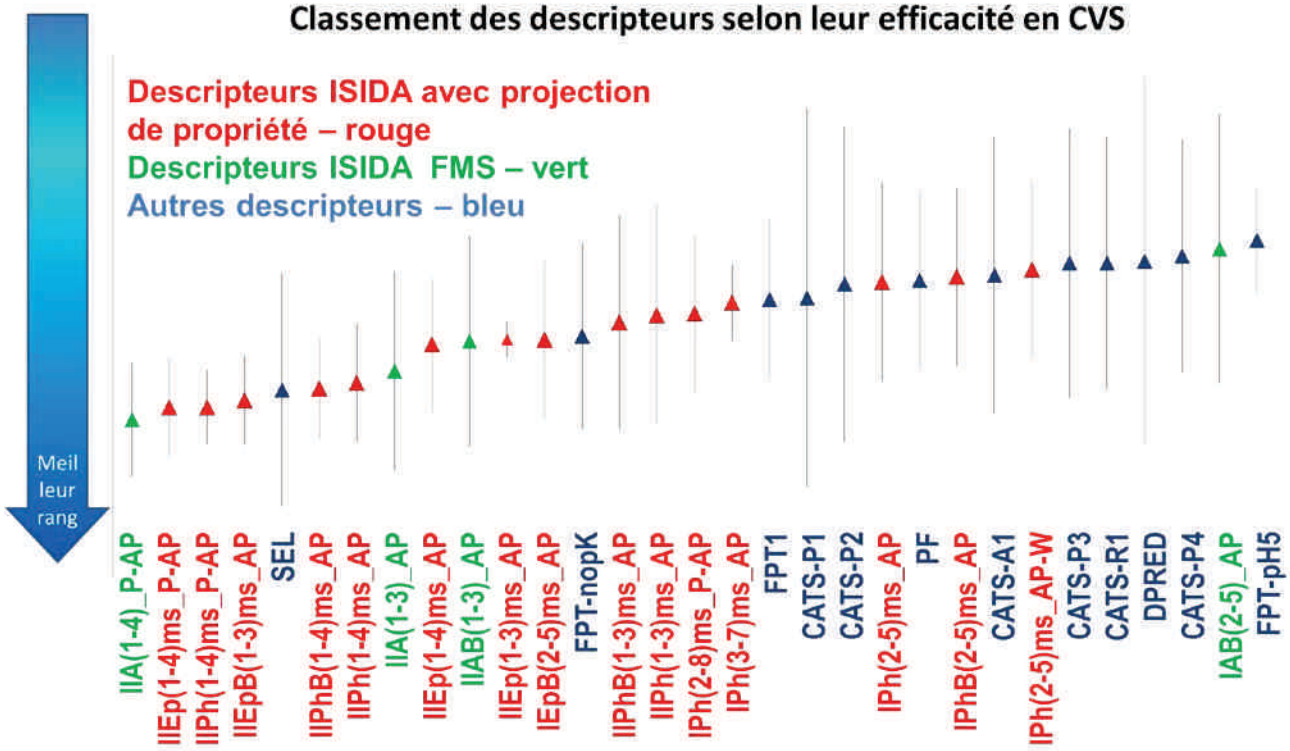


Figure 4. Résultat du classement des différents espaces de descripteurs dans l'étude CVS sur 5 protéases

3.2 Différentes études QSAR

Un modèle QSAR est défini selon la propriété cible, les données utilisées pour le jeu d'entraînement, les descripteurs utilisés, la méthode d'apprentissage, les performances sur un jeu externe de test ou avec une procédure de validation externe. Ceci est résumé dans le Tableau 1 pour chacune des études ainsi qu'une comparaison à la littérature.

L'évaluation d'un modèle QSAR est résumée par quelques paramètres statistiques. Si la propriété choisie est de type binaire qualitative, les performances du modèle peuvent être évaluées par la Précision Balancée (Balanced Accuracy - BA) (voir équation 1). Si la propriété est quantitative et prend des valeurs réelles, celles-ci pourront être évaluées par l'erreur quadratique moyenne (Root-Mean-Square Error - RMSE) (voir équation 2) entre les valeurs de molécules connues ($Y_{expérimental}$) et celle prédites par la fonction ($Y_{prédit}$).

$$BA = 0{,}5 \times \left(\frac{Nb.de\ prédictions\ correctes(classe1)}{Nb.total\ de\ composés\ (classe1)} + \frac{Nb.de\ prédictions\ correctes(classe2)}{Nb.total\ de\ composés(classe2)}\right) \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum(Y_{prédit} - Y_{expérimental})^2}{nb.de\ composés}} \qquad (2)$$

Il est préférable d'avoir une BA la plus proche de 1 possible qui correspondrait à une prédiction parfaite des classes. Pour la RMSE, il est préférable de l'avoir la plus faible possible ; 0 correspondrait à une prédiction parfaite de la propriété.

Dans le Tableau 1, les performances sont présentées par la RMSE ou la BA selon que les propriétés sont quantitatives (indiqué avec un R) ou qualitatives (indiqué avec un C). Plusieurs espaces de descripteurs ISIDA en combinant les différentes propriétés et topologies ainsi que les options ont été utilisés. Dans le Tableau 1, seules les topologies, les propriétés projetées sur le graphe moléculaire et la stratégie de marquage d'atome ont été indiquées par soucis de simplification.

Tableau 1. Résumé des applications des descripteurs ISIDA à la modélisation QSAR

Propriété	Descripteurs ISIDA	Taille du jeu d'entraînement	Méthode d'apprentissage	Taille de jeu externe de test ou procédure de validation	RMSE (R) ou BA (C)	Comparaison à la littérature
Concentration inhibitrice médiane de la hERG[a] (pIC_{50})	Topologie : I, II, Projection : A, Ph, Ep	562	Consensus de régressions multilinéaires	1889 (191 pouvant bloquer la fonction de la hERG)	0.66 (C)[b]	Le modèle de Li et al. est considéré comme référence avec une BA=0.60. L'identification des molécules pouvant bloquer la hERG est plus important que la prédiction correcte des non-bloquants. Dans ce domaine, notre meilleur modèle est plus performant, avec un rappel[c] de 0.76 alors qu'ils obtiennent 0.57.
Coefficient de partage octanol-eau (logP)	Topologie : I, II, Projection : A, Ph, Ep	3225	Consensus de régressions multilinéaires	Test1:9677	0.75 (R)	Une RMSE de 0.75 pour la prédiction du logP est raisonnable. Par exemple, les prédictions logP de ChemAxon obtiennent une RMSE de 0.76 sur ce jeu alors qu'il est très probable que ces molécules sont connues par le logiciel.
				Test2: 226	0.78 (R)	Ce second test provient d'une étude d'étalonnage de différentes méthodes à laquelle nous nous sommes comparés. La meilleure méthode obtenait un RMSE de 0.80 que nous surpassons légèrement.
Index chromatographique d'hydrophobicité (CHI)	Topologie : I, II, III Projection : A, Ph, Ff, Pc, Ep, Lp	485	Consensus de machines à vecteurs supports	195	16.4 (R)	Le modèle se comporte raisonnablement bien. Il n'existe néanmoins pas de modèle de référence pour cette propriété. Une RMSE de 16.4 correspond à une RMSE d'environ 0.8 en terme LogP. Certaines molécules du jeu externe sont douteuses car mesurées différemment de celle du jeu d'entrainement. **Cette étude a permis de mettre en évidence l'utilité du QSAR pour détecter des mesures erronées.**
Affinité de liaison de GPCRs[d]	Topologie : I, II, III Projection : A, Ph, Ff	10000 (5000 actifs)	Consensus de machines à vecteurs supports avec approche chemogénomique	10000 (5000 actifs)	0.87 (C)	Les modèles de référence sur la même étude avec des descripteurs Accelrys (ECPF) obtiennent au mieux une BA de 0.85. Nos modèles sont légèrement plus performants mais en particulier les espaces de descripteurs Ff ont un meilleur comportement dans le calcul de similarité.
Constante d'équilibre de l'acidité (pK_a)	Topologie : I, II, III Projection : A, Ph, Pc MA1	188 (142molécules)	Consensus de régressions multilinéaires, régression des moindres carrés partiels	Validation croisée à trois paquets	1.23 (R)	Ils n'y a pas assez de données et les molécules comportent des fonctions très diverses. Les prédictions de ChemAxon arrivent à une RMSE de 0.97. Cette étude est toujours en cours de développement mais les résultats sont encourageants.
Constante d'équilibre de complexe par liaison hydrogène pour accepteurs (pK_{BHX})	Topologie : I, II, III Projection : A, Ph, Ff, Pc MA1, MA2, MA3	542 (537molécules)	Consensus de machines à vecteurs supports, Consensus de régressions multilinéaires	452 (425 molécules)	0.26-0.29 (R)	Les performances des modèles varient selon la définition du domaine d'applicabilité. : 0.26 avec 75 valeurs prédites et 0.29 avec 129 valeurs prédites. Les modèles obtiennent également de bons résultats sur les molécules bifonctionnelles. Il est difficile de se comparer à la littérature car ce sont les premiers modèles avec des chemotypes aussi divers et un test aussi grand. Le meilleur modèle de la littérature trouvée est celui de Besseau et al. qui adressent uniquement des accepteurs avec un azote et obtient une RMSE de 0.13 sur 142 composés.

[a] hERG = human « *Ether-à-go-go*-Related Gene », [b] Les données du jeu externe sont qualitatives [c] Rappel = (Nb. de positifs prédits correctement)/(nb. total de positifs) [d] GPCR = récepteurs couplés aux protéines G

4. Conclusion

1. Les nouveaux descripteurs ISIDA constituent une généralisation de descripteurs déjà existants et permettent d'extraire un vaste spectre d'informations chimiques. Ils augmentent les chances de construire un bon modèle dans les phénomènes chimiques qui sont par nature très complexes. Ils peuvent tenir compte de certaines particularités d'un système par le biais des fragments locaux pour les molécules polyfonctionnelles ou par la prise en compte du pH pour les problèmes en solution. Un plus grand nombre de modèles différents pertinents permet de construire un meilleur modèle consensus, cad. la moyenne des prédictions des modèles construits avec différents descripteurs est utilisée comme prédiction.

Les nouveaux descripteurs ISIDA ont à plusieurs reprises dépassé d'autres descripteurs. De plus, ils sont facilement interprétables, car faisant référence directe à la structure chimique. Ils ont servi avec succès à l'élucidation de plusieurs relations structure-propriété importantes pour la chimie médicinale.

2. Différentes modélisations effectuées avec les nouveaux descripteurs montrent leurs avantages par rapport aux descripteurs précédents.

- Pour le CVS, ils augmentent considérablement le nombre d'espaces pertinents.
- En utilisant nos descripteurs « locaux », les premiers modèles QSAR prédictifs pour la constante d'équilibre de complexe par liaison hydrogène pour accepteurs (pK_{BHX}) des molécules polyfonctionnelles ont été développés.
- Les nouveaux descripteurs ISIDA ont permis de trouver des molécules dont les mesures étaient erronées dans la modélisation par QSAR de l'index chromatographique d'hydrophobicité. Ces observations ont été confirmées expérimentalement.
- Dans le cadre de la modélisation chémogénomique, ils surpassent légèrement d'autres descripteurs mais en particulier, les descripteurs à identifiants de champs de force et caractéristiques pharmacophoriques ont une meilleure distribution des valeurs de la matrice de similarité. Contrairement aux autres descripteurs qui ont tendance à considérer la plupart des molécules comme similaires (distribution avec un pic dans les valeurs très similaires), les valeurs de similarité obtenus avec les descripteurs ISIDA s'étendent en allant de très dissimilaire à très similaire.
- De par leurs nombreuses possibilités, ces descripteurs peuvent s'adapter pour un problème donné, en codant les aspects physico-chimiques les plus pertinents pour ce problème. Nous avons bon espoir que les incréments obtenus d'un modèle QSAR puissent constituer une source de très bons descripteurs pour un problème lié à la propriété du modèle utilisé, bien que les premiers tests aient été peu conclusifs. Il reste encore beaucoup de travail à faire dans ce sens.

3. Afin de rendre les modèles d'hydrophobicité et de l'affinité des accepteur de liaison hydrogène accessibles aux utilisateurs, ils ont été implémentés dans les services web du laboratoire gratuitement.
(http://infochim.u-strasbg.fr/webserv/VSEngine.html)

Contents

Chapter 1

Introduction

1.1 Chemoinformatics and molecular descriptors

Chemistry, in all its varying aspects, is a field of experiments, which produces data. Since the advent of combinatorial chemistry and of robotised experiments known as High-Throughput Screening (HTS) in combination with the possibility of storage on computer, huge databases have been generated and are in need to be analysed. Some of those databases are freely available as for example PubChem (http://pubchem.ncbi.nlm.nih.gov) from the National Library of Medicine, U.S. National Institutes of Health, which became public in 2004 or ChEMBL (https://www.ebi.ac.uk/chembldb/), from the European Bioinformatics Institute. The information on various assays and chemical properties on a huge number of molecules is available and in order to extract chemical knowledge from those, the interdisciplinary field of chemoinformatics[1–4] has emerged[5].

Chemoinformatics uses computational tools in order to manage, interpret and extract knowledge from chemical information. The term chemoinformatics or cheminformatics appeared in the 1990s. It is commonly admitted to have been first defined by Frank Brown in 1998[6] as: "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organisation.". Historically, chemoinformatics is a field that emerged from pharmaceutical research, however, it can be applied to any field of chemistry (given that the tools are triggered towards the problem). Its major use, however, lies in the medicinal chemistry and pharmaceutical fields, more specifically in drug discovery processes.

Drug discovery is a costly and long procedure. On average it takes 15 years for a drug to reach the market with a cost counted in several hundreds or thousands of millions of Euros[7]. Use of chemoinformatics strategies is a key step in the development of drugs in order to find potential binding candidates to target proteins and to evaluate their properties in Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME/T) so as to reduce the high experimental costs involved in drug discovery.

The classical principles, tools and machine learning algorithms incorporated in chemoin-

formatics were often developed for solving data mining in other domains[8,9], nevertheless, research efforts are still needed in applying those "in silico" to chemical problems. In particular, the computer-based representation of chemical structures is essential for the success of chemoinformatics approaches, and several levels of representations are used. The molecular graph (representing molecules as graphs, with atoms in the nodes connected by edges-bonds) is essential for storing and drawing/displaying molecular structures, but is not well suited for analysis and knowledge extraction. One obvious reason is that the graph representation depends on the arbitrary atom numbering scheme (node "1" of the graph may be arbitrarily assigned to any atom – there are many numbering schemes possible to represent a same structure). Furthermore, a graph is a topological concept – easily apprehended by the human mind, but not well suited for straightforward algebraic operations. Therefore, a secondary level of structure representation is molecular descriptors: structural information is extracted from the molecular graph, and encoded in a straightforward numerical format – typically, by a vector of numbers, D_i, in which each component i stands for a specific structural feature. Such descriptor vectors are well-suited for mathematical structure-activity analysis, and notably for

(a) Similarity-based Virtual Screening (SimVS), based on the classical "similarity principle" [10]: similar compounds are likely to display similar properties. In mathematical formulation, similar molecules are molecules with covariant descriptor vectors. It is therefore possible to calculate the degree of covariance between the "reference" vector describing an active compound and vectors of molecules from a database, in order to select the most covariant ones, and implicitly the most likely to be active.

(b) Quantitative Structure-Activity/Property Relationships (QSAR or QSPR), where machine learning techniques are employed to search for possible empirical equations expressing a molecular property Y (say affinity to a therapeutic target) as a mathematical function of selected vector elements D_i. If such a function (the simplest being a linear dependence $Y = a_0 + \sum a_i * D_i$) is established and shown to return close approximations of the Y value of known molecules, it can be used to predict Y values of virtual compounds, because D_i terms can obviously be calculated prior to actual molecule synthesis. Compounds with desired Y values can therefore be prioritised for synthesis and testing.

The art of conceiving molecular descriptors consists in deciding what kind of structural features one should focus on. Obviously, SimVS and QSAR only work if property-relevant information is captured in vector D. Historically, the start of molecular descriptors is often associated to the work of Hammett in 1937. He described reaction rates and equilibrium constants of reactions involving substituted benzoic acids according to two parameters based on the substituents and their position (meta or para). The idea is that the change between one reaction to another is only the substituent, therefore, the change in energy can be related to the substituent. This analysis was named linear free-energy relationships (LFER). In 1964, Hansch and Fujita introduced a parameter based on the differences in the n-octanol/water partitioning coefficient (LogP) introduced by substituents[11]. They were able to successfully relate biological activity, such as the effect of insecticides on houseflies, to it. These pioneer works constitute the beginning of molecular descriptors and QSPR/QSAR simultaneously. Since, many descriptors have been introduced for a wide variety of purposes. Todeschini and Consonni regroup a great number of them in their

book[12] where they reference over 3300 publications about molecular descriptors.

Descriptors can be classified into different categories[13]. Among them, properties such as Hammett, Hansch and Fujita used, are important and useful but require costly experiments. As mentioned previously, knowledge can be extracted from the molecular graph and these type of descriptors are often categorised as follows:

- **1D-descriptors** are calculated from the chemical formula. Examples are the molecular weight and the atom count[14].
- **2D-descriptors** are calculated from the 2D molecular graph of the molecule. In this category, the topological indices, such as Wiener[15] and Randić indices[16], descriptors encoding property information such as atom pairs[17], multiplets of pharmacophore[18–20], BCUT descriptors[21], ECPF descriptors[22] and fragment descriptors[23].
- **3D-descriptors** are calculated from the 3D structure of the molecule. Some of the 2D-descriptors(multiplets, BCUT) have analogue 3D counterparts by changing the topological distance to the actual distance in the 3D structure. This category also includes quantum-chemical calculation-based descriptors, size, steric, surface and volume descriptors[24–29] and WHIM descriptors[30]. Another type of 3D-descriptors are based on calculated properties such as the electrostatic, hydrophobic or hydrogen-bonding potentials, in "all" points of space surrounding the molecule (often on a grid). The GRID[31] and the Comparative Molecular Field Analysis (CoMFA) descriptors[32] approaches are the most popular of these last descriptors.

Note that, these descriptors can be used to model a property and the prediction of the model can then be used as a descriptor. This permits to inject knowledge learned on behalf of a related property into the model of the current property, by means of so-called feature nets, a specific instance of "inductive learning transfer".

1.2 The ISIDA property-labelled descriptors

The "Laboratoire de Chémoinformatique" of the University of Strasbourg has developed fragment descriptors consisting of sequences and augmented atoms of the atoms' element symbol[33,34] as part of their chemoinformatics suite, named In SIlico design and Data Analysis, in short ISIDA. Fragment descriptors are counts of substructures of a molecular graph – for example, D_1=number of C=O groups, D_2=number of C-N-C fragments, etc. Substructures counts are intuitively understandable and although they are simple they perform well in QSAR[35]. Interpretation of molecular properties by identifying and analysing different fragments has always been in chemists' minds. For example an organic chemist in need to determine whether a compound is soluble in water will look at the molecular graph and key out charged groups for their hydrophilic effect and longer chains of carbons to oppose it.

By "counting" these different fragments a very rough approximation of the solubility can be made and thus answer the initial question. As a matter of fact, many theoretical methods to evaluate the octanol-water partition coefficient (LogP) are based on increments of different fragments or types of atoms[36,37].

The substructural ISIDA descriptors perform well in modelling as they seem to catch essential connectivity-related information. This may be both a strength and a limitation, depending whether the studied property is connectivity-controlled or not. For example, querying for similar compounds in fragment-based SimVS searches will typically retrieve nearest neighbours based on a same scaffold as shown in Figure 1.1.

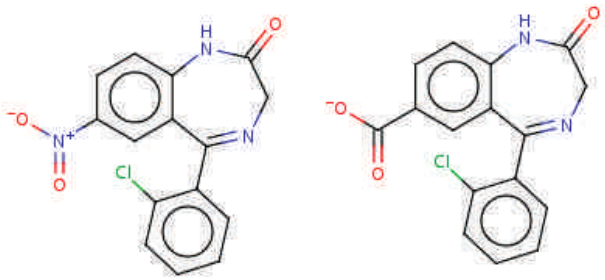


Figure 1.1: Similarity searches based on fragment descriptors are bound to conserve the scaffold, but may allow for variations of substituents – in this case, replacement of a neutral nitro by a charged carboxylate. This matches the empirical perception of molecular similarity by medicinal chemists.

However, connectivity is not the only aspect controlling molecular properties. For example, ligand binding to a biological receptor is controlled by specific anchoring points, defining the "pharmacophore". As far as two molecules have similar pharmacophore patterns, i.e. similar arrangements of physico-chemically equivalent groups ("pharmacophore features": hydrophobic, hydrogen bond partners, charges, etc.), they might be eligible to bind a same target even if they are based on different skeletons. Pharmacophore descriptors, counting, instead of fragments, the occurrence of combinations (pairs, triplets) of pharmacophore features in molecules, may capture this alternative aspect of molecular similarity[38], not easily perceived by the human mind.

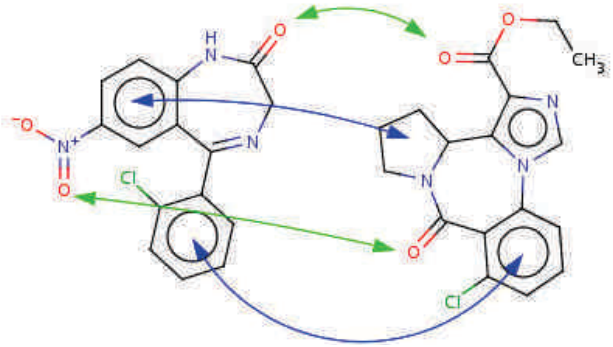


Figure 1.2: Pharmacophore descriptor based analogue in SimVS, which is a benzodiazepine ligand like the reference compound, but does not base on the benzodiazepine skeleton. However, it does have equivalent pharmacophore groups located similarly to the ones in the reference (arrows highlight the equivalences).

Figure 1.2 is an example of "scaffold hopping" (discovery of an alternative scaffold having a similar biological activity) obtained with Fuzzy Pharmacophoric Triplet counts (FPT),

also developed in the "Laboratoire de Chemoinformatique" [20,39]. Scaffold hopping is very appreciated by medicinal chemists, because it may allow "escaping" the patent space covered by a series of scaffold-centric analogues, produce molecules with different pharmacokinetic properties, etc. However, none of these two complementary views on molecular description/similarity is intrinsically better than the other. Indeed, biological activity is jointly controlled by pharmacophore and connectivity issues – remote jumps in structure space, as supported by pharmacophore descriptors are potentially rich in benefits, but nevertheless risky: dramatic changes in molecular connectivity may cause loss of activity (for various reasons: change in flexibility/binding entropy, electron density variations in aromatic rings, etc).

The first aim of this work was, logically, to bridge the gap between these two extreme views: the strict fragment-based, and the fuzzy pharmacophore-centric point of view, by studying hybrid descriptors combining the advantages of both approaches. These were named ISIDA Property-Labelled Fragment (IPLF) descriptors and the work was published in *Molecular Informatics* [40]. A full description of the ISIDA Property-Labelled Fragments (IPLF) descriptors is given in chapter 3. This development relied on all the previous experience of the laboratory team in this field, notably the management of ionization effects. These may be important for understanding otherwise inexplicable "activity cliffs" (significant activity differences of apparently very close molecules) [41]. Figure 1.3 is an example of such activity cliffs detailed in the FPT publication [20]. Thus, this aspect was also integrated as part of the calculation workflow in the IPLF descriptors.

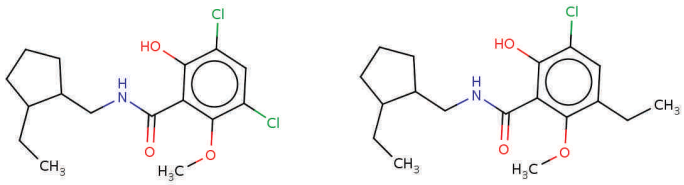


Figure 1.3: State-of-the-art similarity evaluations would all agree that these compounds are virtually identical. The FPT-based similarity scoring does not, due to its pK_a-sensitive pharmacophore feature flagging scheme, and is right not to return a similarity score close to perfect matching, because these molecules are actually displaying significant differences in terms of biological activities (please refer to the original publication [20] for details).

As a next step, "colouring" the molecular graph was pushed beyond pharmacophore types or atomic symbols and properties including partial charges, topological electrostatic potentials (based on partial charges), force field typing and octanol-water partition coefficient (LogP) increments were mapped onto the molecular graph. Moreover, since the modelling of certain properties (hydrogen bonding strength, acidic dissociation constant) requires focusing on specific atomic centres, IPLF were extended to support locality-related information. Indication of explicit formal charges was also added in order to represent the different deprotonation states of a molecule with several acidic centres. Additionally, increments extracted from QSAR models [42] were used to colour fragments by local incremental contributions to modelled properties.

However intellectually appealing and rigorously designed, the ultimate utility of descriptors cannot be proven but by building predictive models and using them to make experi-

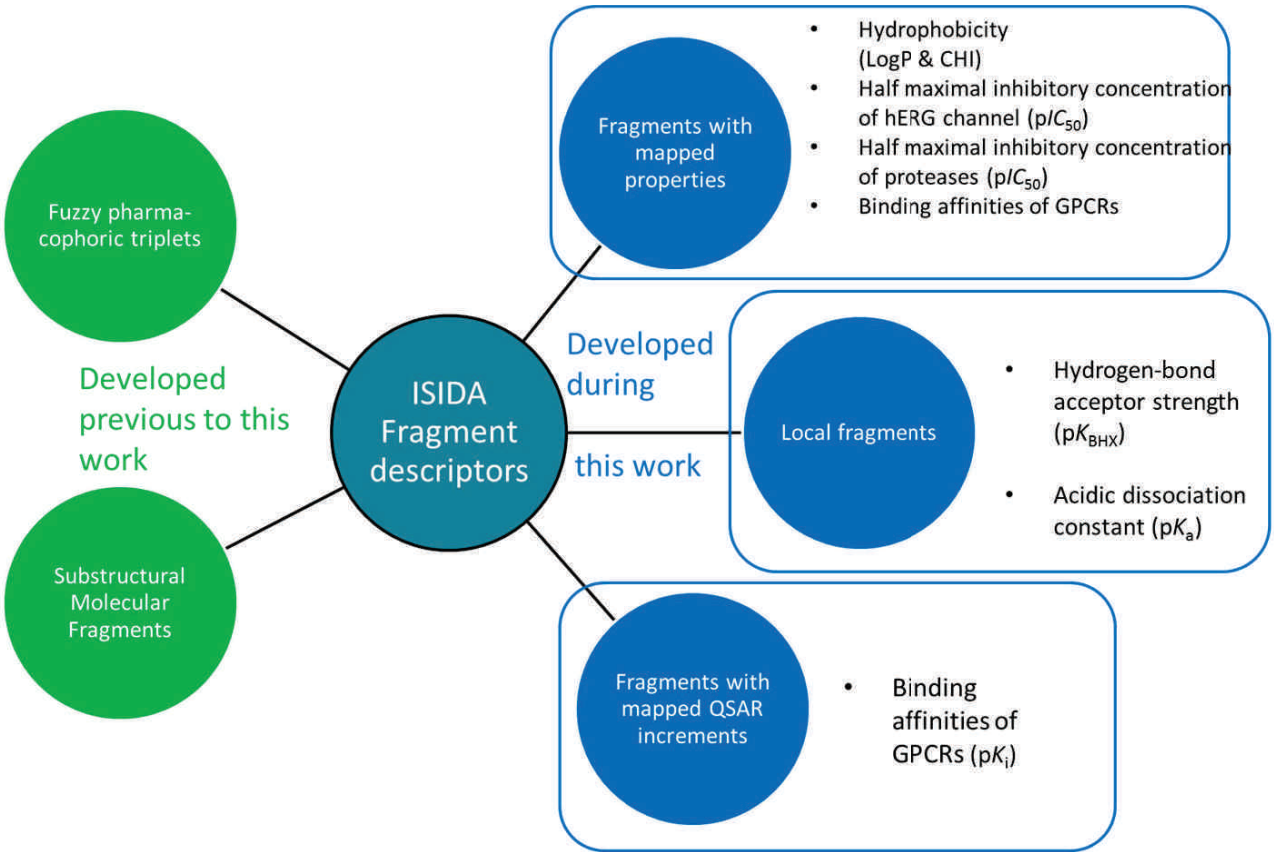


Figure 1.4: ISIDA fragment descriptors

mentally verifiable predictions. This key part of this work includes one SimVS study on the binding affinity of proteases (see section 4.2) and different QSAR/QSPR endpoints relevant to the pharmaceutical realm:

- Hydrophobicity (see section 4.2 for LogP and section 4.3 for the Chromatographic Hydrophobicity Index (CHI))
- Half maximal inhibitory concentration to the human Ether-à-go-go-Related Gene (hERG) channel (see section 4.2)
- Binding affinity to G Protein-Coupled Receptors (GPCR) (see section 4.4)
- Hydrogen bonding acceptor strength (see section 5.3)
- Acidic dissociation constant (see section 5.2)

All details on the extension of ISIDA descriptors are given in section 3 and Figure 1.4 gives an overview of the available types of descriptors developed during this thesis and their application. General methodology about SimVS and QSAR are given in 2. The appendices contain the manual for the ISIDA descriptor calculation and the supporting information of the articles included in this thesis.

Bibliography

[1] Gasteiger, J. *Anal Bioanal Chem* **384**(1), 57–64 (2006).

[2] Engel, T. *J Chem Inf Model* **46**(6), 2267–2277 Nov (2006).

[3] Leach, A. R. and Gillet, V. J. *An Introduction to Chemoinformatics*. Springer Verlag, (2007).

[4] Varnek, A. and Baskin, I. I. *Mol Inf* **30**(1), 20–32 (2011).

[5] Russo, E. *Nature* **419**, 4–7 Sept (2002).

[6] Brown, F. *Annu Rep Med Chem* **33**, 375–384 (1998).

[7] Khanna, V. and Ranganathan, S. *Drug Development Research* **72**(1), 74–84 (2011).

[8] Hann, M. and Green, R. *Curr Opin Chem Biol* **3**(4), 379–383 (1999).

[9] Wegner, J. K., Sterling, A., Guha, R., Bender, A., Faulon, J.-L., Hastings, J., O'Boyle, N., Overington, J., van Vlijmen, H., and Willighagen, E. *Commun ACM* **55**(11), 65–75 Nov (2012).

[10] Johnson, A. M. and Maggiora, G. M. *Concepts and Applications of Molecular Similarity*. John Willey & Sons: New York, (1990).

[11] Hansch, C. and Fujita, T. *J Am Chem Soc* **86**(8), 1616–1626 Apr (1964).

[12] Todeschini, R. and Consonni, V. *Handbook of Molecular Descriptors*. Wiley-VCH, (2008).

[13] Raevsky, O. A. *Russ Chem Rev* **68**(6), 505–524 (1999).

[14] Carbó-Dorca, R. and Gallegos Saliner, A. *J Comp Chem* **30**(13), 2099–2104 (2009).

[15] Wiener, H. *J Am Chem Soc* **69**(1), 17–20 (1947).

[16] Randić, M. *J Am Chem Soc* **97**(23), 6609–6615 (1975).

[17] Carhart, R. E., Smith, D. H., and Venkataraghavan, R. *J Chem Inf Comput Sci* **25**(2), 64–73 May (1985).

[18] Kearsley, S. K., Sallamack, S., Fluder, E., Andose, J. D., Mosley, R. T., and Sheridan, R. P. *J Chem Inf Comput Sci* **36**, 118–127 (1996).

[19] Schneider, G., Neidhart, W., Giller, T., and Schmid, G. *Angew Chem Int Ed Engl* **38**(19), 2894–2896 Oct (1999).

[20] Bonachera, F., Parent, B., Barbosa, F., Froloff, N., and Horvath, D. *J Chem Inf Model* **46**(6), 2457–2477 Nov-Dec (2006).

[21] Pearlman, R. and Smith, K. *J Chem Inf Comput Sci* **39**(1), 28–35 (1999).

[22] Rogers, D. and Hahn, M. *J Chem Inf Model* **50**(5), 742–754 May (2010).

[23] Varnek, A. and Baskin, I. I. *Chapter 1 "Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening" in Chemoinformatics Approaches to Virtual Screening.* Royal Society of Chemistry, (2008).

[24] Higo, J. and Gō, N. *J Comp Chem* **10**(3), 376–379 (1989).

[25] Stanton, D. T. and Jurs, P. C. *Anal Chem* **62**(21), 2323–2329 (1990).

[26] Katritzky, A., Mu, L., Lobanov, V., and Karelson, M. *J Phys Chem* **100**(24), 10400–10407 (1996).

[27] Weiser, J., Weiser, A., Shenkin, P., and Still, W. *J Comp Chem* **19**(7), 797–808 (1998).

[28] Labute, P. *J Mol Graph Model* **18**(4-5), 464–477 (2000).

[29] Xi, Z., Yu, Z., Niu, C., Ban, S., and Yang, G. *J Comp Chem* **27**(13), 1571–1576 (2006).

[30] Todeschini, R. and Gramatica, P. *SAR QSAR Environ Res* **7**, 89–115 (1997).

[31] Goodford, P. J. *J Med Chem* **28**(7), 849–857 (1985).

[32] Cramer, R. D., Patterson, D. E., and Bunce, J. D. *J Am Chem Soc* **110**, 5959–5967 (1988).

[33] Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. *J Comput Aid Mol Des* **19**(9-10), 693–703 Jul (2005).

[34] Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I., and Marcou, G. *Curr Comput Aided Drug Des* **4**(3), 191–198 Sept (2008).

[35] Lounkine, E., Batista, J., and Bajorath, J. *Curr Med Chem* **15**, 2108–2121 (2008).

[36] Mannhold, R., Poda, G., Ostermann, C., and Tetko, I. *J Pharm Sci* **98**(3), 861–893 Mar (2009).

[37] Ghose, A. K. and Crippen, G. M. *J Comput Chem* **7**(4), 565–577 (1986).

[38] Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. *J Med Chem* **39**(16), 3049–3059 (1996).

[39] Bonachera, F. and Horvath, D. *J Chem Inf Model* **48**(2), 409–425 Feb (2008).

[40] Ruggiu, F., Marcou, G., Varnek, A., and Horvath, D. *Mol Inf* **29**(12), 855–868 Dec (2010).

[41] Maggiora, G. M. *J Chem Inf Model* **46**(4), 1535 (2006).

[42] Marcou, G., Horvath, D., Solov'ev, V., Arrault, A., Vayer, P., and Varnek, A. *Mol Inf* **31**, 639–642 (2012).

Chapter 2

Methodology

2.1 Introduction

2.1.1 Similarity-based Virtual Screening

Virtual screening is a major aspect of chemoinformatics and aims to search for new bioactive molecules using *in silico* methods. Two main categories of virtual screening exist: (a) Virtual screening based on the structure of the target protein and (b) Virtual screening based on the structure of the ligand. In this work, only ligand-based virtual screening based on molecular descriptors has been used in order to benchmark the ISIDA fragment descriptors. As mentioned in the introduction, similarity-based virtual screening (SimVS)[1–6] is based on the similarity principle: "A molecule similar to a known active molecule should be active as well". Similarity between molecules is evaluated using a particular descriptor space, which is constituted of the selected descriptors, in combination to a similarity metric (see section 2.3.4), which is the mathematical expression used to measure the similarity. Known actives are compared to molecules in a database and molecules considered to be potentially active are the most similar ones found. The usual SimVS study follows the following steps (see Figure 2.1))

1. Selection of known active molecules and of a candidate database to mine for similar analogues.
2. Standardisation of structures (explained in 2.2)
3. Calculation of molecular descriptors
4. Calculation of the similarity matrix
5. Assessment of the similarity of each candidate with respect to the query

In order to test the relevance of the descriptors used, the database contains molecules with known activity.

The tendency of neighbouring molecules in descriptor space to be close neighbours in the activity space is called the Neighbourhood behaviour[7]. It can be evaluated by certain criteria (see section 2.3.5) and the best descriptor space/similarity metric pair can be determined by it. The methodology also allows to answer a key question of SimVS, "how

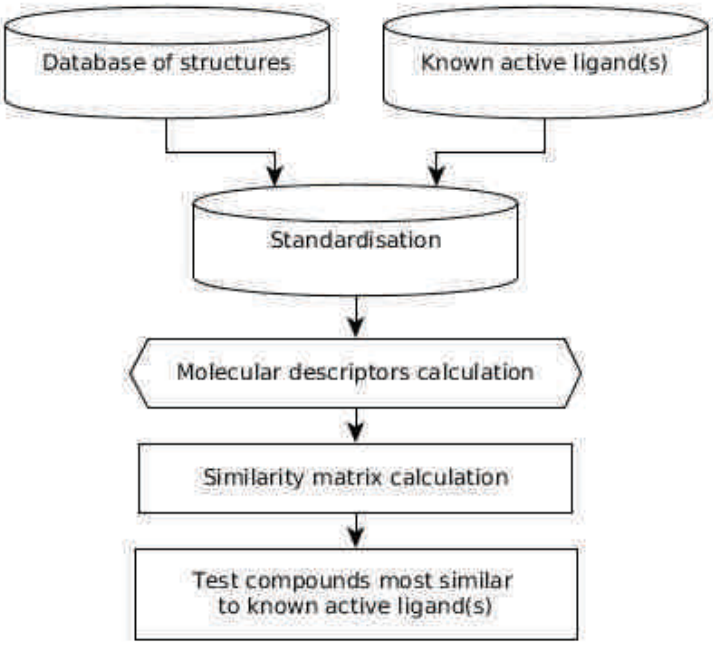


Figure 2.1: Typical workflow for ligand-based virtual screening using molecular descriptors and a similarity metric

similar is similar", e.g. how to set the similarity radius such as to ensure that candidates within that neighbourhood of the query are both numerous and optimally enriched in actives.

2.1.2 Quantitative Structure-Activity/Property Relationships

A QSAR/QSPR model[8] is a mathematical expression of a relationship between an activity/property and the structure of a molecule. The first attempt to analyse biological activity in a QSAR frame was in 1964 by Hansch and Fujita[9]. Since, many descriptors and machine learning approaches have been developed and the field has evolved from trying to predict congeneric series of molecules to more global models to apply to a wider diversity of molecules. Building the model is done in several steps (see Figure 2.2):

1. Collection of data, consisting of structural information of the molecules and an associated activity/property. Data may be collected from different sources but care should be taken that experimental conditions do not vary.
2. Data curation and standardisation (explained in 2.2)
3. The whole data may be split into two sets:
   - One set to build the model: the training set.
   - One set to externally test the model and validate it: the test set.

   Ideally, the test set should be unknown before prediction, but it is rarely the case.
4. Calculation of molecular descriptors.

5. Applying machine learning methods to obtain predictive models
6. Validation/Selection of the models with cross-validation (see section 2.5.3.1) and Y-randomisation (see section 2.5.4).
7. Definition of an Applicability Domain (AD) (see section 2.6)
8. Validation of the models on an external test set.

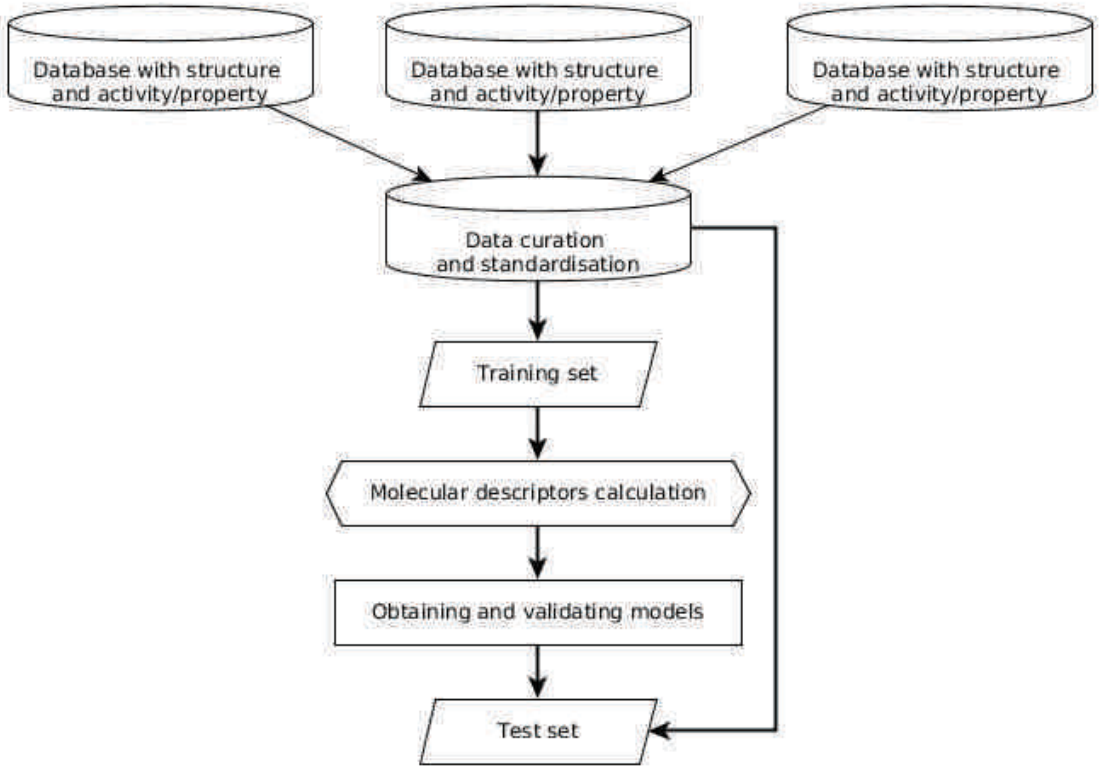


Figure 2.2: Typical workflow for the building of QSAR/QSPR models

According to the OECD principles[10], a QSAR model "should be associated with the following information:

1. a defined endpoint
2. an unambiguous algorithm
3. a defined domain of applicability
4. appropriate measures of goodness-of–fit, robustness and predictivity
5. a mechanistic interpretation, if possible"

These points constitute key steps in the realisation of a QSAR model. The first point concerns the activities and properties of the molecules and they have been introduced in the previous chapter and a more thorough explanation of each endpoint will be given when the related model is presented. Regarding points 2 to 4, these steps will be presented in this chapter(see sections 2.3, 2.4, 2.6 and 2.5). The last point is usually very difficult in

QSAR, and has led chemists to consider QSAR as a black box and despise it. However, using the ISIDA fragment descriptors which were developed it is possible to evaluate the contribution of each atom in the molecular graph (see 2.8).

Note that a QSAR model may be used as a virtual screening tool to find new bioactive molecules by predicting structures and choosing to test compounds predicted as active.

2.1.3 Chemogenomics-Based Virtual Screening

In drug discovery, SimVS is used for the identification of highly potent compounds against a particular target. Recently, the search for the latter has been shifted to explore selective and multi-target ligands using chemogenomics approaches. Chemogenomics, an interdisciplinary field, aims to explore all target families and their ligands. ChemoGenomics-Based Virtual Screening (CGBVS) is a QSAR-based approach to chemogenomics which describes both the target and the ligand. It has been proven to enable to find more active compounds than a classical SimVS[11,12].

2.2 Data curation and standardisation of molecular structures

Prior to the actual development of the QSAR model, it is crucial to curate the data. Verification of the structures is of utmost importance as the next step, the calculation of descriptors depends on it (except in the case of the use of experimentally evaluated properties). Small changes in the representation of a molecule, may result in significant differences in the accuracy of the final prediction[13]. The structures should be chemically correct and a standard representation should be used for groups with several possibilities such as nitro groups. As advocated by Fourches et al.[14], the structures must be standardised to have a unique representation for a same molecule. Ideally, different tautomeric forms of a same species should be recognized as such, and replaced by the dominant form involved in the action mechanism. However, it is very difficult to predict the dominant tautomeric form in solution, not to mention the potentially different one actually binding to a receptor. Tautomerization effects are an open and critical point in chemoinformatics research. Duplicates need to be detected and removed. These may directly influence the evaluation of the models. If possible, values should be verified. Database errors are very common and referring to the original work is always best.

2.3 Descriptors

During this thesis, the developed descriptors were used essentially but others were as well for comparison. In the case of the CGBVS study, a protein descriptor was used.

2.3.1 Molecular descriptors

2.3.1.1 ISIDA Substructural Molecular Fragments

ISIDA Substructural Molecular Fragments (SMF) descriptors[15,16] are based on the 2D molecular graph. Sequences, pairs or atom-centred fragment representing an atom and its neighbours are computed from the different structures and their total count in each structure corresponds to the descriptor value. Sequences and atom-centred fragments can be calculated at varying lengths.

2.3.1.2 Fuzzy Pharmacophore Triplets

Fuzzy Pharmacophore Triplet (FPT) descriptors[17,18] are pharmacophore-based descriptors using triplets as a fragmentation scheme. The atoms are coloured according to their pharmacophoric properties: hydrophobic, aromatic, hydrogen-bond acceptor/donor, positively and negatively charged ions. A fixed set of triplets is calculated. The fuzziness in FPT descriptors is generated from the counting by micro-species (see section 3.3.2) and by adding a contribution in the descriptor value of similar triplets to the one initially found.

2.3.1.3 MOE 2D descriptors

These descriptors are a collection of various descriptors based on the molecular graph calculated by the Molecular Operating Environment (MOE) 2011 program developed by the Chemical Group Computing. The 181 descriptors include physical properties, van der Waals surface area (the subdivided surface areas), the atom and bond counts (subdivided according to various criteria), the Kier and Hall chi connectivity and kappa molecular shape indices, distance and adjacency matrices (Balaban's connectivity topological index, Wiener path number, Wiener polarity number), pharmacophore atom types and partial charges-based descriptors. The whole list is available on the Chemical group Computing website (http://www.chemcomp.com/journal/descr.htm) under the 2D descriptors caption.

2.3.1.4 ChemAxon Pharmacophore Fingerprints

ChemAxon Pharmacophore Fingerprints (PF) are pharmacophore-based atom pairs. They are generated using the default options with the generatemd tool in JChem[19].

2.3.1.5 Chemically Advanced Template Search descriptors

The Chemically Advanced Template Search (CATS) descriptors[20–25] are pharmacophore-based fingerprints. The nodes in the molecular graph are coloured according to atom types: hydrogen-bond donor and acceptor , positively and negatively charged and lipophilic. The atom pairs are counted within a distance of 10 bonds. This results in a 150-dimension fingerprint (15 combinations of pairs $\times$ 10 different topological distances). Each count is then divided by the number of heavy atoms. These descriptors dependent on the topological distance are name CATS2D. A 3D counterpart using the Euclidean distance has been developed[22] and is named CATS3D. Another extension of the CATS descriptors is

the inclusion of colouring the aromatic atoms to differentiate them from the lipophilic atom type. The extended version was simply named CATS-2 and the original version CATS-1.

Note: The CATS2D-2 are very similar to the IPLF pharmacophore-coloured paired sequences of length 2 to 11 (nomenclature: IPh(2-11)_P, see chapter 3 for details), except for the scaling by the number of atoms.

2.3.1.6 Ligand-based Quantification of Interaction Distributions

LIgand-based QUantification of Interaction Distributions (LIQUID) descriptors[26,27] are fuzzy pharmacophore-based descriptors in the 3D space. Three different pharmacophore types are assigned: hydrophobic, hydrogen-bond acceptor and donor in the shape of an ellipsoid centred on the atom and modelled by Gaussian densities. The six corresponding pairs are searched for within 20 radii ranging from 1 to 20Å. In total, $6 \times 20 = 120$ descriptors are generated.

2.3.2 Protein descriptors

2.3.2.1 Local Alignment descriptors

Local Alignment (LA) descriptors[28] measure the similarity between two protein sequences by summing up scores obtained from aligning locally the sequences and allowing gaps. In CGBVS, these protein-protein dissimilarity values may be formally used to "embed" each protein as a point in a protein descriptor space. Its coordinates are determined such as to ensure that protein-protein distance calculations according to the metric of that latent space return LA dissimilarity scores.

2.3.3 Descriptor scaling

Descriptor spaces made of combinations of several types of descriptors should be scaled as they might have different numerical ranges. It is important[29] in order to have comparable variations of descriptors' values. Scaling can be done by assuming a Gaussian distribution of the descriptor values and centering them on zero with a unit standard deviation. It is also known as "normalisation". In order to transform each descriptor value D_i to its scaled value $D_i^{'}$ the following equation (2.1) is used:

$$D_i^{'} = \frac{D_i - \bar{D}}{\sigma} \tag{2.1}$$

where $\bar{D}$ is the mean value of the descriptor's values and σ is their standard deviation.

Another scaling, which does not assume a Gaussian distribution, would be to scale the value using the minima and maxima of each descriptor. It is known as "standardisation" and each descriptor value is transformed using the following equation (2.2):

$$D_i^{'} = \frac{D_i - D_{min}}{D_{max} - D_{min}} \tag{2.2}$$

where D_{min} is the minima of the descriptor's values and D_{max} is the maxima.

Note: the standardisation of descriptor values should not be confused with the standardisation of the molecular graph.

2.3.4 Similarity/Dissimilarity metrics

The similarity metric permits to evaluate the similarity between two molecule in a given descriptor space. Indeed, the change of descriptor space will represent the molecules differently and the distance between the two molecules may change[5]. The distance between the molecules can be viewed as the dissimilarity of the molecules.

Three different types have been used in this work: the Euclidean distance (EUCLID)[29], the Dice coefficient-based distance (DICE)[30] and the Fraction of Differences (FDIFF)[27].

The different distances between two molecules in a descriptor space of n-dimensions represented by their respective vectors m and M are calculated as:

$$EUCLID = \sqrt[2]{\sum_{i=1}^{n}(D_i(M) - D_i(m))^2}$$

$$DICE = 1 - \frac{2\sum_{i=1}^{n} D_i(M)D_i(m)}{\sum_{i=1}^{n} D_i(M)^2 \sum_{i=1}^{n} D_i(m)^2}$$

The FDIFF is a count of the fraction of features that are differently populated in a pair of molecules. The FDIFF count is incremented by 1 for each descriptor element absent in one molecule but present in the other. The total count of elements with such a condition is then divided by the dimensionality n of the descriptor space. Thus, FDIFF varies between 0 and 1.

2.3.5 Neighbourhood behaviour criteria

The Neighbourhood Behaviour (NB) criteria[7,31,32] are a quantitative expression of the classical similarity principle: "similar molecules have similar properties"[33]. The NB is associated to a descriptor space and a dissimilarity metric.

The NB optimality criterion $\Omega(d)$ by Horvath and Jeandenans[31] is defined as follows at a dissimilarity threshold d:

$$\Omega(d) = \frac{kN_{FS}(d) + N_{PFD}(d)}{kN_{FS}^{(null)} + N_{PFD}^{(null)}}$$

where

- $N_{FS}(d)$ is the count of false similar at the dissimilarity threshold d for a molecule pair (M,m). $N_{FS}(d)$ correspond to pairs of compounds that are structurally less dissimilar than d, but are in violation of the similarity principle. The violation is defined when the measured properties on both molecules is high. Their difference $|Y(m) - Y(M)|$ exceeds an experimentally admissible threshold.

- $N_{PFD}(d)$ is the count of potentially false dissimilar at the dissimilarity threshold d.FTN
- k is a weight higher than 1 to give more importance to keeping the number of false similar (N_{FS}) as low as possible rather than keeping the number of potentially false dissimilar low (N_{PFD}). Molecules reflected as potentially false dissimilar may actually be dissimilar to the query compound but nevertheless actives. (k=5 in this work)
- $kN_{FS}^{(null)}$ and $N_{PFD}^{(null)}$ correspond to the number of false similar and potentially false dissimilar in a random distribution.

Given a descriptor space and an associated metric, significant NB (translating into $\Omega << 1$) occurs if in the list of pairs (m,M) ranked by their calculated dissimilarity score, only pairs with minimal property differences $|Y(m) - Y(M)|$ are being ranked at the top of the list. Then, the optimal dissimilarity radius d*, minimizing Ω, is chosen such as to encompass the entire "head" of the ranked list, with a minimal number of FS. Note, however, that in small and biased data sets, sometimes a random ordering of compound pairs may, by pure chance, also place only property-related pairs at the top of the list. Therefore, it is a good practice to check how likely it is to generate such an apparently NB-compliant ranking by pure chance: if this is easy, than low Ω values obtained with the actual metric may not be interpreted as a genuine NB compliance either. In order to take this fortuitous fluctuation effect into account, the NB optimality criterion was extended to the Local Ascertained Optimality Score (LAOS)[27] which takes into account effects of random models by assessing 20 data scrambling simulations (see section 2.5.4):

$$LAOS(d) = \Omega^{\bar{rand}}(d) - \sigma(\Omega^{rand}(d)) - \Omega(d)$$

where

- $\Omega^{\bar{rand}}(d)$ is the mean Ω value over the 20 scrambling simulations
- $\sigma(\Omega^{rand}(d))$ is the standard deviation over the 20 scrambling simulations

2.4 Machine Learning

Machine learning[34,35] is a field of computer science trying to determine the maximum likelihood functional form of the dependence of an explained variable Y with respect to relevant parameters x_i, $Y = f(x_i)$ given a set of observed instances j (Y^j, x_i^j). Explanatory variables in chemoinformatics are descriptors. In principle, machine learning is an open-ended search over the entire set of possible mathematical functions $f(x_i)$ - in practice, however, this may be restricted to preselected functional forms. Various statistical criteria may be used to check whether a function $f(x_i)$ is a good explanation for the observed instances (Y^j, x_i^j). The functions $f^*(x_i)$ maximizing the selected relevance criterion may be considered as the so-far best hypothesis for the mathematical model linking Y to its parameters. In this work, several algorithms have been used, including multi-linear regression, partial least square regression, support vector machines and artificial neural networks.

2.4.1 Multi-Linear Regression

The Multi-Linear Regression (MLR) is a generalisation of the simple linear regression with p explanatory variables. MLR assumes that the modelled property y depends linearly of the different descriptors $(D_1, D_2, ...D_i, ..., D_n)$ according to the following equation:

$$y = a_0 + \sum_{i=1}^{n} a_i D_i$$

The most common algorithm of MLR tries to optimise the coefficients a_i in order to reduce the sum of the squares of the errors as performance criterion.

2.4.1.1 Stepwise selection of variables

The stepwise selection[36], as its name indicates, picks out relevant descriptors in a stepwise manner, instead of doing the MLR on the whole set of available descriptors. The forward and backward approaches have been used in this work. First, all the regression models are systematically assessed in terms of combinations involving a limited number of descriptors. For example, all the bilinear equations $Y = a_i D_i + a_j D_j + a_0$, where $i = 1...N-1$ and $j = i+1...N$ are assessed. The one of best statistical robustness is kept. Then, the current equation is iteratively grown, by scanning over not yet used descriptors and entering the one that causes a maximal and significant growth of model quality criteria (Student's t-test), until no more such descriptors are found. Then, the model may be pruned by backward stepping: searching for the used variable that may be removed without causing a significant quality drop. Note that the relative importance of an explaining variable at the moment of its co-opting into the model may have changed as a consequence of the further growth of the equation.

2.4.2 Partial Least Square Regression

Partial least squares or Projection to Latent Structures (PLS) regression was introduced in the 60s by the statistician Herman Wold[37]. Is has since been widely used in many areas and described in many articles and books[38–40].

The PLS approach finds a linear regression model by projecting the problem to a new space; it is similar to principal component regression. A PLS model tries to find a combination of the descriptors (D_i) which explains to the maximum the variation in the property space (Y). It is based on the assumption that a link exists between latent variables (c) based on the observable variables (the descriptors D_i) and the predicted variables (the property to predict Y). The PLS permits a dimension reduction by expressing the descriptor space of dimension n in a number C of components (c):

$$c_j = \sum_{i=1}^{n} a_i D_i$$

where D_i are the descriptors found explaining the maximum variance in the property y The property y is then linearly related to those components:

$$y = \sum_{j=1}^{C} b_j c_j$$

where b_j are chosen coefficients in order to reduce the performance criterion, usually the sum of the squares error.

The advantage of PLS is that by analysing the descriptor space, it summarizes redundant information, especially for highly correlated descriptor vectors, and reduces the space dimension. ISIDA fragment descriptors have often very large dimensions for descriptor spaces and by their nature, descriptors are highly intercorrelated; therefore, PLS is a method of choice for these descriptors.

2.4.3 Support Vector Machines

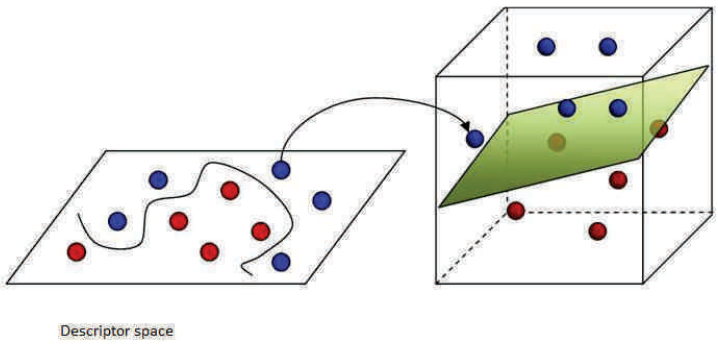


Figure 2.3: Change from the feature space into high-dimensional space

Support Vector Machine (SVM)[41–43] was introduced by Vapnik in the 1990s[44,45]. SVM, in order to handle non-linearly separable problems as presented in Figure 2.3, modifies the features' space to represent them into a higher dimensional space and then apply a learning algorithm.

Note that a transformation like the one depicted in Figure 2.3 is invariant to a roto-translation of the considered reference system and therefore does not require the knowledge of the absolute coordinate values of the instances, but only their relative positions with respect to each other. Therefore, the input of the matrix of distances between the items is sufficient. The Kernel matrix $K(m, M)$ corresponds to the computed dissimilarity score between molecules M and m. The actual coordinates of M and m, e.g. their descriptors may always be "back-engineered" by embedding if $K(m, M)$ is given. However, in the chemogenomics problem, the dissimilarity between the different pairs of compound-protein interactions (CPIs), $K(P : C, p : c)$, needs to be computed. In this context, kernel-driven modelling may be much more advantageous than the descriptor-based one: while it may be unclear how to define the descriptor vector of a putative protein-ligand complex P:C, one may fall back to a straightforward definition of the dissimilarity metric between two complexes P:C and p:c as $K(P : C, p : c) = K^p(p, P) \times K^c(C, c)$, where K^p and k^c may be any of the commonly employed methods to estimate protein-protein and small-molecule dissimilarity. This approach is known as the "Kernel trick".

In the case of a classification problem, the learning algorithm seeks to find a hyperplane capable of distinguishing between the classes. The decision hyperplane in SVM is built

by maximising the distances, called margins, between the instances and the plane. It is described by coefficients attributed to each instance of the training set. These latter are named support vectors. Classification errors can be tolerated in order to obtain simpler planes. It is controlled by the cost parameter which allows errors within a certain distance. The further away from the decision hyperplane the error is located, the higher is the cost. This type of SVM is known as C-SVM.

In the case of a regression problem, a regression function is optimised instead of finding a hyperplane. It is optimised in order to obtain an error below a threshold ϵ. This type of SVM is known as ϵ-SVM.

SVM is an interesting machine learning method because it does not minimise a quadratic equation. One of its advantages is that regularisation of overfitting does not depend on the number of descriptors[46]. This is important in the case of fragment descriptors such as the ISIDA descriptors, because of the large number of fragments generated. However, as any method it has its disadvantages and SVM cannot estimate its own accuracy as it isn't probabilistic.

2.4.4 Artificial Neural Networks

Artificial neural networks[47,48] are inspired from the nervous systems of animals. Interconnected neurons make up a system to treat information given to the network. In machine learning, the network is usually constructed with three layers of "neurons": one to input the information, i.e. the descriptors, a intermediate hidden layer and an output layer for the property or properties to predict. Figure 2.4 illustrates such a network. Neurons in adjacent layers are interconnected and these connections are associated to an adaptive weight w_i. The neurons on the hidden layer, sum up the weighted signals from the input layer. Each neuron has an activation function φ, output=φ(input), controlled by various adjustable parameters. Output intensity is usually normed between 0 and 1, corresponding to "inactive" vs. "firing" neurons in biology. For example, if φ is a sigmoid, the intensity of input signal corresponding to the inflexion point, and the steepness of the transition are fittable. A priori, the choice of the functional form per se (linear, sigmoid, Gaussian, etc.) may be fittable, but this requires specific network optimization heuristics, able to handle both categorical (function form classes) and continuous variables (weights, functional parameters). By default, fitting is restricted to the latter, given a predefined network geometry, with predefined activation functions and can be handled by a gradient-based minimisation of the calculated/experimental RMSE of the property, which is a complex differentiable function of the cited parameters.

2.4.5 Consensus modelling

A consensus model (CM) consists of assembling different models obtained by different building strategies by either using their predictions arithmetic mean as prediction to regression problems or a majority vote as prediction in the case of classification problems. Models may differ in strategy by the data, the descriptors, the machine learning method or its parameters selected. It has been shown that more robust models with a higher predictability result from building a CM[16,49,50].

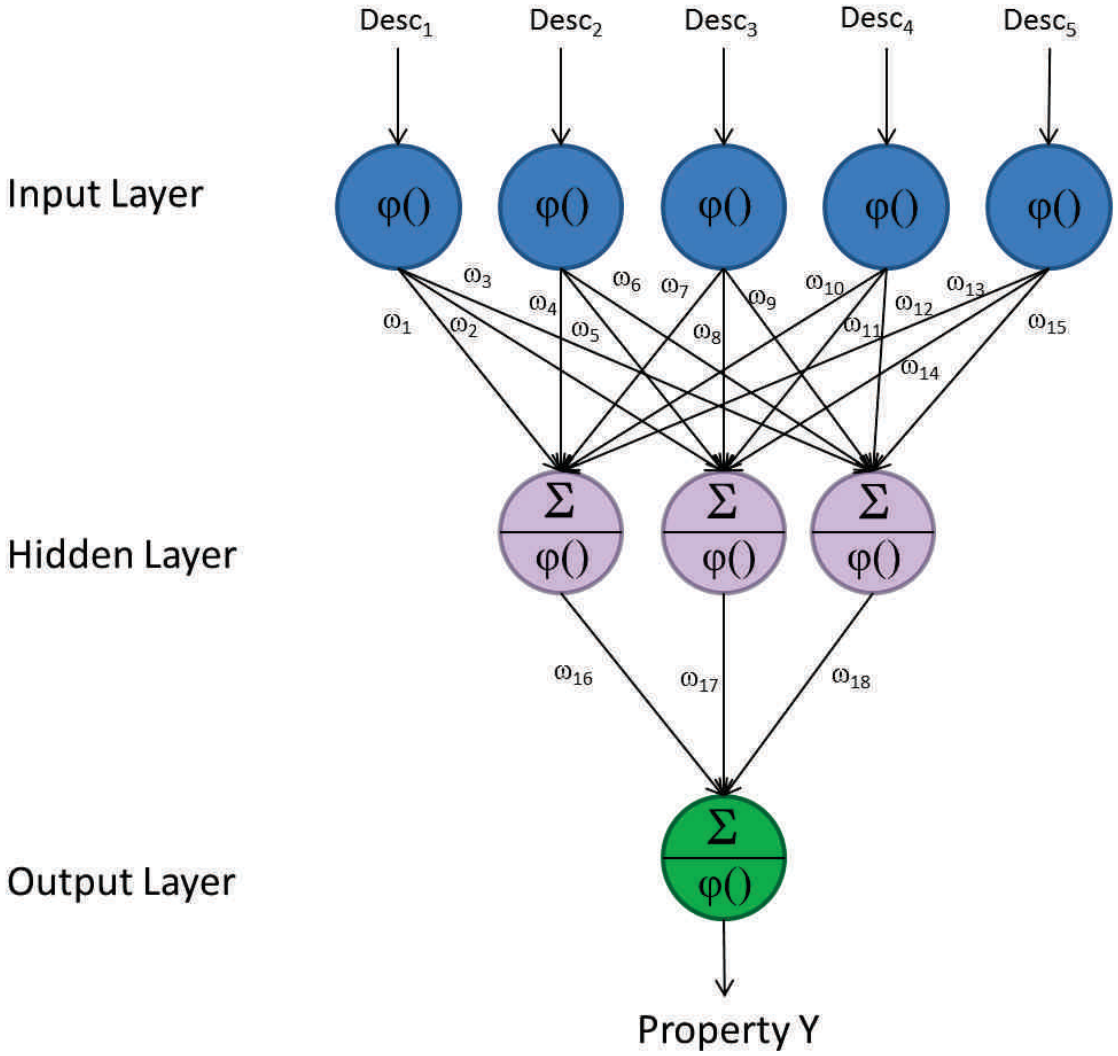


Figure 2.4: Example of an artificial neural network

2.4.5.1 Bagging

To make models differ by their training set, an algorithm can be applied that randomly generates different subsets from the initial data set. Models are then build on these different subsets and combined into a CM. This process is known as bagging or bootstrap aggregation[35].

2.5 Evaluation of models performance and validation

How well a model performs is evaluated by different criteria comparing the predicted value of the model to the actual experimental value. A model should be validated externally, i.e. with data not included in the training set[51]. This section will introduce the different statistical criteria used in this work and explain different validation methods such as cross-validation and y-randomisation.

Table 2.1: Confusion Matrix for a binary class problem

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

2.5.1 Regression models performance criteria

Pearson's correlation coefficient:

$$R^2_{corr} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(f_i - \bar{f})}{\sqrt[2]{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(f_i - \bar{f})^2}}$$

Determination coefficient:

$$R^2_{det} = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Root Mean Squared Error:

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}}$$

Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - f_i|}{n}$$

where y_i is the experimental value of compound i,
$\bar{y}$ is the mean value of the experimental values,
f_i is the predicted value of compound i, $\bar{f}$ is the mean value of the predicted values and n is the number of compounds.

2.5.2 Classification models performance criteria

In this section, only criteria in the case of binary classification problems, i.e. only two classes are defined, will be set out as the only classification models produced during this thesis were to differentiate binders from non-binders in ligand-protein interactions. Binders are usually referred to as active molecules as they may induce a biological response. This class of compound is defined as the positive class, while non-binders are referred to as inactive and belong to the negative class.

2.5.2.1 Confusion Matrix, balanced Accuracy and Matthew's correlation coefficient[52]

The comparison of the classifier's predictions and the actual classes of the compounds can be summarised in a matrix called the confusion matrix (see Table 2.1) where:

- instances of the Positive class **correctly** predicted are named True Positives (TP),
- instances of the Negative class **correctly** predicted are named True Negatives (TN),

- instances of the Positive class **wrongly** predicted are named False Negatives (FN),
- instances of the Negative class **wrongly** predicted are named False Positives (FP).

From the confusion matrix, average evaluators of the model's performance can be calculated:

True positive rate, aka. Sensitivity or Recall(Positive):

$$TPrate = \frac{TP}{TP + FN}$$

False positive rate :

$$FPrate = \frac{FP}{TN + FP}$$

Specificity or Recall(Negative):

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Balanced Accuracy (BA):

$$BA = \frac{Recall(Positive) + Recall(Negative)}{2} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$$

The accuracy corresponds to the classifier's version of the mean absolute error. However, in the case of an unbalanced set, it does not give a good estimation the prediction performances in both classes but rather on the most dominant one. The Balanced Accuracy (BA) will compensate for the unbalance in the set. Both vary between 0 and 1. When the data is perfectly predicted, they are equal to one and if every instance is wrongly predicted, then they are equal to 0.

Matthew's correlation coefficient:

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt[2]{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

The MCC corresponds to the classifier's version of Pearson's correlation coefficient. It varies between -1 and 1. When a perfect prediction of the data is achieved, the MCC is equal to 1. It is equal to 0 when half the predictions are wrong and is equal to -1 when all positives are predicted as negatives and vice versa.

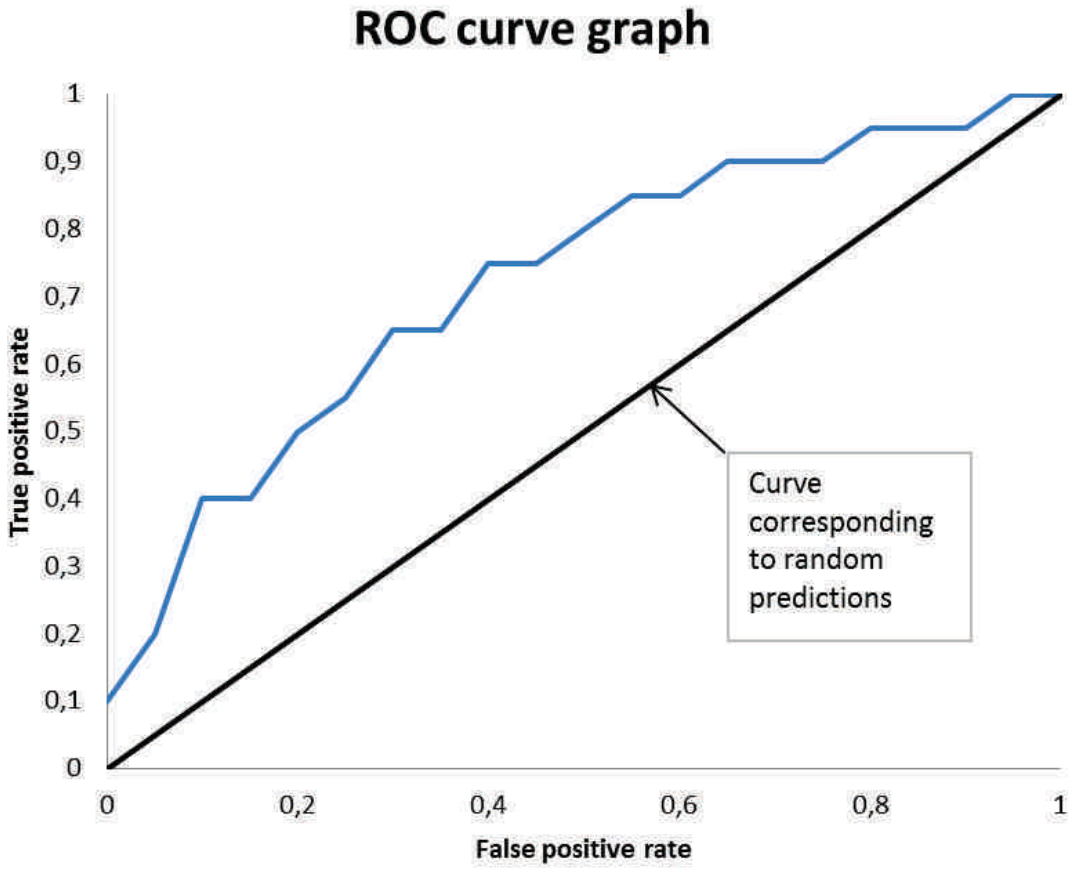


Figure 2.5: Example of a ROC curve

2.5.2.2 Receiver Operating Characteristics graph

Receiver Operating Characteristics (ROC) graphs[53] are two-dimensional graphs where the TP rate (y-axis) is plotted against the FP rate (x-axis). It represents the trade-off between benefits (in the form of TP rate) and avoidable costs (in the form of FN rate), i.e. entities that will be synthesized and tested for nothing, because they are not active as predicted. An example of such plot is given in Figure 2.5. Some classifiers produce a continuous output such as the probability of an instance to correspond to a certain class. In this case, a threshold is associated with the prediction of the classifier above which compounds are considered as active (positive class) and below which they are considered as inactive (negative class). By varying the threshold, it is possible to obtain different TP and FN rates and to produce a curve on the ROC graph, which is referred to as the ROC curve. In the case of SVM, the distance from the decision hyperplane is used to sort the instances[54]. The further away, the better the prediction is considered.

The area under the ROC curve (AUC) is considered as a measure of performance of the classifier. Ideally, AUC should be equal to 1. It would prioritize all true actives ahead of all inactive molecules. If the curve corresponds to the diagonal, which corresponds to random predictions, the AUC will be equal to 0.5. It is associated to the performance of a random model. Therefore, only classifiers located in the upper triangle are considered

to perform better than random predictions.

2.5.3 External validation and cross-validation strategy

A model should always be validated on external data, that is the above mentioned statistical parameters should not only be calculated on data used on the model but additional instances, kept aside to verify whether the model may interpolate or extrapolate new correct predictions from the examples used at the training stage[51]. Ideally, the data should be entirely external to the modelling. However, the cross-validation strategy permits to have an evaluation of a model's performance without external data.

2.5.3.1 Cross-validation

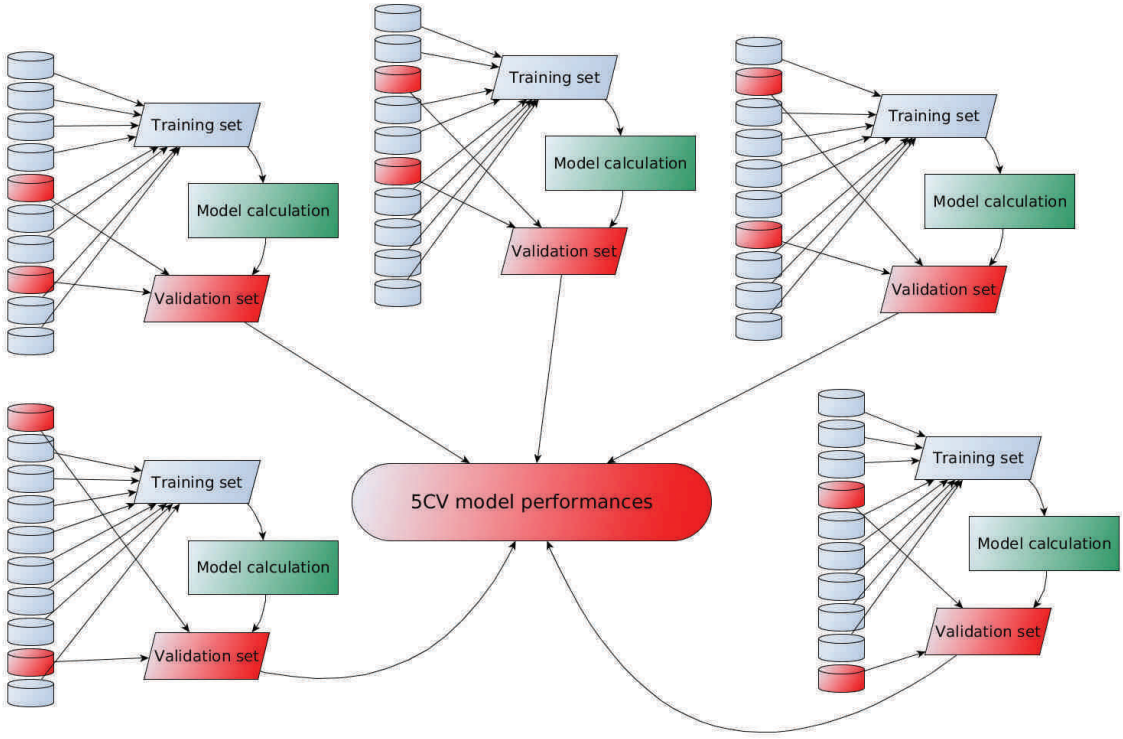


Figure 2.6: 5-fold cross-validation (5CV) procedure

The cross-validation (CV) procedure[55] consists in dividing randomly the initial data set into n folds of approximately the same size. Then, a training set consisting of n-1 folds is used to train a model and the remaining fold, the validation set, is predicted by this model. This step is repeated until all the folds have been predicted. Statistical parameters can then be calculated on the whole set as each fold has been predicted externally.To insure the statistics are well evaluated and not a lucky draw of the training and validation set, the CV should be performed several times and a mean of the different statistical parameters is used. It is recommended[55] to use the lowest number of folds possible, i.e. 2. It is very common in chemoinformatics to use 5-fold CV and it is has been used for comparison purposes with collaborators in this thesis.

The concept of 5-fold CV is illustrated in Figure 2.6. The initial set of data contains 10 instances, which are then randomly split each time into training set (8 instances) and validation set (2 instances). The model is calculated on the training set and is used to predict the validation set, which then permits the calculation of statistical parameters to evaluate the model's performance. CV permits to detect over-fitting models, which are caracterised by a very good training set prediction and a poor performance on external test sets. Over-fitting is usually regarded as a model which has learned the training set perfectly instead of extracting general rules about the property.

2.5.4 Y-randomisation or scrambling

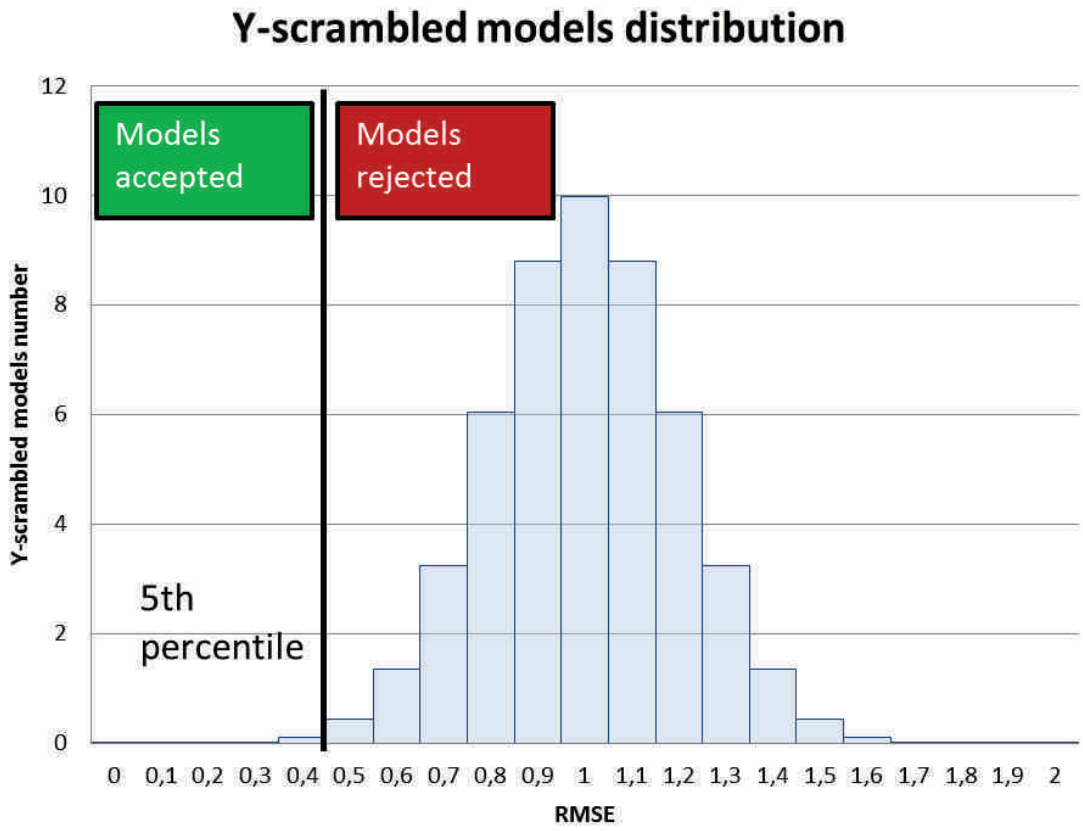


Figure 2.7: Scrambling models distribution and boundary for acceptance of models

Fortuitous models may be build based on selecting descriptors which may have nothing to do with the problem at hand but fit the data. This happens especially when the number of descriptors is high. These models may have good CV performances and seem to not overfit, however, their external predictions will be bad as it is not based on any proper learning of the problem. The purpose of y-randomisation[56], also named scrambling, is to maximise the odds that the model is not a serendipitous event. The instances' values or class labels are shuffled randomly and CV is run on the permuted data. The procedure is done several times and a statistical performance criteria calculated on the CV is selected to evaluate the performances of the y-randomised models. A confidence interval at a certain level such

as 95% can be calculated on the selected criteria by assuming a normal distribution of the y-randomised models. In order to ensure that a non-randomised model is not fortuitous, it needs to perform better than the upper bound of the calculated confidence interval in terms of the selected criteria. For example, as shown in Figure 2.7 if the selected criteria is RMSE, then since the best models have a lower RMSE than the others, the 5th percentile instead of the 95th should be taken.

The idea is that if many columns of random numbers are added to the descriptor matrix, some of these will by chance mimic the Y column. If this chance is high, scrambling Y will continue to produce good models, because another random column will mimic the scrambled Y vector. By verifying how well scrambled models perform, those fortuitous models can be removed. However, if a descriptor column is serendipitously correlated to Y due to a bias in the compound collection, then Y-scrambling is of no help. For example, using fragment descriptors, all the actives, and only actives contain a fragment F in the training set. Many inactive molecules with this fragment F exist, but they are not known, or not added to the training set. Reversely, there are actives without the fragment F. However, until the set is expanded, the rule that molecules containing F are actives cannot be detected as fake, no matter how much the data is scrambled.

2.5.5 Outliers

When validating a model, it is common to find molecules with great prediction errors having a tendency to be far-off the distribution of the others' distribution when plotting the predicted values against the experimental values. These are often designated as outliers[8,57] in chemoinformatics. Statistical definition and tests to identify outliers exist but it is common, particularly in chemoinformatics, to identify them simply by their discrepancy in comparison to other molecules[58]. Several causes for outliers may be considered:

- Experimental problems
  - The experimental value is wrong. Errors of "copying" often occur and if possible the source should be checked.
  - Different experimental procedures or conditions have been used and may not be compatible.
  - A disturbing element occurred during the experiment and the molecule may have been modified or an impurity was measured for example.
  - ...
- Modelling problems
  - Too few instances of this kind are seen by the machine learning, hence the model does not "understand" those molecules.
  - The representation used (=standardisation and descriptor space) did not catch the essential difference between the outlier and the well-predicted. In this case, the descriptor space may not be appropriate.
  - The molecular graph is not the one corresponding to the value.

– ...

Outliers may also be due to natural deviations in populations... and many other reasons. It is often very difficult to determine the cause of an outlier but they should always be checked.

2.6 Applicability Domain

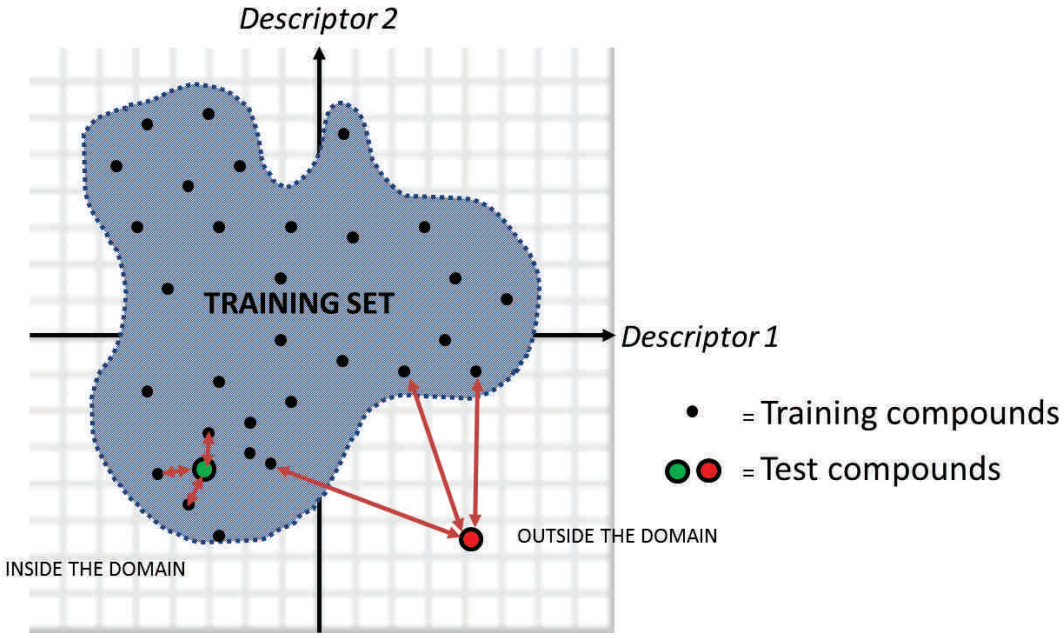


Figure 2.8: Representation of the Applicability Domain boundary in chemical space

A model is based on a number of experimental observations from particular molecules. It is expected that the model should be able to predict molecules similar to those observed but not dissimilar ones as it cannot be guaranteed that the proper relation have been learned for those. The Applicability Domain (AD) defines this boundary and expresses the scope and limitations of a model, i.e. molecules for which the prediction is expected to be reliable. Netzeva at al. define it as: "the response and chemical structure space in which the model makes predictions with a given reliability" [59]. The concept of AD is illustrated in Figure 2.8 where the chemical space is represented by Descriptor1 and Descriptor2. The test set compound 1 (in green) is inside the AD boundary and its prediction is considered reliable while the test set compound 2 (in red) is outside and therefore, its prediction is considered as unreliable.

AD is a very important issue, if it is not evaluated correctly it may lead to wrong conclusions [8,60]. Although the problem of AD is being explored intensively [59,61,62], the problem is far from being resolved [63].

During this work, AD of individual models has been evaluated using two definitions: Bounding box and Fragment control.

The Fragment Control [64] approach can only be used on fragment-based descriptors which are determined from the structures of the molecules in the training set and not predefined. A test molecule is considered outside the AD if a new fragment occurs in its structure, i.e. unseen in the training set.

The Bounding Box [62],aka. Min-max method, uses the minimum and maximum values of each descriptor encountered in the training set. A molecule is considered outside of the AD of a model if one of its descriptor values is outside of the minimum-maximum range for this descriptor. It is an indication that the molecule is chemically different than those encountered. Bounding box includes the fragment control AD as well.

2.6.1 Applicability of Consensus Models

In the case of consensus models, the variance of predictions of the different models constitutes a criteria of trustworthiness of the overall prediction[65,66], which has been used in this work.

2.7 Machine learning Software

2.7.1 Stochastic QSAR Sampler

The Stochastic QSAR Sampler (SQS)[17] generates a consensus model of MLR models using a genetic algorithm[67,68]. The genetic algorithm is inspired by evolutionary theory by Darwin. Different populations of individuals are generated from user-defined descriptors. The individuals are made of chromosomes which correspond to a descriptor, thus, each individual is made of a subset of descriptors. During the initialisation phase, the initial pool of descriptors provided by the user undergoes certain non-linear transformations and in the end, a final set of 5000 descriptors are selected by eliminating highly correlated vectors. Populations of individuals are built by selecting descriptors randomly from these 5000 descriptors. Individuals represent possible descriptor selection/transformation schemes in the employed descriptor space. This information, the "chromosome", lists the specific descriptor terms that will be used in the encoded model (as such or after submission to one of the optional non-linear transformations). The chromosome unambiguously defines a model, in which selected/transformed descriptors are assigned coefficients by a linear regression procedure. The better the performance of this model in 3-CV, and the less variables it contains, the "fitter" the chromosome is considered. Its probability to generate offspring by cross-overs with other fit chromosomes (i.e. combine successful descriptor selections) thus increases. This translates into accumulation of well-cross validating models in the population, over time.

2.7.2 ISIDA/QSPR

The ISIDA/QSPR program[69] developed by Vitaly Solo'vev uses MLR analysis with combined forward and backward stepwise variable selection techniques[16,70,71]. It uses different types of ISIDA fragment descriptors and generates multiple MLR models on each

type of descriptor. The MLR models are selected according to the leave-one-out cross-validated correlation coefficient. The leave-one-out cross-validation is a special case of cross-validation where the validation set contains only one compound. It thus corresponds for set of N compounds to a N-fold cross-validation. Models are selected at a user-defined threshold of the leave-one-out cross-validated correlation coefficient and assembled in a consensus model. The program automatically applies an applicability domain using bounding box and fragment control when doing an overall cross-validation or predicting external values.

2.7.3 ASNN

The ASsociative Neural Network (ASNN) program[72–74] developed by Igor Tetko, builds a consensus model on 100 neural networks. For each neural network, the set is divided into a training and an internal validation set in order to assess the performance of the network externally and to avoid over-fitting. At the prediction stage, the consensus prediction $y_{\bar{M}}$ of molecule M is corrected using the values of the nearest neighbouring molecules in the model. The N nearest molecules are determined by the correlation coefficient between the descriptor vectors of the molecule to predict (molecule M) and the ones in the training set (molecules m). The corrected prediction $y_{\bar{M}}{}'$ correction is calculated with the following equation:

$$y_{\bar{M}}{}' = y_{\bar{M}} + \frac{1}{N}\sum_{m=1}^{N} y_{exp,m} - y_{\bar{exp}}$$

where $y_{exp,m}$ is the experimental value of molecule m and $y_{\bar{exp}}$ is the mean of the experimental values of the N nearest neighbour molecules.

2.7.4 Weka

Weka[35 75] is a data mining program in Java upheld by the Machine Learning Group at the University of Waikato. It contains several machine learning algorithms and their implemented PLS algorithm has been used in this work.

2.7.5 LibSVM

The LibSVM package[76] developed by Chih-Chung Chang and Chih-Jen Lin, has been used in this work to build SVM models. Their code being open-source, J.B. Brown (Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University) made a link to their method inside his tools to generate SVM models from in-house prepared CPIs similarity matrices. These tools were used in the chemogenomics project on GPCRs (section 4.4).

2.8 Interpretability of QSAR models using fragment descriptors

Interpretation of QSAR model is sought by chemists in order to give mechanistic explanation of the observed phenomenon. However, achieving this, simultaneously with

prediction efficiency is rare. Prediction efficiency is often achieved using complicated algorithms such as ANN, SVM or a consensus model based on different methods which cannot be readily interpreted. Though SVM with a linear kernel could be quite easily reinterpreted to attribute the coefficient to a descriptor like in MLR. Interpretability of QSAR models is often achieved by using a few physico-chemically meaningful descriptors and MLR to observe which coefficients are significant. Another approach is to use atomic or fragment increments which sum up over the whole molecule to the predicted property, as was done, for example, by Ghose and Crippen for the prediction of LogP[77,78]. The approach developed by Marcou et al.[79] enables the interpretation of fragment descriptors into atomic increments by analysing partial derivatives of the predicted value.

In the case of a linear model, the predicted property (y) is calculated as follows:

$$y = \sum_{i}^{n} a_i D_i$$

where D_i is the value of fragment descriptor i and a_i is its contribution. The atomic contributions ac can be calculated from the fragment contribution by summing up all the fragment contributions in which the atom appears in the corresponding fragment:

$$ac_A = \sum_{j}^{p} a_j$$

where j are the p fragments in which the atom A is found.

In the case of a non-linear regression model, the idea by Marcou et al. is to make each descriptor value vary by δ ($\delta = 1$ when fragment counts are used) in turn and analyse the effect on the predicted property by the function f to obtain the fragment contribution:

$$a_i = \frac{\partial f(\{D_1, D_2, ..., D_i, ..., D_n\})}{\partial D_i} = \frac{f(\{D_1, D_2, ..., \mathbf{D_i}, ..., D_n\}) - f(\{D_1, D_2, ..., \mathbf{D_i} - \delta, ..., D_n\})}{\delta}$$

The approach has been implemented into the ISIDA tools by Gilles Marcou as ISIDA ColorAtom and works only with models build on ISIDA fragment descriptors. Furthermore, the ISIDA ColorAtom program permits the user to colour the molecule by the increments for visualisation, as shown in Figure 2.9.

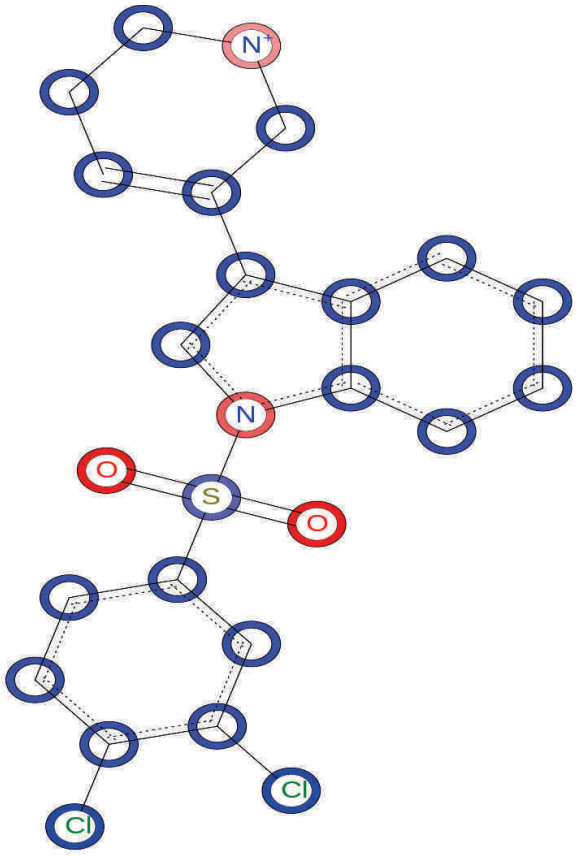


Figure 2.9: Example of a "coloured" molecule using the ISIDA ColorAtom program. This is an example from a ligand of the serotonin 6 (5-HT6) receptor used in our GPCR study of these increments (see 6.2)

Bibliography

[1] Martin, Y. C., Kofron, J., and Traphagen, L. M. *J Med Chem* **45**(19), 4350–4358 Sep (2002).

[2] Bender, A. and Glen, R. C. *Org Biomol Chem* **2**(22), 3204–3218 (2004).

[3] Maldonado, A. G., Douvet, J., Petitjean, M., and Fan, B. *Mol Divers* **10**(1), 39–79 Feb (2006).

[4] Eckert, H. and Bajorath, J. *Drug Discovery Today* **12**(5-6), 225–233 (2007).

[5] Willett, P. *Annu Rev Inform Sci* **43**, 3–71 (2009).

[6] Geppert, H., Vogt, M., and Bajorath, J. *J Chem Inf Model* **50**(2), 205–2016 (2010).

[7] Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. *J Med Chem* **39**(16), 3049–3059 (1996).

[8] Tropsha, A. *Mol Inf* **29**(6-7), 476–488 (2010).

[9] Hansch, C. and Fujita, T. *J Am Chem Soc* **86**(8), 1616–1626 Apr (1964).

[10] *OECD Principles for the Validation for the Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. URL: http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf*, (2004).

[11] Brown, J., Niijima, S., Shiraishi, A., Nakatsui, M., and Okuno, Y. In *IEEE International Conference on Bioinformatics and Biomedicine Workshops,* Gao, J., Dubitzky, W., Wu, C., Liebman, M., Alhaij, R., Ungar, L., Christianson, A., and Hu, X., editors (, Philadelphia, PA, 2012).

[12] Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., and Okuno, Y. *Mol Syst Biol* **7**, 472–484 (2011).

[13] Young, D., Martin, T., Venkatapathy, R., and Harten, P. *QSAR Comb Sci* **27**(11-12), 1337–1345 (2008).

[14] Fourches, D., Muratov, E., and Tropsha, A. *J Chem Inf Comput Sci* **50**(7), 1189–1204 JUN (2010).

[15] Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. *J Comput Aid Mol Des* **19**(9-10), 693–703 Jul (2005).

[16] Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I., and Marcou, G. *Curr Comput Aided Drug Des* **4**(3), 191–198 Sept (2008).

[17] Bonachera, F. and Horvath, D. *J Chem Inf Model* **48**(2), 409–425 Feb (2008).

[18] Bonachera, F., Parent, B., Barbosa, F., Froloff, N., and Horvath, D. *J Chem Inf Model* **46**(6), 2457–2477 Nov-Dec (2006).

[19] *ChemAxon Pharmacophore Fingerprint, http://www.chemaxon.com/jchem/doc/user/PFpChemAxon*, (2014).

[20] Schneider, G., Neidhart, W., Giller, T., and Schmid, G. *Angew Chem Int Ed Engl* **38**(19), 2894–2896 Oct (1999).

[21] Schneider, G., Lee, M.-L., Stahl, M., and Schneider, P. *J Comput Aid Mol Des* **14**, 487–494 (2000).

[22] Fechner, U., Franke, L., Renner, S., Schneider, P., and Schneider, G. *J Comput Aid Mol Des* **17**(10), 687–698 Oct (2003).

[23] Renner, S. and Schneider, G. *Chem Med Chem* **1**(2), 181–185 (2006).

[24] Schneider, G., Schneider, P., and Renner, S. *QSAR Comb Sci* **25**(12), 1162–1171 (2006).

[25] Reutlinger, M., Koch, C. P., Reker, D., Todoroff, N., Schneider, P., Rodrigues, T., and Schneider, G. *Mol Inf* **32**(2), 133–138 (2013).

[26] Tanrikulu, Y., Nietert, M., Scheffer, U., Proschak, E., Kristina, G., Schneider, P., Weidlich, M., Karas, M., Michael, G., and Schneider, G. *Chembiochem* **8**(16), 1932–1936 Nov (2007).

[27] Horvath, D., Koch, C.and Schneider, G., Marcou, G., and A., V. *J Comput Aided Mol Des* **25**, 237–252 Jan (2011).

[28] Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. *Bioinformatics* **20**(11), 1682–1689 JUL (2004).

[29] Leach, A. R. and Gillet, V. J. *An Introduction to Chemoinformatics*. Springer Verlag, (2007).

[30] Dice, L. R. *Ecology* **26**(3), 297–302 (1945).

[31] Horvath, D. and Jeandenans, C. *J Chem Inf Comput Sci* **43**(2), 680–690 Mar-Apr (2003).

[32] Horvath, D. and Jeandenans, C. *J Chem Inf Comput Sci* **43**(2), 691–698 Mar-Apr (2003).

[33] Johnson, A. M. and Maggiora, G. M. *Concepts and Applications of Molecular Similarity*. John Willey & Sons: New York, (1990).

[34] Alpaydın, E. *Introduction to Machine Learning*. The MIT Press Cambridge, Massachusetts London, England, second edition, (2010).

[35] Witten, I., Franck, E., and M.A., H. *Data Mining - Practical Machine Learning Tools and Techniques*. Elsevier, 3rd edition, (2011).

[36] Hocking, R. R. *Biometrics* **32**(1), 1–49 (1976).

[37] Wold, H. *"Estimation of principal components and related models by iterative least squares" in Multivariate Analysis, p. 391-420*. New York: Academic Press, (1966).

[38] Helland, I. S. *Scand J Statist* **17**, 97–114 (1990).

[39] Lingjaerde, O. C. and Christophersen, N. *Scand J Statist* **27**(3), 459–473 (2000).

[40] Tenenhaus, M., Vinzia, V. E., Chatelinc, Y.-M., and Laurob, C. *Comput Stat Data An* **48**(1), 159–205 (2005).

[41] Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, (2000).

[42] Schoelkopf, B. and Smola, A. J. *Learning With Kernels-Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, (2002).

[43] Ivanciuc, O. *Applications of Support Vector Machines in Chemistry in "Reviews in Computational Chemistry"*, volume 23. Wiley-VCH, Weinheim, (2007).

[44] Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, (1995).

[45] Vapnik, V. *Statistical Learning Theory*. New York: John Wiley & Sons., (1998).

[46] Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer, Berlin, (1995).

[47] Haykin, S. *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, (1998).

[48] Artemenko, N. V., Baskin, I. I., Palyulin, V. A., and Zefirov, N. S. *Russ Chem Bull* **52**(1), 20–29 (2003).

[49] Ali, K. M. and Pazzani, M. J. *Machine Learning* **24**, 173–202 (1996).

[50] Merkwirth, C., Mauser, H., Schulz-Gasch, T., Roche, O., and Stahl, M. andLengauer, T. *J Chem Inf Comput Sci* **44**(6), 1971–1978 Oct (2004).

[51] Tropsha, A., Gramatica, P., and Gombar, V. *QSAR Comb Sci* **22**(1), 69–77 (2003).

[52] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. *Bioinformatics* **16**(5), 412–424 (2000).

[53] Fawcett, T. *Pattern Recogn Lett* **27**(8), 861–874 JUN (2006).

[54] Lin, H.-T., Lin, C.-J., and Weng, R. C. *Mach Learn* **68**, 267–276 (2007).

[55] Dietterich, T. G. *Neural Comput* **7**, 1895–1923 (1998).

[56] Rücker, C., Rücker, G., and Meringer, M. *J Chem Inf Model* **47**(6), 2345–2357 (2007).

[57] Maggiora, G. M. *J Chem Inf Model* **46**(4), 1535 (2006).

[58] Aggarwal, C. *Outlier analysis*. Springer Verlag, (2013).

[59] Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J., Tong, W., Veith, G., and Yang, C. *ATLA* **33**, 155–173 (2005).

[60] Brandmaier, S., Peijnenburg, W., Durjava, M. K., Kolar, B., Gramatica, P., Papa, E., Bhhatarai, B., Kovarich, S., Cassani, S., D'Onofrio, E., Rahmberg, M., Öberg, T., Jeliazkova, N., Golsteijn, L., Comber, M., Ruggiu, F., Novotarskyi, S., Sushko, I., Kunwar, P., Abdelaziz, A., and Tetko, I. V. *ATLA, Altern Lab Anim* **42**, 13–24 (2013).

[61] Jaworska, Joannaand Nikolova-Jeliazkova, N. and Aldenberg, T. *ATLA* **33**, 445–459 (2005).

[62] Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., and Todeschini, R. *Molecules* **17**, 4791–4810 (2012).

[63] Varnek, A. and Baskin, I. I. *J Chem Inf Model* **52**(6), 1413–1437 May (2012).

[64] Varnek, A., Fourches, D., Kireeva, N., Klimchuk, O., Marcou, G., Tsivadze, A., and Solov'ev, V. *Radiochim Acta* **96**, 505–511 (2008).

[65] Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., and Varnek, A. *J Chem Inf Model* **48**(9), 1733–1746 Sep (2008).

[66] Horvath, D., Marcou, G., and Varnek, A. *J Chem Inf Model* **49**(7), 1762–1776 Jul (2009).

[67] Mitchell, M. *An Introduction to Genetic Algorithms. fifth edition.* Massachusetts Institute of Technology, (1999).

[68] Coley, D. A. *An Introduction to Genetic Algorithms for Scientists and Engineers.* World Scientific Publishing Co. Pte. Ltd, (1999).

[69] Solov'ev, V. P. and Varnek, A. (2011). ISIDA (In Silico Design and Data Analysis) QSPR program v5.76, Strasbourg - Moscow.

[70] Solov'ev, V. P., Varnek, A., and Wipff, G. *J Chem Inf Comput Sci* **40**, 847–858 (2000).

[71] Varnek, A. and Solov'ev, V. P. *Comb. Chem. High Throughput Screening* **8**, 403–416 (2005).

[72] Tetko, I. V. and Tanchuk, V. Y. *J Chem Inf Comput Sci* **42**, 1136–1145 (2002).

[73] Tetko, I. V. *J Chem Inf Comput Sci* **42**, 717–728 (2002).

[74] Tetko, I. V. *Neural Processing Letters* **16**, 187–199 (2002).

[75] *Weka v. 3.7.6, University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/*, (2012).

[76] Chang, C.-C. and Lin, C.-J. *ACM Trans Intell Syst Technol* **2**(3), 1–27 (2011).

[77] Ghose, A. K. and Crippen, G. M. *J Comput Chem* **7**(4), 565–577 (1986).

[78] Viswanadhan, V. N., Ghose, A. K., Reyankar, G. R., and Robins, R. K. *J Chem Inf Comput Sci* **29**, 163–172 (1989).

[79] Marcou, G., Horvath, D., Solov'ev, V., Arrault, A., Vayer, P., and Varnek, A. *Mol Inf* **31**, 639–642 (2012).

Chapter 3

ISIDA descriptors

As mentioned previously, the ISIDA fragment descriptors were extended to include a pH-dependent count and mappings of different properties prior to their fragmentation. The three different marked atom strategies, the triplet fragmentation, the formal charge indication and a newer system for the representation of bonds were also implemented. This chapter is dedicated to explain the details of the different possibilities of ISIDA descriptors. They are chiefly composed of a combination of a mapped property onto the molecular graph, a fragmentation scheme and a counting strategy. A nomenclature was developed to characterise each of them and are coded according to the following:

TopologicalFragmentationMappingTypeBondInclusion
(LowerLength-UpperLength)CountingType_Options.
Indications of the nomenclature will be given throughout the explanations and summarised at the end.

3.1 Property-mapping on the molecular graph

Fragment descriptors are classically calculated on the molecular graph with the nodes (aka vertices) indicated as the atom symbol and the edges as the bond orders. However, nodes (and even edges) may be associated to other properties and they can be mapped to give a different "colouration" of the graph. This process can be considered as a form of inductive learning transfer where the information is added via the definition of the descriptors. In this work, three different approaches have been differentiated to map properties onto the molecular graph: properties defined by substructures, increments calculated from substructures and increments calculated from QSAR modelling.

Note: Nodes without any associated property will be recognised as "flagless" and indicated by a "§" symbol. Nodes may also have several properties associated to them – all the combination will then be taken into account when creating fragments (see section counting).

3.1.1 Mapping of properties defined by substructure

Atom symbol (Nomenclature:A) : The nodes are simply represented by the atomic symbol of the atom, which corresponds to "classic" fragments.

Pharmacophoric properties (Nomenclature:Ph) : Pharmacophoric flags are derived from pharmacophore modelling. Pharmacophores are defined[1] as "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response" according to the IUPAC recommendations. Different rules may be enumerated to obtain these from the substructures and they are not universal. Rules used in this work are as follows (see Figure 3.1 for an example):

- Aromatic atoms are flagged as "R"
- Carriers of positive charges are flagged as "P"
- Centres of negatively charged functional groups are flagged as "N"
- Any oxygen or nitrogen bound to a hydrogen is flagged as "D" (HB donor)
- Any oxygen or nitrogen or negative sulphide or thiourea (=S) is flagged as "A" (HB acceptor)
- Any carbon or halogen except if concerned by the rules above is flagged as "H" (hydrophobe)

The atomic pharmacophoric types were attributed with the ChemAxon's PMapper[2]

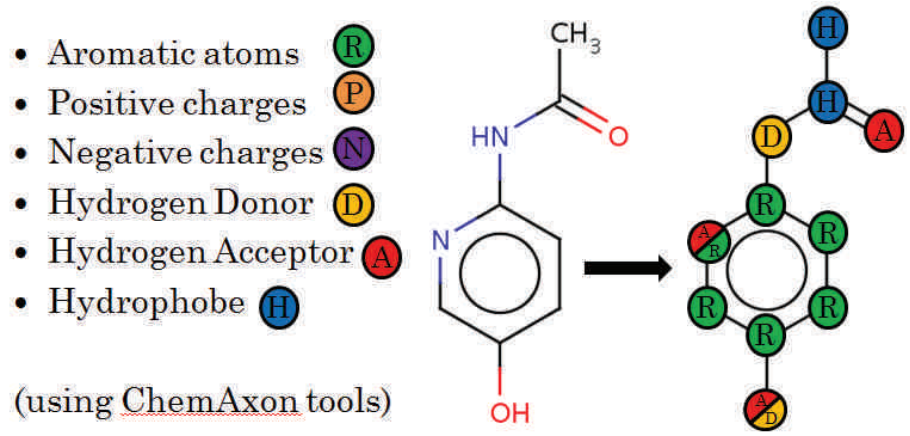


Figure 3.1: Pharmacophoric rules and example of pharmacophoric graph colouration

Force field typing (Nomenclature:Ff) : Force fields contain definitions of types of atoms by their substructures to differentiate different types of atoms for which the interaction parameters differ. In this work, the consistent valence forcefield's definitions were used[3].

Benson atoms (Nomenclature:Ba) : Benson atoms differentiate different types of carbon. These are presented in Table 3.1. If a carbon belongs to two classes, the most important one according to the Priority column will be used. These were not used in the studies of this work, but they were implemented in previous definitions of the ISIDA fragments.

Table 3.1: Benson atom definition

Priority	Type of atom	Symbol
1.	Aromatic C	CB
2.	Triple-bonded CN (C#N)	CN
3.	Triple-bonded C (C#)	CT
4.	Twice double-bonded C (=C=)	CA
5.	Ketone (C=O)	CO
6.	Double-bonded C (C=)	CD

3.1.2 Mapping from property increments

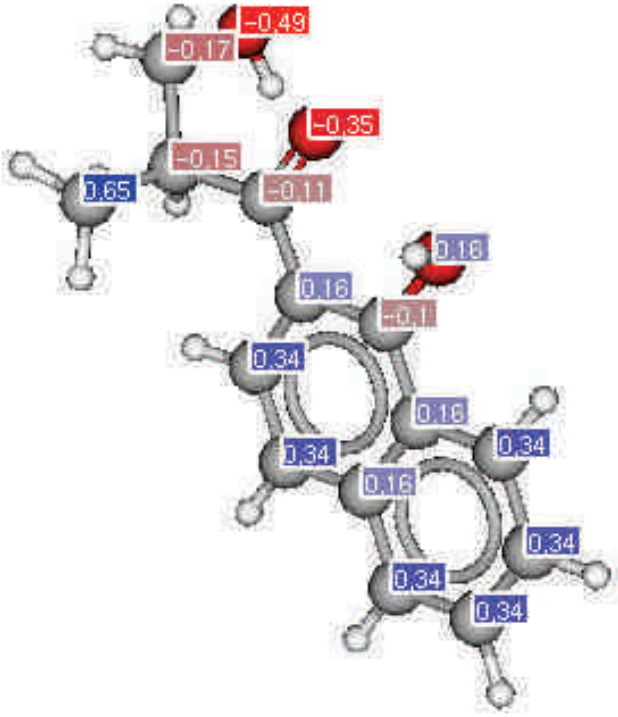


Figure 3.2: Example of LogP increments

Increments (see Figure 3.2 for an example) are a good way to map a property onto a molecular graph, however, being continuous, they cannot be used directly to label fragments - they would produce an infinite spectrum of possible entities. In order to bin

these values, the distributions of the increments on concerned sets have been visualised using histograms and kernel density estimations[4,5] to establish boundaries between the bins. Figure 3.3 The idea is that atoms with similar values have a similar impact on the property and should be represented by the same symbol or flag. The bins may overlap and the corresponding atom will thus have two symbols associated to it. Bins boundaries were chosen preferably in areas of low density.

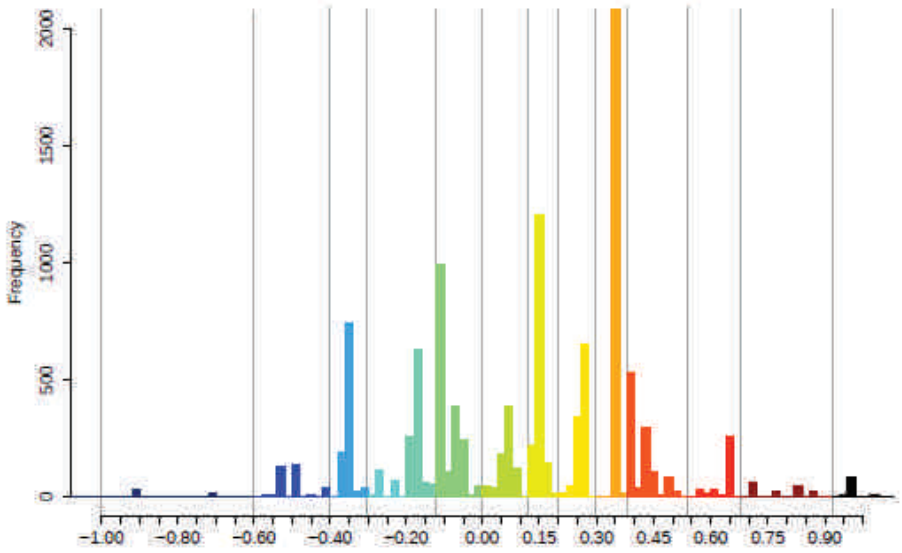


Figure 3.3: Example of distribution of increments and chosen bins

Increments based on QSAR models have been tried on GPCRs and are reported in section 6.2. Increments based on substructures defined obtained by various methods have been used in several benchmark and are described in the next section. Details about the distribution are given in the relevant studies were these mappings have been used.

3.1.3 Increments calculated from substructures

Partial charges (Nomenclature:Pc) were calculated according to Gasteiger's method[6–11] which is based on the electronegativity of the σ and π bonds. Many other models of partial charges exists but we only used Gasteiger's partial charges.

Topological electrostatic potential colouration (Nomenclature:Ep): The topological electrostatic potential V_i of each atom i are calculated from the partial charges according to:

$$V_i = \frac{q_i}{d_0} + \sum_{j \neq i} \frac{q_j}{d_i j}$$

with q_j the partial charge on atom j, q_i the partial charge on atom i, $d_i j$ the topological distance between atom i and j and d_0 an empirically determined virtual distance to take into account the concerned atoms charges.

LogP increments (Nomenclature:Lp) were calculated using the Ghose-Crippen approach[12,13]. The atoms are classified according to their element, their oxidation state and the surrounding atoms into 120 categories with an associated increment value.

3.1.4 Formal Charge indication

An option permits to add the information of the formal charge on an atom behind its symbol in the fragment. This option is useful to differentiate protonation states. When used, it is indicated in the nomenclature in the Options section with FC.

3.1.5 Bonds

Fragments may be generated including or excluding the information about the bond order. The different possible bonds indicated are summarized in Table 3.1.5. When bonds are included in the fragment description, it is indicated by a B in the BondInclusion section. Fragments may also be generate with only the bonds represented and without any symbol corresponding to the atoms. In this case, the MappingType section is left empty and a B is indicated in the BondInclusion.

Note that some of the bonds presented in Table 3.1.5, such as creation of a single bond or Triple bond to aromatic, are specially designed for Condensed Graph of Reaction[14–17]. These bonds are named dynamic bonds and two options permit to single out fragments containing them:

- All dynamic bonds (Nomenclature:AD): Only fragments containing only dynamic bonds are kept.
- One Dynamic Bond (Nomenclature:OD): Fragments with at least one dynamic bond are kept.

Although I have helped in the implementation and the new nomenclature of such bonds, these were never used in my work; only the four first bond orders were present in the sets.

Bond Type	Symbol
Simple	-
Double	=
Triple	+
Aromatic	*
Single or Double	.
Single or Aromatic	:
Double or Aromatic	”
Any bond type	?
Special bond type	_
Single bond in cycle	.
Double bond in cycle	:
Triple bond in cycle	#
Hydrogen bonds	~
Unknown bond	YY
Single bond creation	81
Double bond creation	82
Triple bond creation	83
Aromatic bond creation	84
Single bond cut	18
Double bond cut	28
Triple bond cut	38
Aromatic bond cut	48
Single bond to double bond	12
Single bond to triple bond	13
Single bond to aromatic bond	14
Double bond to single bond	21
Double bond to triple bond	23
Double bond to aromatic bond	24
Triple bond to single bond	31
Triple bond to double bond	32
Triple bond to aromatic bond	34
Aromatic bond to single bond	41
Aromatic bond to double bond	42
Aromatic bond to triple bond	43

3.2 The different fragmentation schemes

ISIDA descriptors include three basic patterns of fragmentation of the molecular graph: a) Sequences, b) Atom-centred fragments and c) Triplets, which are explained in the following paragraphs.

a. Sequences (Nomenclature: I) are strings of successive connected atoms and/or bonds in the molecular graph (see Figure 3.4). It corresponds to the shortest possible path between each pair of atoms.

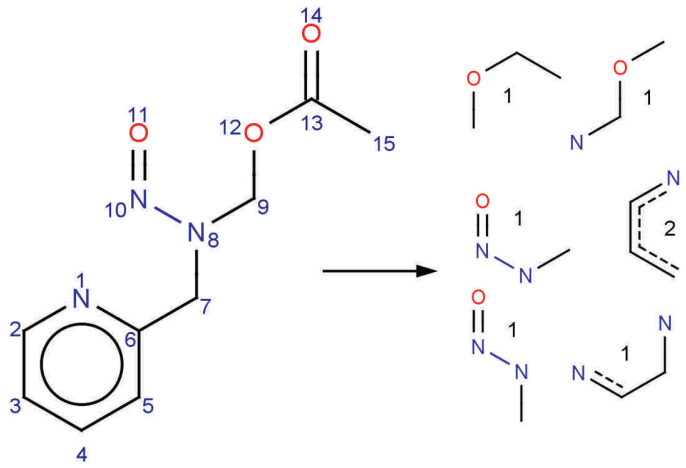


Figure 3.4: A few examples of sequences of length 4 and their count from a molecular graph

b. Atom-centred fragments (Nomenclature: II) start from an atom and encode the connected atoms to a certain topological distance(see Figure 3.5). These include so-called neighbouring atoms (topological distance = 1) or augmented-atoms as well as extended augmented atoms (topological distance > 1).

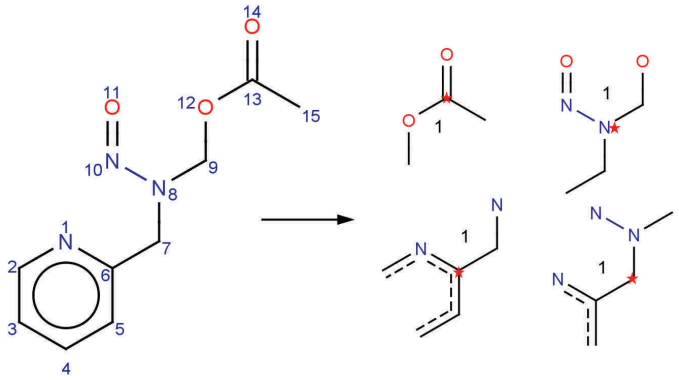


Figure 3.5: A few examples of atom-centred fragments of sphere 2 and their count. The central atom is indicated by a star

c. Triplets (Nomenclature: III) are all the possible combinations of 3 atoms in a graph with the topological distance between each pair indicated. For example, the triplet formed by the atom number 1, 11 and 13 in Figure 3.5 will yield a triplet of the type: N5O5C6 where d(1,11)=5, d(11,13)=5 and d(1,13)=6.

3.2.1 Fragment Length

The fragments are searched at defined minimum and maximum lengths. This length is defined by the number of atoms in the fragment instead of the topological distance. Hence, the lengths indicated correspond to the topological distance + 1. For example, if

atom counts are included, the minimum length will be 1. The boundaries are indicated in the nomenclature at the (LowerLength-UpperLength) section.

In the case of sequences, all possible lengths between the minimum and maximum will be searched for. In the case of triplets, the lengths indicate the minimum and maximum distances between the triplet's vertices.

In the case of atom-centred fragments, the length determines the distance between the central atom and the final vertex of a possible path starting at the central atom. All distances between the minimum and maximum length are explored. A path starting from the central atom may end before reaching the imposed distance. By default, the path will still be indicated in the fragment although it will not correspond to the length of the other paths. An option exists to restrict the fragment to contain only paths of the demanded length. It is indicated in the nomenclature in the Options section with a R.

3.2.2 Atom pairs

Atom pairs are fragments where the symbol of two atoms are given with the topological distance separating them. These are seen in ISIDA descriptors as a combination of sequences and the atom pairs option. When used, it is indicated in the nomenclature in the Options section with a P. This option can also be used in combination with atom-centred fragments. The resulting fragments can be assimilated to a "n-uplets" where the size n of the monitored multiplet is not rigidly dictated by the user (as in pharmacophore pairs, triplets, quadruplets) but is free to reflect the actual topological environment of a molecule. In the publication presenting IPLF descriptors[18], these were named "trees".

3.2.3 Marked atom strategy

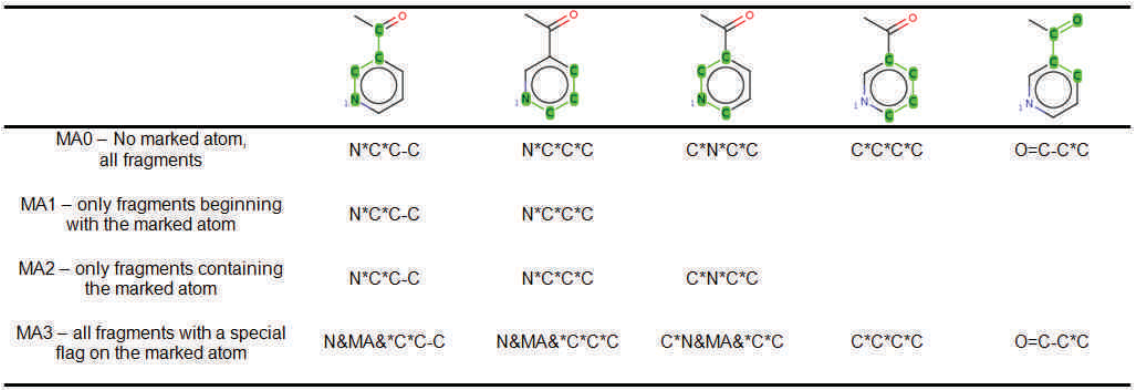


Figure 3.6: Example of the different Marked atom strategies with sequences of length 4 in 3-acetyl pyridine with the N as marked atom. If the path is not represented in the descriptor, the field is left empty

The Marked Atom (MA) strategy consists of indicating an atom of importance, in particular for "local" properties such as the acidic dissociation constant or the hydrogen bond strength. For these properties, the information of which atom is involved in the reaction/interaction is essential. The information of such an atom can be injected into the description by different means; three strategies have been developed:

1. : Fragments start with the MA. In the case of atom-centred fragments, the central atom corresponds to the marked atom.
2. : Only fragments containing the MA are kept.
3. : A special flag (&MA&) is added to the symbol(s) representing the MA. All fragments are generated.

Descriptors using such a strategy were named local descriptors and an example of the strategies are given in Figure 3.6.

3.2.4 Path exploration

By default, paths between two atoms (for examples between the central atom and one extremity of an atom-centred fragment) are computed in order to obtain the shortest possible path between them. In case, several possibilities correspond to the shortest distance, all possibilities will be enumerated as a fragment. It is however possible to compute all the possible paths using an option. When used, it is reflected in the nomenclature with an AP in the Options section.

3.2.5 Wildcard

The wildcard option permits to explore all possible fuzziness of fragments between the fully described fragment and the paired fragment. The nature of intermediate interconnecting atoms in a fragment may be optionally mentioned or ignored. In this sense, a "pair" of a Hydrophobe H at 5 bonds away from a hydrogen bond Acceptor A may be seen as the fuzzy common denominator of all pairs HHHHHA, HHAHHA, ...,H?HHHA,HH???H,etc., where "?" is added as a wildcard to state that the atom type at that point does not matter. When used, it is reflected in the nomenclature with in the Options section with a W.

3.3 Counting strategies

3.3.1 Occurrence count

By default, the occurrence of a fragment in a molecule corresponds to its value in the vector representing this molecule. Nodes (atoms) may have several associated properties represented by a symbol/"flag". In this case, subgraphs on which the fragments are based including such an atom will generate different fragments and they will all be counted.

3.3.2 pH-dependent counting

In order to obtain a pH-dependent count, the micro-species and populations are generated (using the ChemAxon pK_a plugin[19]) and fragments are generated on each of the micro-species. Each of their occurrence is weighted by the micro-species population in percentage. In the end, all weighted occurrences of each fragment from all micro-species are summed to represent the molecule. An example of this counting strategy with pharmacophoric mapping is shown in Figure 3.7. The top left fragment, A*R*R*N, has a

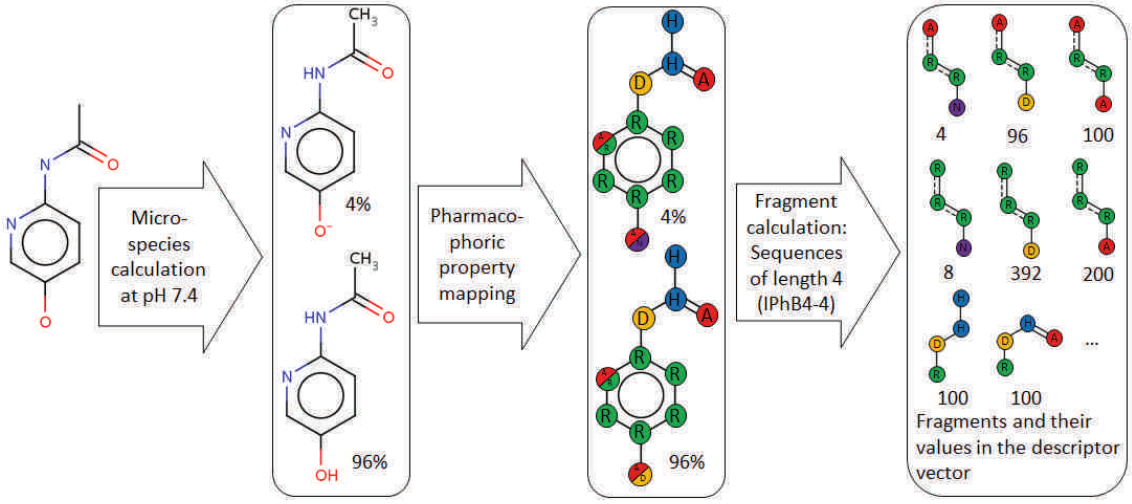


Figure 3.7: Calculation workflow for ISIDA descriptors as sequences with pharmacophoric mapping

value of 4 because it appears in the first micro-species with a population of 4% once. The fragment in the centre, R*R*R*D, occurs twice in the first micro-species and 4 times in the second, thus its value is equal to $2 \times 4 + 4 \times 96 = 392$. Notice that certain atoms have 2 flags and these are accounted for as an occurrence just as explained in the previous paragraph.

3.4 Nomenclature summary

To characterize the different ISIDA fragment descriptors, they are coded according to the following:
TopologicalFragmentationMappingTypeBondInclusion
(LowerLength-UpperLength)CountingType_Options
Where:

1. TopologicalFragmentation is a Roman number and corresponds to the following fragmentation types:
   - I - Sequences
   - II - Atom-centred fragments
   - III - Triplets
2. MappingType s a chain of letters starting with a capital and followed by only lower case letters. The following codes have been used up to now:
   - A - Atom symbol
   - Ph - Pharmacophoric properties
   - Ff - Force field typing

- Ba - Benson atoms
- Pc - Partial charges
- Ep - Topological electrostatic potentials
- Lp - LogP increments

3. BondInclusion simply indicates whether bonds were used with a capital B. If only bonds are used then no ColourationType will appear.
4. LowerLength and UpperLength are the number of atoms to be included at minimum and maximum respectively. Note that the number of atoms correspond to the topological distance + 1, hence, for example, a LowerLength=2 and UpperLength=5 will create fragments with at minimum a topological distance of 1 and maximum a topological distance of 4.
5. CountingType corresponds to the type of weight used to count the occurrences of fragments. When none is indicated then the simple fragment count is used (weight =1):
   - pH-dependent (ms): Micro-species population in percentage at a given pH (by default 7.4) will be used to weight the occurrence of the fragments in their corresponding micro-species.
6. Options indicate special options used during the fragmentation and are listed below. When several are used, they are separated by a hyphen (-).
   - P - Pairs are generated (only the extremities of the fragment are shown and the topological distance between them)
   - R - Restricted path length (only for atom-centred fragments)
   - AP - All Path exploration
   - FC - Formal Charge representation
   - MAX - Marked Atom strategy, where X stands for the used strategy:
     (a) Fragments starting with MA only
     (b) Fragments with MA only
     (c) Special flag is added to MA
   - AD - Fragments with All Dynamic bonds
   - OD - Fragments with One Dynamic bond

Example: II_PhB(3,5)ms_P-FC are paired atom-centred fragments (trees) with a pharmacophoric mapping and bond information as well as formal charges are indicated. Their length vary from 3 to 5 atoms. The micro-species and their population were used to generate and count the fragments.

Note: Previous publications[14,18] indicated to a different nomenclature used with previous versions of the ISIDA Fragmentor. This “new“ nomenclature is an attempt to unify all

the possibilities of the programs in one nomenclature.

Bibliography

[1] Wermuth, C.-G., Ganellin, C. R., and Mitscher, L. A. *Annual Reports in Medicinal Chemistry* **33**, 394 (1998).

[2] *ChemAxon PMapper, JChem 5.7.1, http://www.chemaxon.com/jchem/doc/user/PMapper.html*, (2013).

[3] Hagler, A. T., Huler, E., and Lifson, S. *J Am Chem Soc* **96**, 5319–5327 (1974).

[4] Rosenblatt, M. *Ann Math Statist* **27**(3), 832–837 (1956).

[5] Parzen, E. *Ann Math Statist* **33**(3), 1065–1076 (1962).

[6] Gasteiger, J. and Marsili, M. *Tetrahedron Letters* **19**(34), 3181–3184 (1978).

[7] Gasteiger, J. and Marsili, M. *Fresenius Zeitschrift Fur Analytische Chemie* **304**(4), 258–259 (1980).

[8] Gasteiger, J. and Marsili, M. *Tetrahedron* **36**(22), 3219–3228 (1980).

[9] Marsili, M. and Gasteiger, J. *Croatica Chemica Acta* **53**(4), 601–604 (1980).

[10] Guillen, M. D. and Gasteiger, J. *Tetrahedron* **39**(8), 1331–1335 (1983).

[11] Gasteiger, J.; Saller, H. *Angew Chem , Int Ed* **24**(8), 687–689 (1985).

[12] Ghose, A. K. and Crippen, G. M. *J Comput Chem* **7**(4), 565–577 (1986).

[13] Viswanadhan, V. N., Ghose, A. K., Reyankar, G. R., and Robins, R. K. *J Chem Inf Comput Sci* **29**, 163–172 (1989).

[14] Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. *J Comput Aid Mol Des* **19**(9-10), 693–703 Jul (2005).

[15] Hoonakker, F., Lachiche, N., Varnek, A., and Wagner, A. In *19th International Conference on Inductive Logic Programming (ILP'09)*, (2009).

[16] De Luca, A., Horvath, D., Marcou, G., Solov'ev, V., and A., V. *J Chem Inf Model* **52**(9), 2325–2338 (2012).

[17] Muller, C., Marcou, G., Horvath, D., Aires de Sousa, J., and A., V. *J Chem Inf Model* **52**(12), 3116–3122 (2012).

[18] Ruggiu, F., Marcou, G., Varnek, A., and Horvath, D. *Mol Inf* **29**(12), 855–868 Dec (2010).

[19] *ChemAxon pKa plugin version 5.3.8*, (2011).

Chapter 4

Applications of ISIDA property-labelled fragment descriptors

4.1 Introduction

The different studies using various property mappings by substructures rules or increments are presented in this chapter, while the two following chapters describe the studies on the local descriptors and mapping by increments from QSAR modelling. The ISIDA Property-Labelled Fragment descriptors (IPLF) were first benchmarked with three properties and corresponding sets already studied at the laboratory to ensure their propensity for model building. These three studies (outlined in section 4.2) include a NB benchmark on the proteases binding affinity and two QSAR models on the octanol-water partition coefficient (LogP) and the binding affinity to the hERG ion channel. After good results in these studies, they were employed in the "Projet Interdisciplaire de Recherche" (PIR - Interdisciplinary Research Project) for the national French chemical library, the "Chimiothèque Nationale" (CN), for the modelling of the Chromatographic Hydrophobicity Index (CHI) (outlined in section 4.3). Finally, they were used in a joint chemogenomics project with the Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, University of Kyoto (Japan) (outlined in section 4.4). Hydrophobicity (LogP and CHI) and the different binding affinities are important to research in pharmaceutical science, in particular in drug discovery, as they play major roles in the pharmacokinetics and pharmacodynamics of the drugs respectively. As these studies were all published, the present chapter will first focus on a discussion of the measured properties, since the understanding of the underlying experimental work, with its limitations and potential sources of errors, is of paramount importance for modelling. Next, the corresponding articles will be introduced, optionally followed by additional details on the therein reported work. Eventually, some not yet published studies will be outlined.

4.1.1 Hydrophobicity

Hydrophobicity is a chemical concept rather difficult to define precisely. It can be considered as the propensity for molecules to repel water molecules which corresponds to the Greek roots of the word: hydro = water and phobic=repellent. It is associated to the concepts of hydrophilicity (which attracts water) and lipophilicity (which attracts lipids). In general, hydrophobicity should take into account all the interactions between molecules which would make them repel water and be attracted to a reference lipidic compound. Several scales were invented to measure these, including LogP and CHI.

Hydrophobicity has been identified as playing a major role in drug action for over a century with the pioneer works of Overton and Meyer[1] in 1899 which observed that the effect of anesthetics was related to the oil-water partition coefficient. The LogP became a standard property to describe the hydrophobicity of compounds due to the works of Hansch et al[2]. One of the main reasons for hydrophobicity to be so determining for drugs is the transportation throughout the body and membrane permeation to enter cells. A molecule should be hydrophobic enough to progress through the lipidic bilayer of the cell membrane but hydrophilic enough to be soluble in the water-base liquids such as the cytosol or the plasma. Being able to assess in advance the hydrophobicity of virtual molecules is primordial to find suitable drug candidates.

The octanol-water partition coefficient, LogP , is defined as the ratio of the concentration of a solute in the octanol (organic phase) and in the water. The solute is usually in its unionised form or at least the same protonation state should be found in both phases,:

$$LogP = log\frac{[solute]_{octanol}}{[solute]_{water}}$$

Albeit logP is thought of as characterising the "neutral" species, compounds with ionisable groups will spontaneously evolve in proteolytic equilibria with water, whilst compounds with complementary polar groups may associate in the organic phase. Such effects, in particularly the latter, may be difficult to predict on the basis of the structure. Ionisation effects may be better controlled if the aqueous phase is replaced by a constant pH buffer (in particular, a buffer mimicking physiological pH). When the latter is used, the corresponding partition coefficient is called the octanol-water distribution coefficient $LogD_{pH}$. Thus, the different micro-species at that pH are present which is important for the evaluation of ADMET properties. LogP and LogD are traditionally measured with the shake flask method. It consists of diluting the compound into a volume of the aqueous solution and one of octanol, then measure the compounds' concentration in each of the phases by UV-Vis spectroscopy or any other suitable detection method. It is time-consuming and cannot be automated. It also requires relatively big amounts of compounds, especially in regards to chromatographic methods. Therefore, the possibility of using chromatography and in particular High Pressure Liquid Chromatography (HPLC) has been desirable and different approaches have been described[3–5] in the literature. Due to the recent development of combinatorial synthesis and the generation of huge sets of compounds for drug

testing, it has become a necessity to develop High-Throughput Screening (HTS) methods for the property profiling of drug candidates and also in regards to hydrophobicity.

The Chromatographic Hydrophobicity Index, CHI , was suggested by Valko et al.[6,7] as a HTS method for the evaluation of hydrophobicity. It is derived from gradient retention times from a fast gradient Reverse-Phase High Pressure Liquid Chromatography (RP HPLC) and approximates the volume percentage of organic solvent in the mobile phase. The CHI is an approximation of the isocratic chromatographic hydrophobicity index (φ_0). In order to explain these two properties, a few definitions are necessary:

- φ: Organic phase concentration in volume percent in the mobile phase. For example, if the mobile phase is a mixture of acetonitrile/water 70/30 v/v then $\varphi = 70$. Note that the HPLC is done with the same concentrations of acetonitrile/water during the whole measurement unlike in the case of a gradient HPLC.
- t_0 : Dead time, aka column void or dwell time. It corresponds to the time between the injection and the retention time of an unretained reference solute.
- t_{iR} : Isocratic retention time. The retention time of a compound obtained by a HPLC with a mobile phase of fixed concentration.
- t_{gR} : Gradient retention time. The retention time of a compound obtained by a gradient HPLC.
- $logk'$: Isocratic retention factor, derived from t_{iR} and t_0 with $logk' = \frac{log(t_{iR}-t_0)}{t_0}$
- φ_0 : Isocratic chromatographic hydrophobicity index. It corresponds to the φ at which the isocratic retention factor $logk' = 0 \Rightarrow t_{iR} = 2 \times t_0$

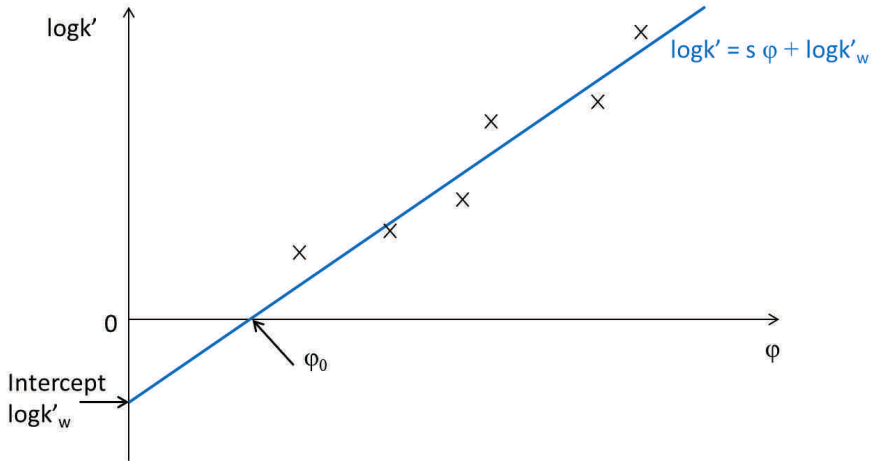


Figure 4.1: Plot of the isocratic retention factor $logk'$ against the mobile phase concentration φ in volume percentage to obtain the isocratic chromatographic hydrophobicity index φ_0

The isocratic chromatographic hydrophobicity index φ_0 corresponds to the volume percentage of organic phase required to achieve an equal distribution of the compound between the mobile and the stationary phase in the column. It will be influenced by the

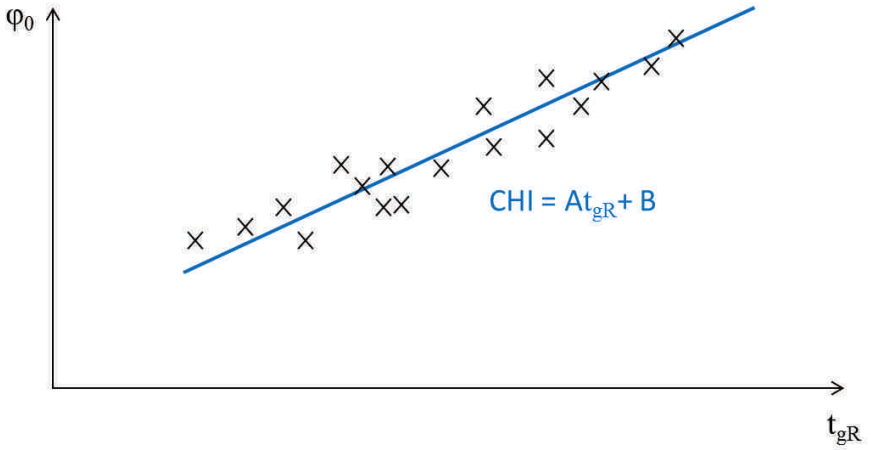


Figure 4.2: Plot of the isocratic chromatographic hydrophobicity index φ_0 against the gradient retention times t_{gR} to obtain the CHI

experimental conditions, in particular pH and the type of organic solvent. Therefore, in most experiments the solvent is indicated and a buffer is used to prevent pH variation. To acquire the value of φ_0, several measurements of the t_{iR} are performed at different isocratic mobile-phase concentrations φ with a RP-HPLC. The results are then plotted as $logk' = f(\varphi)$ and fitted to obtain a linear equation (see example on Figure 4.1). The linear fitted curve gives the following equation: $logk' = s \times \varphi + logk'_w$ where $logk'_w$ denotes the intercept and s the slope. When $logk' = 0$, then $0 = s \times \varphi_0 + logk'_w \Leftrightarrow \varphi_0 = -\frac{logk'_w}{s}$. The measurement of φ_0 gives a good estimate of the hydrophobicity of compounds using HPLC and has been shown to be correlated to logP on 500 drugs[4]. However, at least three different measurements per compounds needs to be executed, which is tedious and not HTS-compatible. Furthermore, it is condition dependent in regards to the pH and the mobile phase but also to the column. For these reasons, Valko et al. went further and tested an alternative index, the CHI derived from the retention times of a fast gradient RP HPLC and φ_0. Valko et al.[6] carried out a fast gradient RP-HPLC on a set of 76 compounds for which they had determined φ_0 and obtained the gradient retention times t_{gR} . A linear fitting of the plot $\varphi_0 = f(t_{gR})$ (see Figure 4.2) is used to calculate the CHI, i.e. $CHI = A \times tgR + B$. The CHI corresponds to the φ_0 predicted by the fitted linear equation. For later experiments, 10 compounds with known CHI are used to calibrate the column and the linear fit to these 10 compounds is used to assess CHI from the gradient retention times.

4.1.2 Binding affinity

A different aspect important to drug discovery is the interaction of a potential drug with a target protein, generally termed as the binding affinity. Evaluation of the strength of this interaction can be done by assessing the dissociation constant (K_D), associated to the CPI (see reaction shown in Equation 4.1).

$$P : C \rightleftarrows P + C \tag{4.1}$$

where P stands for the protein, C for the compound (ligand) and P:C for the complex formed between the two. It is often presented in a log scale, where $pK_D = -log(K_D)$.

Another way to assess the binding affinity is to evaluate the concentration of a drug needed to inhibit a biological activity by half. It is known as the half maximal inhibitory concentration (IC_{50}) and is often presented in its corresponding log scale, $pIC_{50} = -log(IC_{50})$. Many drugs exert a biological effect as a consequence of an inhibition of the function of a target such as an enzyme. Although, it may not always be the case, the studies of binding affinities have been coined with the term inhibition. IC_{50} does not account for the binding affinity directly, nevertheless, they are related as shown by Cheng and Prussof[8].

Binding affinity is often derived from assays evaluating IC_{50}. Many different assays exists[9] such as the commonly employed competition assays. In this case, a reference radioactive ligand is used and another ligand is added at different concentrations to see how the new ligand displaces the radioactive ligand. The resulting displacing curves are analysed to determine the IC_{50}. Another more direct assay would be to measure the induced inhibition or activity at different concentrations and determine the IC_{50} when half of the protein are inhibited or activated. For example, if the inhibition leads to apoptosis of the cells in the in vitro culture, a cell counting kits can be used to follow the effect. Certain cell counting kit consist of dying a viable cell in orange, thus, the number of cells can be counted using UV-Vis spectroscopy. The IC_{50} can then be measured as corresponding to the concentration where half of the cells are still alive.

A boundary value of IC_{50} or K_D is often defined in order to classify compounds into active/non-active and thereby change the problem to a two-class classification one.

In this thesis, one specific target and two target families have been studied: the hERG potassium channel, proteases and GPCRs.

4.1.2.1 Human Ether-à-go-go Related Gene channel

The hERG channel conducts potassium ions out of the heart's muscle cells. Several cases of noncardiac drugs leading to arrhythmia and sudden deaths have been identified in the mid-1990 for which, the hERG channel inhibition has been shown to be the cause[10,11].The hERG channel is, therefore, an essential target in the drug discovery and development when evaluating cardiac toxicity. The assay is of considerable cost and therefore, early recognition of potential inhibitors by in silico methods is of notable interest and several models have been published in recent years[12].

4.1.2.2 Proteases

The protein family of proteases, aka peptidase or proteinase, includes any enzyme that performs proteolysis. Proteases are involved in a multitude of physiological phenomena in all organisms. Their functions range from the digestion of proteins by either breaking a specific bond or decompose it entirely, to involvement in signalling pathways. In our study, 5 different serine proteases were involved: Chymotrypsin (found in the duodenum), Factor Xa (FXA) (found in the liver), Trypsin (found in the pancreas), Tryptase (found in mast cell), Urokinase-type Plasminogen Activator (uPA). Serine proteases use a serine in the active site as a nucleophile to catalyse the proteolysis[13]. Proteases may be targeted for

various pharmaceutical goals. For example, a famous serine protease is the thrombin which is involved in the coagulation process and is targeted by anti-thrombosis drugs.

4.1.2.3 G Protein-Coupled Receptors

The GPCRs are a family of proteins located in the cell membrane whose function is to transmit information into the cell[14]. The signals transmitted by the GPCRs include photons, odours, tastes, hormones, and neurotransmitters. About 400-500 GPCRs recognize non-sensory ligands and are potential drug targets[15]. Most marketed drugs target a GPCR[16],which are therefore intensively studied. An example of a well known GPCR is the dopamine receptor family which is involved in the reward pathways in the brain. Drugs such as cocaine and amphetamines target it.

4.2 Initial benchmarks of IPLF descriptors

The first three benchmarking studies of the IPLF descriptors were published in 2010 in Molecular Informatics[17]. The publication is reproduced in the following section. Afterwards, more details on the methodology for the NB study are given which are not explicitly detailed in the publication and results with the descriptor spaces are reproduced indicating the new nomenclature which was explained in the previous chapter. The old nomenclature are coded according to the following:
TopologicalFragmentationBondInclusionWildcard Options
MappingTypeLowerLengthUpperLength
Translation of the old nomenclature into the new is given in Table 4.1.

Table 4.1: Old and New ISIDA descriptors nomenclature

Topological fragmentation schemes notation		
Old	New	Comment
aa	II	atom-centred fragments
tree	II..._P	trees are atom-centred fragments with the pairs option
seq	I	sequences
pair	I..._P	atom pairs (sequences with the pairs option)
Labelling strategies		
Old	New	Comment
SY	A	Atom symbols
PHTYP	Ph	Pharmacophoric properties
EPTYP	Ep	Topological electrostatic potentials
Other options		
Old	New	Comment
b	B	bond information
w	W	wildcard

Note that the program used in these study, Fragdesc (see 7.4), only computes all paths so the AP option should be indicated in the new nomenclature of the descriptor. Also, by default, the micro-species dependent counting was used with the pharmacophoric properties and topological electrostatic potentials labelling strategies.

The publication published in Molecular Informatics[17] on the first benchmarking studies of IPLF descriptors follows in the next pages. It is reproduced with authorisation from all authors.

molecular informatics

DOI: 10.1002/minf.201000099

ISIDA Property-Labelled Fragment Descriptors

Fiorella Ruggiu,[a] Gilles Marcou,[a] Alexandre Varnek,[a] and Dragos Horvath*[a]

Contribution to the 2nd Strasbourg Summer School on Chemoinformatics, VVF Obernai, France, June 20–24, 2010

Abstract: ISIDA Property-Labelled Fragment Descriptors (IPLF) were introduced as a general framework to numerically encode molecular structures in chemoinformatics, as counts of specific subgraphs in which atom vertices are coloured with respect to some local property/feature. Combining various colouring strategies of the molecular graph – notably pH-dependent pharmacophore and electrostatic potential-based flagging – with several fragmentation schemes, the different subtypes of IPLFs may range from classical atom pair and sequence counts, to monitoring population levels of branched fragments or feature multiplets. The pH-dependent feature flagging, pursued at the level of each significantly populated microspecies involved in the proteolytic equilibrium, may furthermore add some competitive advantage over classical descriptors, even when the chosen fragmentation scheme is one of the state-of-the-art pattern extraction procedures (feature sequence or pair counts, etc.) in chemoinformatics. The implemented fragmentation schemes support counting (1) linear feature sequences, (2) feature pairs, (3) circular feature fragments a.k.a. "augmented atoms" or (4) feature trees. Fuzzy rendering – optionally allowing nonterminal fragment atoms to be counted as wildcards, ignoring their specific colours/features – ensures for a seamless transition between the "strict" counts (sequences or circular fragments) and the "fuzzy" multiplet counts (pairs or trees). Also, bond information may be represented or ignored, thus leaving the user a vast choice in terms of the level of resolution at which chemical information should be extracted into the descriptors. Selected IPLF subsets were – tree descriptors, in particular – successfully tested in both neighbourhood behaviour and QSAR modelling challenges, with very promising results. They showed excellent results in similarity-based virtual screening for analogue protease inhibitors, and generated highly predictive octanol-water partition coefficient and hERG channel inhibition models.

Keywords: Molecular descriptors · Fragment counts · Pharmacophore features · Electrostatic potential · Virtual screening · Neighbourhood behaviour · QSAR · logP · hERG · Protease inhibition

1 Introduction

For many fields such as drug discovery and toxicology, prediction of compound properties and biological activity using in silico approaches is of paramount importance.[1–4] Standard methods include quantitative structure-activity relationships (QSAR)[5–8] and similarity-based virtual screening.[9–12] However, although an enormous amount of descriptors[13–16] are available and data mining[17–19] methods are implemented for chemoinformatics tasks, significant differences of activity still arise among molecules perceived as similar by the computational tool (activity cliffs).[20,21] Activity cliffs, apparently violating the similarity principle "similar molecules are likely to have similar properties", are a complex problem, and it is not straightforward to distinguish the "genuine" situations from the cases related to an inappropriate chemical space and activity landscape definition (inappropriate molecular representation, i.e. descriptors erroneously stating that the molecules were "similar"). Therefore, research efforts are still made to improve descriptors.

Medicinal chemists view the molecule as interconnected substructures (groups) and interpret activity by increments of positive and negative contributions of these fragments. Many substructural descriptors, belonging to the large family of fragment counts,[22] exist and are readily calculable from the molecular graph. These are based on a wide range of fragmentation schemes, ranging from detection of predefined groups (MACCS[23] keys, DayLight fingerprints[24]) to open-ended detection of all fragments of a specified type (pairs,[25] linear or branched fragments of specified minimal and maximal sizes). The latter include the ISIDA fragment counts[26–30] which feature sequences and ramified substructures, called augmented atoms, including or not bond type information, according to the users' needs.

Fragment counts capture the information about the nature of involved atoms by simply reporting the atom symbols, which may, at first sight, be insufficient to render their actual, chemical context-dependent properties. For example, the presence of a C–C–C–N–C sequence in a mole-

[a] *F. Ruggiu, G. Marcou, A. Varnek, D. Horvath*
Laboratoire d'Infochimie, UMR 7177 Université de Strasbourg-CNRS
Institut de Chimie, 4, rue Blaise Pascal, 67000 Strasbourg, France
phone: + 33687934703
**e-mail: horvath@chimie.u-strasbg.fr*

cule would not tell us whether the N atom is an amino group or an amide N. This information may not be altogether lost, but might be inferred from the analysis of the entire set of populated fragments – it is implicitly "hidden" in the descriptors. However, the idea to represent atoms by specific labels, which are more informative than their symbols and tailored to the needs of the specific property prediction problem, is not new. As a molecule is often thought of as a "coloured" graph, the vertices being tainted according to the nature of occupying atoms, alternative labelling procedures may be also called new "colouring schemes".

Fragment descriptors based on different colouring schemes were already introduced – so, for example, the extended-connectivity fingerprints[31] (ECFPs) by Accelerys. They are conceptually similar to augmented atom and integrate properties such as Daylight atomic invariants rule, functional pharmacophore role, Sybyl atom types or *a*log*P* atom codes. However, these are not readily interpretable (they rely on a cumbersome fragment labelling scheme) and only ramified fragments can be generated.

In particular, pharmacophore type-based colouring has been extensively used to capture the nature of putative interactions (hydrophobic contacts, hydrogen bonds, salt bridges) a functional group in a ligand might be involved in when binding to a protein. This colouring scheme is transversal to colouring by atomic symbols (various atoms may belong to a same pharmacophore type). Pharmacophore descriptors typically focus on the relative position of pharmacophore groups in the molecule, not paying much attention to the manner in which these are interconnected – the pharmacophoric pairs,[32,33] triplets[34,35] and 4-point descriptors,[36] which enables them to perform successful "scaffold hopping" (discovery of new scaffolds porting the given pharmacophore pattern seen in known actives).[37–39] By contrast, the above-mentioned pharmacophore-coloured ECFPs are genuine fragment counts. Apparently, pharmacophore multiplets and fragment counts seem to be conceptually different – in fact, they can be unified within the concept of "fuzzy" fragments, in which the nature of terminal fragment atoms is always explicit, while the nature of the intermediate, interconnecting atoms may be optionally mentioned or ignored. In this sense, a "pair" of a Hydrophobe H at 5 bonds away from a hydrogen bond Acceptor A may be seen as the fuzzy common denominator of all pairs H–H–H–H–H–A, H–H–A–H–H–A, …, H–?–H–H–H–A, H–H–?–?–?–H, etc., where "?" is added as a wildcard to state that the atom type at that point does not matter.

This work is set to create a conceptually unified descriptor calculation scheme, based on two key degrees of freedom taken into account in the calculation strategy:

The colouring scheme – allowing for the use of arbitrary atom typing schemes (symbols, pharmacophore types, electrostatic properties, lipophilicity, etc.). Furthermore, in order to ensure for a chemically relevant colouring, pH-dependent typing[34] is performed. Except for symbol-based colouring, pharmacophore and electrostatic property-based flags may differ, and will be rendered as a function of the actual protonation scheme of the compound at given pH.

The fragmentation fuzziness parameter, ranging from explicit fragment enumeration (including or not the bond type information) to atom multiplet counts in which the nature of interconnecting atoms is ignored. Note: in the latter, "fuzziness" will be consistently used in the sense of voluntarily ignoring some – or all – interconnecting atom features at the coloured fragment detection/counting step. This is different from the concept of fuzziness as tolerance with respect to a limited amount of variation of the distances separating two features.[40,41] This latter type of fuzziness is not (yet) supported in the current implementation of our descriptors.

This should allow for a comprehensive coverage of all possible descriptor schemes from strict fragment counts to fuzzy "scaffold-hopping" terms, times all the considered colouring schemes – in order to let machine learning pick the description level best suited to a given problem. Therefore, we generically refer to our descriptors as ISIDA property-labelled fragment descriptors (IPLFs). The list of fragments seen to occur in a molecule (respectively the combined list of all fragments found in all its microspecies being populated at the given pH, for pH-dependent flagging[34]) and their cumulated population levels define the molecular fingerprint, to be used in similarity scoring and QSAR studies.

Three studies: a neighbourhood behaviour (NB) benchmarking study on five proteases,[42] and two quantitative structure-activity relationships (QSAR) for the logarithm of the *n*-octanol-water partition coefficient[43] (log*P*) and the human Ether-a-go-go Related Gene (hERG) potassium channel blocking potency[44,45] were made to evaluate the capacity to encode relevant information of the descriptors and the robustness of the generated descriptor spaces.

2 Methods

IPLF generation (Figure 1) implies two distinct steps:

"Type": Atomic typing – applying all the considered colouring schemes to each molecule in the data set, after analysis of its possible pH-dependent protonation states, in order to produce a labelled output file, where each atom is flagged according to each scheme.

"FragDesc": Fingerprint build-up, allowing the user to combine any of the supported fragmentation schemes with any of the provided colouring schemes to obtain the corresponding molecular fingerprints.

2.1 The Typing Program

The typing program is written in java and makes use of several ChemAxon tools such as the standardizer,[46] the pharmacophore mapper,[47] and various property calculation

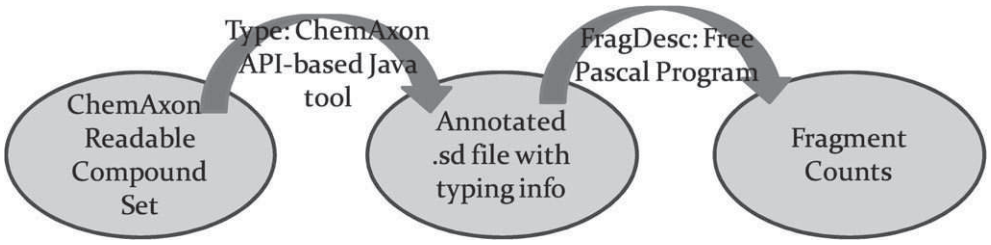


Figure 1. Main workflow of the IPLF build-up.

plugins (ionization,[48] charge calculation[49]). The input may be of any format supported by ChemAxon. Each molecule is treated at a turn.

The basic application work flow (Figure 2) starts with the conversion of the input molecule into a standardized internal representation (using the standardizer). Then, the molecule's microspecies are calculated by the pK_a plugin and only those having a population level above a threshold of 1% are retained. The populations of each retained microspecies are given in percent and rounded up to an integer value giving a total of 100. The idea pf pK_a-dependent pharmacophore typing has been first used with fuzzy pharmacophore triplets, and extensively benchmarked in previous publications. This may prove essential[11,50] to explain otherwise counterintuitive "activity cliffs", but has virtually no impact[42] on molecules with simple protonation patterns (with a single dominant microspecies that can be readily pinpointed by empirical rules, such as "aliphatic amines are protonated"). Structure-activity relationships may, intriguingly, sometimes loose[51] predictive power due to rigorous pK_a-sensitive flagging, compared to some chemically wrong but "lucky" protonation scheme assignment. If this latter introduces a specific systematic error for the "actives" or "inactives" of the training set, therefore spuriously facilitates their separation. The present work will therefore not reopen the already well-addressed topic of the impact of the pK_a-sensitive flagging scheme on descriptors.

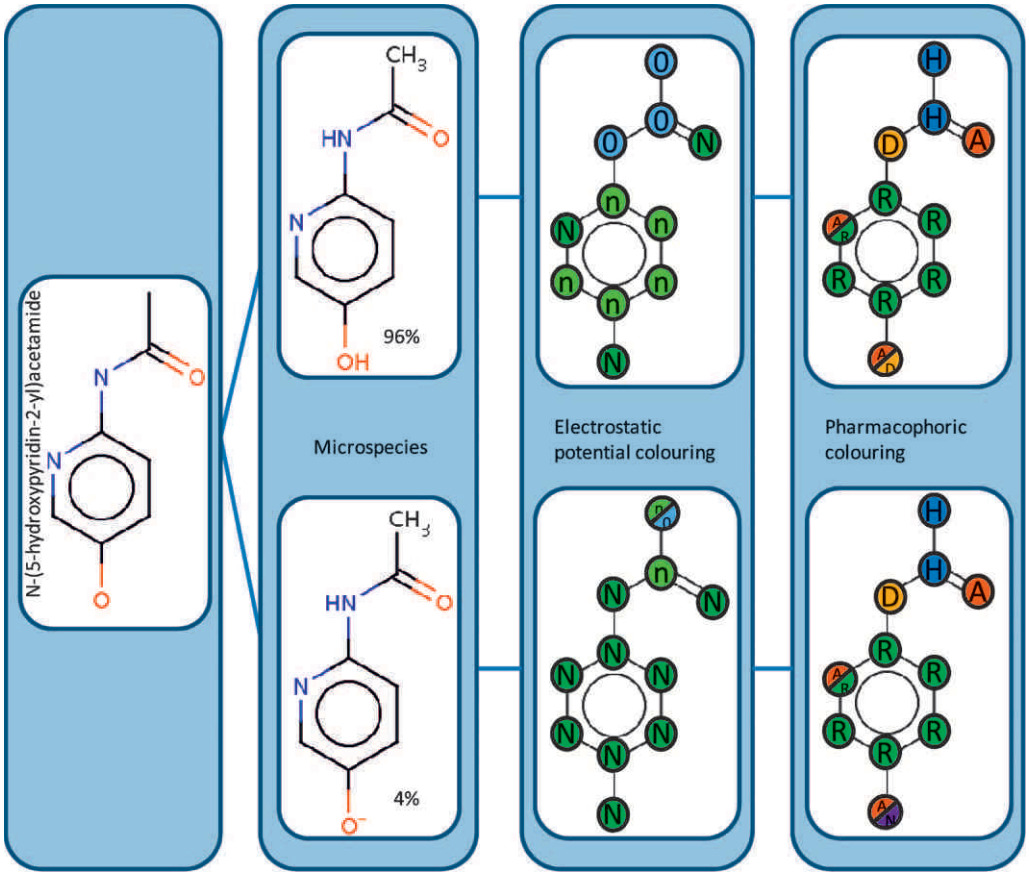


Figure 2. Typing work flow on *N*-(5-hydroxypyridin-2-yl)acetamide.

Next, a loop over each such microspecies is performed, successively calling the various flagging procedures with the particular microspecies as argument (several more, including force field-based and lipophilicity-based typing are available, but only the ones actually used in the study will be explained below):

2.1.1 Electrostatic Potential-Based Flagging

First, the ChemAxon's charge calculation plugin[49] calculates all charges based on Gasteiger's partial electronegativity orbital equalization method. The electrostatic potential V_i of each atom *i* are calculated according to:

$$V_i = \sum_{j \neq i} \frac{q_j}{d_{ij}} + \frac{q_i}{d_0}$$

with q_j the partial charge on atom *j*, q_i the partial charge on atom *i*, d_{ij} the topological distance between atom *i* and *j* and d_0 a virtual distance to take into account the concerned atoms charges ($d_0 = 0.4$ after some empirical adjustments, aimed at ensuring that polar positive and polar negative heavy atoms in typical organic compounds were classified in agreement with chemical common sense).

The *V* values are then binned into 5 categories:

- N: negative ($V_i \leq -0.28$),
- n: slightly negative ($-0.32 < V_i \leq -0.08$),
- 0: neutral ($-0.12 < V_i \leq +0.12$),
- p: slightly positive ($+0.08 < V_i \leq +0.32$) and
- P: positive ($V_i \geq +0.28$)

The overlapping bins permit ambiguous cases to be represented by both flags (for example $V_i = -0.3$ would return a "N/n" flag for atom *i*. This is later on taken into account when generating the fingerprints (both possible patterns corresponding to flags N and n for that atom will be generated – see below).

2.2.2 Pharmacophore Flagging

Next, the atomic pharmacophoric types are attributed by the pharmacophore mapping tool PMapper, according to the following custom rules:

- Aromatic atoms are flagged as "R"
- Carriers of positive charges are flagged as "P"
- Centres of negatively charged functional groups are flagged as "N"
- Any oxygen or nitrogen bound to a hydrogen is flagged as "D" (HB donor)
- Any oxygen or nitrogen or negative sulphide or thiourea (=S) is flagged as "A" (HB acceptor)
- Any carbon or halogen except if concerned by the rules above is flagged as "H" (hydrophobe)
- Atoms not matching either of above rules are labeled "F" (featureless)

Note that, like in previously reported work,[34] the pharmacophore flagging rules are specifically applied to automatically generated microspecies in which formal charges had been explicitly assigned to cations and anions, and therefore differ from the ones expected to apply to default neutral representations of compounds.

2.2.3 Typing Output

Eventually, the output SDF is written with the different property-fields corresponding to for each microspecies. After the molecular structure fields, the program will insert as many sets of atom label fields as relevant microspecies were found: for each microspecies *i*, the field "> <POP*i*> " contains the participation, in percent, of that microspecies in the proteolytic equilibrium at specified pH, followed by microspecies-specific label fields "> <**label***i*>" where **label** stands for the respective colouring scheme (pharmacophore types PHPTYP or electrostatic potential flags EPTYP were the only used in this work). Flagging types for atoms are separated by a semicolon, and in case an atom has several features, these are separated by a slash, as exemplified in Figure 3.

2.3 The Fragment Calculator

The Fragment calculator was entirely written in the object-oriented language Free Pascal, in the Lazarus environment. It reads the above-mentioned annotated SD file and finds,

```
....
M  END
> <POP1>
96

> <PHTYP1>
H;H;A;D;R;R;R;R;A/D;R;R

> <EPTYP1>
o;n/o;N;o;n/o;n/o;n/o;n/o;N;n/o;n/o

> <POP2>
4

> <PHTYP2>
H;H;A;D;R;R;R;R;A/N;R;R
....
$$$$
```

Figure 3. Typical output of the typing program.

in each molecule, existing sequences, pairs, augmented atoms or trees of the selected features. Iteratively, for each populated microspecies associated to the current molecule in the input file, the nodes in the molecular graph are coloured according to the features they adopt in the microspecies (see Figure 2) and all possible fragments of the chosen fragmentation scheme are produced and counted. First, all the subgraphs matching the user specifications are detected in the molecular graph, and internally represented under a canonical form, based on the current atom numbering scheme *a, b, c, d…* It is hence avoided that, say, a same sequence "*a-b-c-d*" is being detected and stored twice, both as "*a-b-c-d*" and "*d-c-b-a*". Atom numbers are then (iteratively, if an atom carries more than one flag) replaced by the respective flags, and the resulting coloured fragment label strings are also canonicalized (for example, irrespectively whether *a, b* and *c* are hydrophobes, while *d* is an acceptor, or *a* is the acceptor while *b, c* and *d* are hydrophobes, both subgraphs will count as representatives of the "AHHH" sequence). Canonical representations avoid any spurious duplicate counts of fragments allowing symmetric feature mappings. Fragment population levels are summed up over all microspecies, with increments equal to microspecies population levels. For each coloured fragment, an unambiguous canonical label – a SMILES-like, intuitive "line formula" of the fragment – is associated. The order in which these fragments are detected – and numbered in the IPLF vector describing the molecule – $IPLF_i(M)$ representing the population level of fragment *i* in molecule *M* – depends on the actual series of molecules. In order to force two independent fragmentation runs to respect a common numbering scheme of the fragments occurring in both sets of molecules, a header file (.hdr) generated by the first fragmentation job, listing the found fragments and their original ordering in the IPLF vector, may be passed as an argument for the second fragmentation job. If so, the IPLF vector generated for the second set of molecules will preserve the initial numbering of previously encountered fragments, and may append new ones (never populated in set one) at positions beyond the maximal number of fragments seen in the first set. In this way, the two IPLF files will be directly comparable – their columns being associated to a same fragment – if the first file is formally completed with empty columns corresponding to the new fragments characteristic of the second set. In practice, the descriptor is not stored under the form of the (expectedly sparse) matrix $IPLF_i(M)$, but as a variable-length list of colon-separated pairs of integers $i{:}IPLF_i$, for each populated fragment *i*. Fragments *i* not listed as such on the line associated to molecule *M* are not populated ($IPLF_i = 0$) in that molecule.

2.3.1 Options of the Fragmentation Strategy

Bond information may be included and are represented in the string by:

- "–" for single bonds,
- "=" for double bonds,
- "#" for triple bonds and
- "*" for aromatic bonds.

Size control: An upper (*u*) and lower (*l*) size limit permits the length and range of fragments to be user-defined. The interpretation of these parameters is context-specific: they represent sequence lengths, whereas in conjunction with augmented atoms, they stand for the "radius" of the selection (number of successively considered coordination spheres, see below).

Fuzziness: With sequences and augmented atoms, the user may toggle the fragmentation procedure in "strict" or "fuzzy" mode. Fuzzy mode implies that, alternatively, fragments in which the typing of intermediate atoms is ignored (they flags being replaced by the "?" wildcard) are also explicitly monitored. The "fuzzy" mode does obviously not apply to pairs and trees, since these are nothing but totally fuzzy sequences and respectively augmented atoms in which the nature of all but terminal (and, in trees, central) atoms are ignored.

Therefore, the generic denomination given to a particular set of IPLFs is of the form "*sbwClu*" where *s*∈{seq, pair, aa, tree} stands for the fragmentation strategy {sequences, pairs, augmented atoms, trees}, optionally followed by either "*b*" to signal that bond information is to be included (ignored by default), or "*w*" to signal enabling of fuzzy fragment enumeration, or both *b* and *w*. Eventually *C*∈{SY, PH, EP} stands for the three colouring schemes {symbols, pharmacophore type, electrostatic potentials}. Size parameters *l* and *u* (integers within [0,9]), follow.

Each time a fragment is found, its count is incremented by the population level of the parent microspecies. Molecules are therefore described by a fingerprint of fragment counts accounting for the occurrences in each microspecies and their population level. In case an atom has several flags, all possible combinations of the flags will be considered and they will be incremented by the microspecies population level. Examples of different fragmentation schemes from the *N*-(5-hydroxypyridin-2-yl)acetamide and pharmacophoric colouring are given in Figures 4–6.

Four main classes of fragments are considered:

A) *Sequences*: are series of features of connected atoms. They will be discussed in more detail, for some of the issues encountered at their generation apply to other fragmentation schemes as well. In this context, the size parameters control sequence length. For example, choosing the lower limit at 3 and the upper at 5 will detect all the possible sequences of 3, 4 and 5 atoms in the graph. Each such atom sequence is the source of one or more feature sequences of a same length, (combinatorially) obtained by replacing each atom by each of the features it possesses. For example, a sequence of four hydrophobic carbons maps to a single feature sequence "HHHH" – however, if the first atom in the sequence would instead have been an amphi-

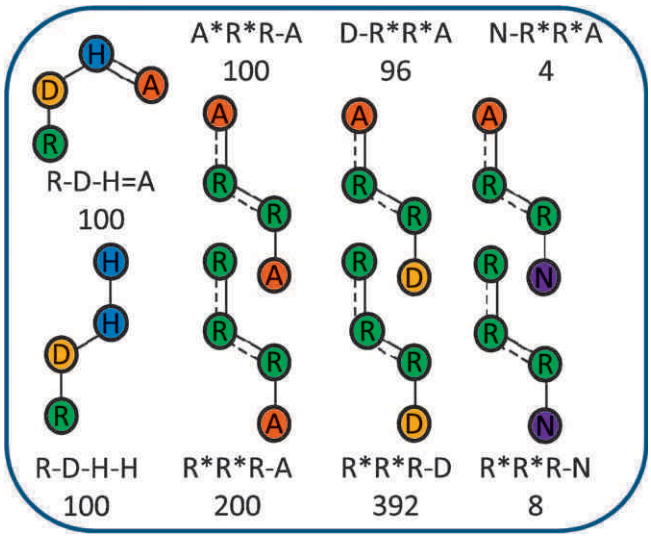


Figure 4. Sequences of length 4 with bond information and their count.

philic hydrogen bond donor and acceptor "A/D", then the same sequence of atoms would have given birth to two feature sequences "AHHH" and "DHHH". As a consequence, the physical atoms in the sequence of four will be counted twice – once as representatives of an acceptor/hydrophobic sequence, then as representatives of a donor/hydrophobic sequence. The total number of coloured fragments in a microspecies therefore largely exceeds the number of actual linear subgraphs, and is closer related to molecular complexity and/or classification uncertainty than to size. If both terminal atoms were of type "A/D", then three different sequences "AHHA", "DHHD" and "AHHD" would be generated, with the latter being counted twice (it also receives population increments associated to the noncanonical representation "DHHHA" which is automatically mapped back to its canonical label). Uncertainty in atom typing is a specific issue of the electrostatic potential typing scheme: potential values close to the border of the main bins are implicit representatives of both bins, therefore contributing to the combinatorial "explosion" of the number of coloured sequences. "Counting" a sequence means incrementing its associated population level by the integer number representing the percentage of participation of the microspecies in which it occurs to the proteolytic equilibrium population at given pH. In Figure 4, the 6 sequences to the right result from "equivalent" atoms but the features are combined differently. As a counting example the sequence R*R*R–D is found 4 times in microspecies 1 and 2 times in microspecies 2 giving the following calculation for the count: $4\times 96+2\times 4=392$.

B) *Pairs*: are disjoined fragments, consisting of any two atoms at a given topological distance separating them. For example, the sequence mentioned above is one of the possible embodiments of the pair represented as R3D, which matches any occurrence of an aromatic at 3 bonds away from a donor. It could have been alternatively represented as a sequence "R??D", with the wildcard "?" stating that the nature of the intermediate atoms does not matter. Pair counts are equal to the sum of counts of embodying sequences.

C) *Augmented atoms*: encode an atom and its environment, i.e. it is a branched fragment centred on a chosen atom and extending to include a user-defined number of successive coordination shells. Their radius (allowed to range between $l=0$, when "augmented" atom descriptors behave like simple atom feature counts, to $u=3$), is the maximum topological distance to the central atom to which the environing atoms are considered. The string representing them is similar to SMILES notation and each branch is canonicalized to obtain a unique representation. In Figure 5 different augmented atoms and their count are shown (bond information is ignored). Cycles are "cut open" to avoid a same atom appear twice in the coordination sphere of a central ring atom.

D) *Trees*: are augmented atoms in which only the core and the leaf features of an augmented atom are given. The intermediate atoms are masked by replacement by the any

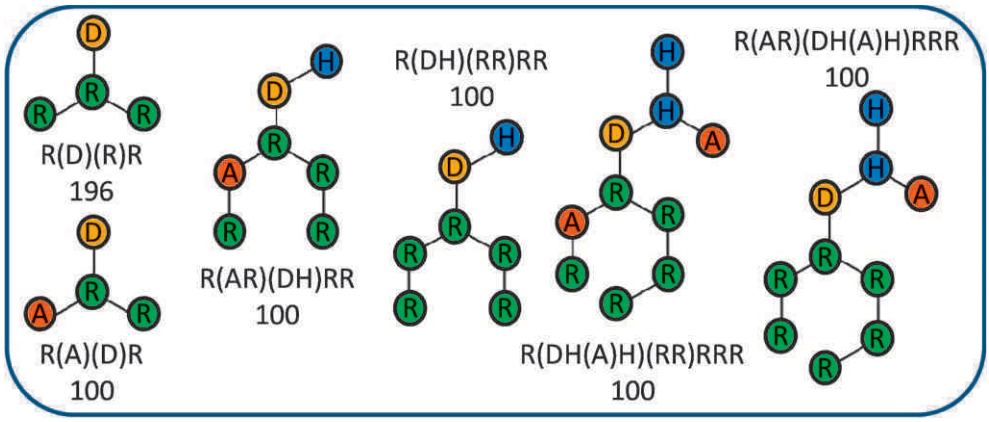


Figure 5. Augmented atoms from radius 1 to 3 centred on the aromatic carbon without bond information and their count.

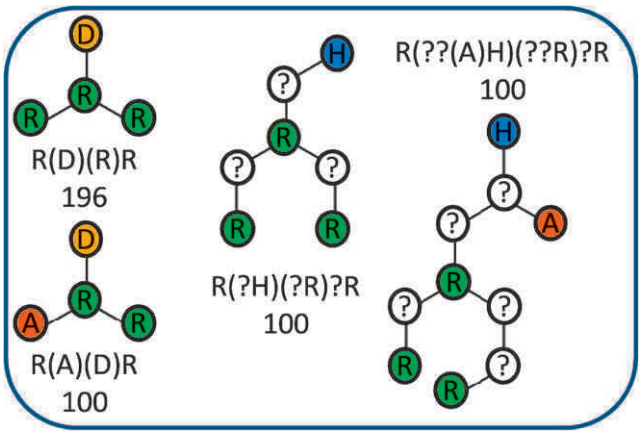


Figure 6. Trees of radius 1 to 3 on the same central atom as for augmented atoms.

feature flag "?". They relate to augmented atoms like pairs with respect to sequences, i.e. represent maximum fuzziness augmented atoms. In Figure 6 the corresponding trees of the augmented atoms in Figure 5 are shown. At radii 0 and 1, augmented atoms and trees are the same as there are no intermediate atoms.

2.4 Benchmarking and Applications of IPLFs

Various types of IPLFs were benchmarked both with respect to their performance in both similarity-based virtual screening (retrieval of active analogues) and QSAR-based molecular property predictions. Descriptor sets including sequences (seq), pairs (pair), augmented atoms (aa) or trees (tree) of different length in combination with atom symbols (SY), pharmacophoric features (PH) or electrostatic potential (EP) flagging were used. The exhaustive list of IPLFs employed in all the benchmarking studies, can be taken from QSAR result Table 1. Please note that terms present in Table 1, but omitted from benchmarking result tables or figures, were not left out of those particular benchmarking studies, but were among the poor performers, not worth mentioning in those contexts.

2.4.1 Neighbourhood Behaviour (NB)

IPLF-based chemical spaces were retrospectively added to an extensive neighbourhood behaviour benchmarking study, already involving more than 50 other classical descriptor sets, in conjunction with three different dissimilarity metrics and two descriptor normalization strategies. The study was based on a consistent data set of a core of 2500 compounds (data courtesy of Prof. Gisbert Schneider and Morphochem AG, Munich) extracted from a combinatorial library, and subjected to binding propensity (pIC_{50}) measurements with respect to 5 different serine proteases: Chymotrypsin, Factor X_a (FXA), Trypsin, Tryptase, Urokinase-type Plasminogen Activator (UPA). In turn, each known active of every protease is considered to be a query compound in a similarity-based virtual screening against the database of remaining 2499 molecules, using the descriptors and a similarity metric.[52] Similar "hits" are selected at the dissimilarity cut-off maximising the Local Ascertained Optimality Score (LAOS), a "noise-free" variant,[42] focusing on individual active queries, of the neighbourhood behaviour optimality score.[9,11,50] It is characteristic of the considered chemical space (defined as a combination of descriptors and dissimilarity calculation rule, or metric), within the specific context (given query compound, considered target) of the virtual screening experiments. For each reported active of every considered target, similarity-based retrieval of active analogues is performed within every considered descriptor space (i.e. every combination of a descriptor set and a similarity metric), and leads to the estimation of the associated LAOS. The LAOS is determined both by the used descriptors and the employed metric. Therefore, a quality factor associated to the descriptors alone (the descriptor "Component Merit" CM, as termed in the original work[42]) is first calculated as a weighted average of all LAOS scores obtained by those descriptors in conjunction to different metrics. Higher LAOS scores count more (weights are taken equal to the LAOS values), to emphasize that a descriptor set is "good" if there exists at least one metric which enables it to achieve high LAOS values. For each active query, a set of descriptor-specific CM values are thus calculated and sorted in decreasing order (bigger CM – better ranking). The sorting allows the conversion of empirical CM scores to rank indices. However, the descriptor set having "won" (ranked #1) the NB contest with respect to an active *A* on a target *T* may perform poorly on other actives *A'* and

Table 1. log*P* prediction accuracy for 9677 compounds.

Descriptors	*RMSE*	R^2_{test}
aabSY02	0.7509	0.8381
aabPH02	0.7801	0.8109
aaSY02	0.7945	0.8188
treePH03	0.8032	0.8148
seqbSY25	0.8033	0.8147
aaPH02	0.8116	0.8109
treeSY03	0.8374	0.7987
seqwPH25	0.8487	0.7932
seqPH25	0.8571	0.7891
seqSY25	0.9062	0.7642
seqwSY25	0.9104	0.7620
seqSY37	0.9796	0.7245
seqPH37	1.0032	0.7110
pairSY28	1.0254	0.6981
pairPH28	1.0576	0.6788
seqbEP25	1.6888	0.1810
treeEP03	1.7538	0.1169
aaEP02	1.8291	0.0393
seqEP37	2.0127	0.0000
seqwEP25	2.6091	0.0000
pairEP28	2.8649	0.0000
seqEP25	3.2551	0.0000

targets *T'*. The average of ranks, witnessed by a descriptor set when browsing through the entire list of all the actives on the five proteases – lower meaning better, ideally equal to 1 for a systematic winner of all virtual screening challenge – illustrates how often chemical spaces using these descriptors did outperform competing spaces. The associated standard deviations of these ranks were also monitored, in order to discriminate between situations in which, at equal average rank performance, some descriptors may do very well for specific queries/targets and perform poorly for others (high variance), whereas others may perform similarly throughout the pool of tests (low variance).

2.4.2 QSAR Models

Two QSAR models: a categorical predictor of the hERG channel blocking risk (discriminating between "blockers" and "nonblockers" at a predefined concentration level of 40 μM) and a quantitative prediction model of the octanol-water partition coefficient log*P*. In both cases, QSAR consensus models were built using the Stochastic QSAR Sampler[53] (SQS). This procedure permits to produce several multilinear regression models (optional nonlinear transformations of the descriptors were not enabled in this study, in order to reduce the model fitting effort). Prediction is made using an average of the predictions of individual models; detailed information is given elsewhere. The applicability domain issue was not considered here, as the aim of these studies is benchmarking – comparison to literature studies referring to global performances with respect to the entire test sets. Reporting to state-of-the-art results from literature obtained on the same, or similar, external prediction sets is in our opinion a better strategy than comparing the herein developed models to equivalent equations based on other sets of descriptors. The latter would be a stricter benchmarking study, but is practically flawed in as far the state-of-the-art descriptors used as baseline terms against the IPLF will be determined by practical constraints (availability of a licence for the corresponding programs, etc.) and might not necessarily be the most appropriate to model the considered properties. Comparison to the work of external expert groups, having intelligently chosen their descriptors in order to maximize the chance to build good models, is a stronger challenge.

With each descriptor set, linear QSAR models were built, using SQS in conjunction with a three-fold cross-validation scheme (model training on 2/3 of the respective sets, followed by validation on the remaining tier, then changing the left-out tier). Eventually, consensus models of all the well-validating equations obtained throughout the cross-validation process were used for prediction of external data sets. Since the focus of this paper is not QSAR modelling per se, model fitting and cross-validation statistics will not be reported – only the performance in the actual predictive challenges will be discussed.

log*P* linear consensus models was trained on a dataset of 3225 molecules and validated on 9677 compounds from the PhysProp[54] database. Root-Mean-Squared prediction Errors (*RMSE*) and associated determination coefficients R^2 with respect to the 9677 external molecules will be reported. Eventually, in order to situate the performance of these models in the context of existing log*P* prediction tools, a second external data set has been analyzed, for which performance of such methods has been extensively reported (284 compounds, out of which – following the same procedures used in the cited log*P* benchmark article[43] – zwitterionic molecules, as well as compounds included in the training set of 3225, were discarded, leaving a rest of 226 molecules for a second predictive challenge).

hERG linear consensus models for different IPLFs were trained on pIC_{50} values for 562 molecules[44] (courtesy T. Oprea, Univ. of New Mexico and O. Taboureau, Technical Univ. Copenhagen). It was then validated on the categorical hERG bioassay data made available on PubChem.[55] The consensus model returned a real-value average of the individual real-value classification score of each individual model, which ranges between 0 and 1. Since hERG blockers are potentially dangerous, failing to recognize actual blockers (false negatives) are a more serious issue than predicting nonblockers to be active (false positives). Therefore, the cut-off for the consensus real-value classification score was empirically set to 0.25 rather than the classical 0.5 (compounds with scores equal or larger than 0.25 were predicted as blockers – class label being rounded up to 1, the rest as nonblockers 0). The herein predicted class was confronted to the PubChem experimental class, in order to calculate specificities, blocker retrieval rates and balanced accuracies of prediction.

3 Results and Discussion

3.1 ISIDA Property-Labelled Fragment Descriptors

Unlike predefined substructural keys such as MACCS, IPLFs may detect and represent novel fragments. They evidence both molecule-specific and common fragments of molecules in a dataset – the variance in structure within this particular dataset – which is very important for the analysis and modelling of SAR.

In the herein presented, unified framework, many reference classes of molecular descriptors are found as particular IPLF subtypes, corresponding to certain combinations of fragmentation schemes and property-labelling. In some cases, the match is quantitative – symbol-labelled fragment counts are, for example, in all respects identical to the widely used ISIDA[26] fragment descriptors. Pharmacophore pair counts, in particular, are extremely popular in chemoinformatics and many more or less different embodiments thereof are known (some of which, such as different version of CATS,[40] and ChemAxon pharmacophore pair counts PF,[33] were players in the NB benchmarking study). The

herein tested "pairPH*lu*" differs from existing ones in terms of its pH-sensitive pharmacophore flagging (and by the user having an actual control on the pair sizes, which is not always the case in other implementations). Original in the current approach is the ability to also generate the conceptually intermediate descriptors between "sequences" and "pairs", by enabling the "fuzzy" handling of nonterminal atoms. While a pair is a maximum fuzziness representation of a sequence being de facto "emptied" of any information concerning intermediate atoms in the path, the herein supported fuzzy rendering scheme considered all the possible "information gaps" within a sequence – from none at all (classical sequence) to full coverage of intermediate atoms (classical pair). Also, the bond information toggle being independent of the fuzziness level, this algorithm may in principle count "pairs" in which the nature of intermediate atoms is ignored, but the orders of the intermediate bonds are explicitly rendered.

There is an obvious analogy between pharmacophore-coloured augmented atoms and the very popular Pipeline Pilot ECFPs,[31] whereas the former nevertheless have two key advantages: the pH-dependent pharmacophore flagging scheme and, arguably more important still, interpretability. H(A)(HP)R(R)R can be straightforwardly understood as a tri-substituted hydrophobe H, bound to (1) an acceptor A, (2) a hydrophobe H, further carrying a cation P, and (3) an aromatic R, further connected to two more aromatics (as expected – aromatic atoms come in rings). By contrast, the rather obscure ECFP numbering scheme requires some quite tedious decoding. The importance of rational, pH-sensitive pharmacophore flagging has been extensively discussed in the context of fuzzy pharmacophoric triplets[34,51] and will not be revisited here. Also, unlike in ECFPs, augmented atom radii are user-defined, including the lower bound, and one can exclude eventually unwanted small fragments.

"Tree" descriptors are original and can be assimilated to property "*n*-uplets" where the size *n* of the monitored property multiplets is not rigidly dictated by the user (as in pharmacophore pairs, triplets, quadruplets) but is free to reflect the actual topological environment of a molecule. If, for example, the root atom of the tree happens to be a terminal atom of a linear chain, that particular tree descriptor will be a simple pair count. When centred on a tri-substituted atom connected to tri-substituted neighbours, at a radius of two, the tree will be a property sextuplet. Pharmacophore trees therefore do not replace rigid-format pharmacophore multiplet counts, but form a super-class of open-ended pharmacophore multiplets. Likewise, fuzzy rendering allows counting of all potential fragments with "wildcard" atoms, thus smoothly bridging the gap between augmented atoms and trees. In principle, this would enable QSAR builders to seek for optimally tuned structural patterns associated to activity – fragments in which some features need to be conserved, while others might be variable, with all this information associated to a single descriptor term. Practically, however, due to the combinatorial explosion of possible fragments – which becomes extremely acute if bond information and/or fuzzy fragment monitoring are enabled – certain IPLFs may span descriptor spaces of huge dimensionality (vectors of $>100\,000$ integers), that may be a real challenge for descriptor selection in predictive model mining (less so, perhaps, in similarity-based virtual screening. Also, "kernel trick"-based approaches such as Support Vector Machines may be tools of choice for data mining in such high-dimensional contexts). Working with "aaw" and "aabw"-type spaces is computationally challenging and was therefore postponed for further study. Furthermore, even if proven feasible, lacking chemical diversity in the training set may be the ultimate reason for which chemoinformatics models based on these terms may prove "too information-rich to be useful" (it may be difficult to find a data set which is chemically rich enough to consistently illustrate the statistical significance of a fragment including both compulsory key features and variable features).

3.2 Descriptor Benchmarking Study

It is obvious that the herein reported benchmarks and applications, albeit extensive and based on robust compound sets of thousands of molecules, only address a very narrow range of the potentially novel opportunities opened by this very versatile molecular description tool.

3.2.1 Neighbourhood Behaviour of IPLFs

The NB benchmarking results, expressed as average ranks and standard deviation bars of each descriptor over the individual NB challenges with respect to every active query of each of the five proteases are summarized in Figure 7. Please refer to the original publication[42] for more detail of the initial set of descriptors used benchmarking – in the following, only the key members among the initial descriptors will be mentioned. Plotting was performed in order of decreasing rank averages, meaning that left-most occurring descriptors outperform the others, on the average. Not only that IPLFs dominate in this chart, but the best actors also have significantly lower standard deviations, meaning that they are systematically found among top performers, all queries and proteases confounded. Unlike classical descriptors, some IPLFs showed general good ranking with a low standard deviation indicating their robustness throughout the different studies. By contrast, the most significant challenger which, strictly speaking, is not entirely a member of the IPLF family is a composite descriptor vector regrouping both selected fuzzy pharmacophore triplets FPT1 and selected ISIDA sequences (identical to seqbSY36 counts according to current notation) which were found to enter a QSAR model of the Tryptase affinity. Or, SEL performs outstandingly well in some NB challenges – but fails in others. Its average rank is the result of an important

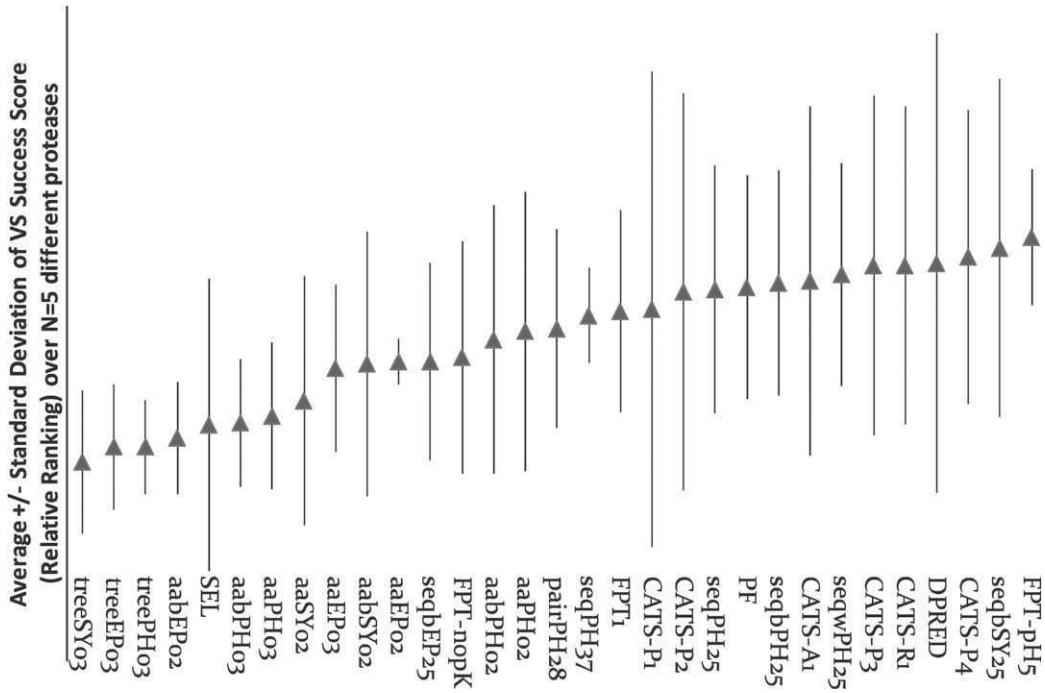


Figure 7. Ranking and standard deviation of different descriptors.

streak of wins with respect to many queries, compensated by a series of serious failures. If partial ranks are computed for each target, SEL would be ranked #1 for UPA, #2 for FXA, #5 for Trypsin, #13 for Tryptase, and #37 for Chymotrypsin). Rank 15 for Tryptase is in itself surprising, since this target served to fit the QSAR model picking members of the SEL descriptor sets out of the extended FPT1 and ISIDA sequence pools. Unsurprisingly, Tryptase ranks the calculated pIC_{50} (DPRED) according to the Tryptase QSAR model as the number one top performer, while positions 2 to 12 are taken by newly conceived lPLFs – none of which could contribute to SEL, for they were not included in the pool of eligible descriptors at the QSAR build-up stage. Compared to SEL, the scorer of the best average rank, treeSY03, is neither ranked #1, but does never fail to make it in the top 20 terms with respect to any of the targets: it occupies positions #2 (Tryptase), #3 (Chymotrypsin), #4 (Trypsin), #14 (UPA) and respectively #18 (FXA).

Tree descriptors manage, irrespectively of colouring scheme, to achieve relatively high and, at the same time, reliable and reproducible performances in similarity-based virtual screening. These feature multiplets appear to be the perfect compromise, providing a structure description at the appropriate detail level. Intriguingly, they are consistently successful throughout the whole series of NB challenges, unlike any of the terms tested in the previous series of simulations, which led to the pessimistic view that robust, universally applicable chemical spaces may not exist, and that each chemical space seems to have its own specific "success queries" and "stumbling stones". Therefore, trees will be descriptors of choice to challenge against other NB tests, involving other targets.

Coherently, augmented atoms follow trees in terms of overall NB – they seem to provide slightly too much structural detail and hence render similarity calculations noisier. Nevertheless, electrostatic potential colouring, reducing the size to two coordination spheres, as well as accounting for bond information, ensures an intriguing fourth place for aabEP02. Unlike in the family of trees of maximal size 3, the aabX02 setup seems to be much more sensitive to the colouring scheme: symbol and pharmacophore-coloured versions significantly lag behind.

3.3 log*P* Study

Lipophilicity, measured by the octanol-water partition coefficient log*P*, is an important physico-chemical property for toxicology, pharmaceutical sciences, environmental research, etc. log*P* is often used to estimate absorption of drugs and is used as a filter in early drug discovery. In silico methods able to correctly assess log*P* of virtual molecules have therefore always been of interest and have been intensively studied.[43,56–58]

Our different models were validated on an external set of 9677 molecules from the PhysProp[54] database. The predictive power of the models is evaluates with the root-mean square error (*RMSE*) and the validation determination coefficient (R^2_{test}) given in Table 1.

Electrostatic potential based models do less well than the atom symbols and pharmacophoric representation

models. This may appear puzzling, since lipophilicity is, after all, a problem related to the overall polarity of the molecules. This relationship is however less straightforward, since it also involves the subtle hydrophobic effect,[59] thought to be of entropic nature, and depending less on the sheer electrostatic potential at the solute-solvent interface, but more on the hydrogen bonding ability. This is not the same: in the EP colouring scheme, many "hydrophobic" carbons and halogens may, due to neighbouring polar groups, be labelled as slightly positive or negative, which may be physicochemically defendable but does not prevent them to act as hydrophobes, in perturbing the bulk water hydrogen bond arrays. This arbitrary split of "hydrophobes" between "*p*", "0" and "*n*", where polarized carbons may end up in a same class as slightly polar, but nevertheless hydrogen-bonding heteroatoms may be the reason for the poor performance of the EP colouring scheme in log*P* modelling.

Augmented atoms and trees seem to perform better than other fragmentation schemes. The best model, based on aabSY02, has an *RMSE* of 0.75 log units, which is well within the state-of-the-art in the field. For example, the ChemAxon log*P* calculator returns an *RMSE* of 0.76 with respect to a subset of 9582 compounds – it fails for the remaining 95 molecules – whereas it cannot be precluded that some of these molecules were in the ChemAxon model training set. By contrast, the experimental accuracy of log*P* (estimated on duplicate entries of a Pfizer proprietary dataset) may be estimated at some 0.35 log units, albeit the one affecting data from the PHYSPROP database, coming from different laboratories, different experimental techniques at different time periods, may be significantly larger. This *RMSE* is difficult to compare to literature-reported values, never based on identical training/validation sets and machine learning techniques (for example, nearest-neighbour-based models actually return the experimental values if the predicted compound is part of the training set, neural nets are supposedly better than linear regression at equal descriptor quality, etc.).

In order to make a meaningful comparison of our results with respect to other methods, the models were challenged to predict log*P* values for a public dataset used for extensive benchmarking, and following the therein reported protocol: after discarding the 22 zwitterions and of further 36 molecules part of our training set, predictions were made for a cured external dataset of 226 molecules. Note that in the cited publication, the ratio of molecules that had to be excluded because involved in calibration was much higher: the external dataset enclosed only 43 molecules. The herein developed log*P* models and different methods benchmarked in the publication are ranked according to their *RMSE* values in Table 2.

The benchmarking results indicate that some of our models outperform or are comparable to other methods, albeit the challenge is biased against our approaches (they are confronted to 5 times as many external molecules, and are simple linear regression models, i.e. their success cannot be attributed to sophisticated machine learning, but must be a consequence of the good quality of descriptors).

3.4 hERG Study

Several cases of noncardiac drugs leading to QT prolongation and sudden deaths have been identified in the mid-1990. hERG channel inhibition has been shown to be the cause.[60,61] The hERG channel is therefore an essential target in the drug discovery and development when evaluating cardiac toxicity. The assay is of considerable cost and therefore, early recognition of potential inhibitors by in silico methods is of notable interest and several models have been published in recent years.

Each model developed on a different descriptor space was externally validated on the PubChem data (1889 molecules with 191 blockers and 1698 nonblockers after duplicate elimination) and their quality was evaluated with the well classified inactive fraction (True negative rate, *TN*, a.k.a. "specificity"), the well classified active fraction (True positive rate, *TP*, a.k.a. "recall") and balanced accuracy (*BA*) summarized in Table 3. Balanced accuracy $BA=(TP+TN)/2$ is a measure to assess how well the model identifies both actives and inactives and corresponds to the average of the true negative and true positive rates in this case.

The best model to classify the actives is pairEP28 ($TP=0.76$ and $BA=0.66$) and second and third best are the models with best balanced accuracy; treePH25 ($TP=0.73$ and $BA=0.68$) and aaSY02 ($TP=0.72$ and $BA=0.68$). All three descriptors are from different properties and different schemes showing no preferences. On the overall one can observe a trend where descriptor spaces issued from electrostatic potentials representation do less well. Except for the pairEP28, they all have low *TP* rates and tend to predict all the molecules as inactives. There is no overall preference for a certain fragmentation scheme.

Li et al. generated hERG potassium channel inhibition models with GRIND descriptors and support vector machines (SVM), using the same sets of data for training and validation.[44]. Their best model was able to achieve a $TP=0.57$, a $TF=0.75$ and a $BA=0.60$. In another recent study,[45] hERG models were fitted on hand of 2644 compounds using ECPF of Daylight atomic invariants rule and functional pharmacophore role with linear discriminant analysis and support vector machines. The external validation on the PubChem data was only able to identify 78 blockers out of 193 corresponding to a *TP* rate of 0.40.

Without taking into account applicability domain and without combining our different descriptors, IPLF with SQS models already outperform models from recent studies.

The hypothesis that the hERG model is successful because it discriminates between hydrophobic binders and polar nonbinders, so actually behaves like a "hidden" log*P* predictor, cannot be discussed in terms of relative profi-

Table 2. Benchmarking of different log*P* methods.

Rank	Method	*RMSE*	Rank	Method	*RMSE*
1	treePH03	0.78	31	pairPH28	1.11
2	aaPH02	0.79	32	*T*log*Pe*	1.12
3	Consensus log*P*	0.80	33	*V*log*P*	1.13
4	aabPH02	0.82	34	pairSY28	1.14
5	Alog*P*Sc	0.82	35	SLIPPER-2002	1.16
6	Milog*P*c	0.86	36	Xlog*P*2c	1.16
7	aabSY02	0.86	37	Quantlog*P*	1.17
8	Sþlog*P*	0.87	38	COSMOFragf	1.23
9	Xlog*P*3c	0.89	39	QikProp	1.24
10	seqbPH25	0.89	40	VEGAg	1.24
11	Clog*P*	0.91	41	LSER	1.26
12	Alog*P*c	0.92	42	*V*log*P*-NOPS	1.39
13	aaSY02	0.92	43	*Q*log*P*	1.42
14	CSlog*P*	0.93	44	CLIPh	1.54
15	Mollog*P*	0.93	45	Mlog*P* (Simþ)	1.56
16	seqbSY25	0.94	46	SPARCi	1.70
17	OsirisPc	0.94	47	NCþNHET	1.71
18	seqPH25	0.95	48	GBlog*P*	1.75
19	seqSY25	0.96	49	aabEP02	1.77
20	seqPH37	0.97	50	treeEP03	1.89
21	treeSY03	0.97	51	aaEP02	1.91
22	seqwSY25	0.97	52	seqEP37	2.09
23	AB/log*P*c	1.00	53	AAM	2.10
24	ACD/log*P*	1.00	54	seqEP25	2.11
25	Alog*P*98	1.00	55	seqbEP25	2.12
26	seqwPH25	1.02	56	Mlog*P*(Dragon)c	2.45
27	seqSY37	1.02	57	seqwEP25	2.51
28	ABSOLV	1.02	58	HINT	2.72
29	KowWINc	1.05	59	LSER UFZ	2.79
30	Vlog*P* OPSd	1.07	60	pairEP28	3.57

Table 3. hERG prediction accuracy for 1889 molecules from PubChem.

Descriptors	*BA*	*TN*	*TP*
aaSY02	0.68	0.65	0.72
seqPH25	0.68	0.62	0.73
treePH03	0.68	0.76	0.59
aaPH03	0.67	0.68	0.67
seqPH37	0.67	0.70	0.63
seqSY25	0.67	0.64	0.69
treeSY03	0.67	0.66	0.67
aaPH02	0.66	0.69	0.63
pairEP28	0.66	0.55	0.76
pairPH28	0.66	0.63	0.68
seqSY37	0.65	0.69	0.60
pairSY28	0.64	0.62	0.66
seqEP37	0.56	0.88	0.25
seqEP25	0.52	0.96	0.08
treeEP03	0.52	0.95	0.09
aaEP02	0.49	0.95	0.04

ciencies of the various descriptor sets. Actually, aabSY02 generate a bad hERG model, not even worth citing in the final result table. However, virtually all top hERG descriptors – and notably aaSY02 – are also top log*P* modellers. The above-mentioned hypothesis may, however, be easily discarded in comparing, for example, the log*P* values for the hERG training molecules, calculated by the aaSY02-based consensus model, to the actual hERG inhibition data, with respect to which they show no correlation at all ($R2 = 0.09$).

The reason for which aabSY02 failed to learn hERG activity is that aabSY02, including bonding information, was too fine-grained for the smaller hERG training set: the too specific instances of bond-sensitive fragments were too rarely encountered to allow meaningful machine learning. By contrast, the much larger and more diverse log*P* set had no problems in supporting successful aabSY02-based machine learning. This nicely illustrate the point that the "proficiency" of descriptors in QSAR should not be understood as an absolute descriptor quality, but rather its ability to match the practical situation and limitations encountered at the knowledge extraction step. The graininess of descriptors must adapt to the amount and quality of experimental data – a hyperfine counting of detail-rich substructures makes no sense if the experimental data does not sustain an analysis at such level. Eventually, the main quality of IPFL is not "QSAR-ability" per se, but the flexibility to adapt to the actual QSAR building context.

4 Conclusions

IPLFs are not conceptually new, but represent a generalization of already existing descriptors and allow to fill in the gaps between what were until now considered to be distinct descriptor categories ("strict" linear and circular fragment counts on one hand, "fuzzy" pair/multiplet counts on the other), whereas a continuous spectrum of possible setups may be envisaged as a "missing link" between these. Coupled to a rigorous pH-sensitive flagging strategy, IPLF configurations which managed to outperform state-of-the-art descriptors – in terms of both good neighbourhood behaviour and QSAR prediction accuracy – were readily identified. IPLF specific-controls allow the user to adapt the fine-graininess of the molecular description to the specific problem: the more training information available, the more useful detailed description levels may become. With the herein considered training sets, all being of an order of magnitude of 10^3 compounds, tree descriptors seemed to offer an optimal compromise and fared consistently well both throughout NB and QSAR tests.

The unified approach provides a clean, homogeneous benchmarking environment in which it will be easy to study the impact of a specific structural aspect, "all other things being equal", while this is not feasible if different descriptor categories have to be obtained from different third-party programs, each coming with specific standardization/flagging rules.

Electrostatic potential colouring appeared as the less interesting option when compared to atom symbol or pharmacophore-based alternatives. This may be partly due to the fact that local electrostatic potentials, furthermore approximated as a function of 2D topological distances, may be less than perfect descriptors of actual local polarity, and furthermore convey little information about nonelectrostatic properties. Alternatively, the potential binning scheme considered here may be faulty – other cut-offs delimiting polar from nonpolar groups might increase EP-based descriptor performance.

References

[1] O. Sperandio, M. A. Miteva, B. O. Villoutreix, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 250 – 258.
[2] H. Koppen, *Curr. Opin. Drug Discovery Dev.* **2009**, *12*, 397 – 407.
[3] J. A. Frearson, I. T. Collie, *Drug Discovery Today* **2009**, *14*, 1150 – 1158.
[4] H. Mauser, W. Guba, *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 365 – 374.
[5] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733 – 1746.
[6] O. A. Santos-Filho, A. Cherkasov, *J. Chem. Inf. Model.* **2008**, *48*, 2054 – 2065.
[7] G. Klebe, *Understanding QSAR: Do we always use the correct structural models to establish affinity correlation?*, http://www.qsar2008.org/home/FA04-10-12-42_h6vpw99c3zxmfq28f4e9/qsar2008.org/public_html/File/abstract%20session%207/Klebe_QSAR_Uppsala_2008.pdf (accessed **2009**).
[8] A. M. Doweyko, *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81 – 89.
[9] G. Papadatos, A. W. J. Cooper, V. Kadirkamanathan, S. J. F. Macdonald, I. M. McLay, S. D. Pickett, J. M. Pritchard, P. Willett, V. J. Gillet, *J. Chem. Inf. Model.* **2009**, *49*, 195 – 208.
[10] H. Eckert, J. Bajorath, *Drug Discovery Today* **2007**, *12*, 225 – 233.
[11] D. Horvath, C. Jeandenans, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680 – 690.
[12] M. A. Johnson, G. M. Maggiora, in *Concepts and Applications of Molecular Similarity*, Wiley Interscience, New York, **1990**.
[13] J. H. Nettles, J. L. Jenkins, C. Williams, A. M. Clark, A. Bender, Z. Deng, J. W. Davies, M. Glick, *J. Mol. Graph.* **2007**, *26*, 622 – 633.
[14] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.
[15] R. Todeschini, M. Lasagni, E. Marengo, *J. Chemom.* **1994**, *8*, 263 – 273.
[16] G. Rucker, C. Rucker, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683 – 695.
[17] X. H. Ma, J. Jia, F. Zhu, Y. Xue, Z. R. Li, Y. Z. Chen, *Comb. Chem. High Throughput Screening* **2009**, *12*, 344 – 357.
[18] R. Vert, J. P. Vert, *J. Machine Learning Res.* **2006**, *7*, 817 – 854.
[19] A. J. Smola, B. Schölkopf, *Stat. Comput.* **2004**, *14*, 199 – 222.
[20] R. Guha, J. H. VanDrie, *J. Chem. Inf. Model.* **2008**, *48*, 646 – 658.
[21] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535 – 1535.
[22] E. Lounkine, J. Batista, J. Bajorath, *Curr. Med. Chem.* **2008**, *15*, 2108 – 2121.
[23] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273 – 1280.
[24] *DayLight Fingerprints and Similarity*, http://www.daylight.com/dayhtml/doc/theory/theory.finger.html.
[25] E. R. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64 – 73.
[26] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. P. Solov'ev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191 – 198.
[27] A. Varnek, D. Fourches, V. Solov'ev, O. Klimchuk, A. Ouadi, I. Billard, *Solvent Extr. Ion Exchange* **2007**, *25*, 433 – 462.
[28] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693 – 703.
[29] V. P. Solov'ev, A. Varnek, *Russ. Chem. Bull.* **2004**, *53*, 1434 – 1445.
[30] V. P. Solov'ev, A. Varnek, G. Wipff, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847 – 858.
[31] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742 – 754.
[32] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed. Engl.* **1999**, *38*, 2894 – 2896.
[33] ChemAxon Screen User Guide. http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Feb. 2009).
[34] F. Bonachera, B. Parent, F. Barbosa, N. Froloff, D. Horvath, *J. Chem. Inf. Model.* **2006**, *46*, 2457 – 2477.
[35] J. S. Mason, A. C. Good, E. J. Martin, *Curr. Pharm. Des.* **2001**, *7*, 567 – 597.
[36] J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, R. F. Labaudiniere, *J. Med. Chem.* **1998**, *38*, 144 – 150.
[37] K. Tsunoyama, A. Amini, M. J. E. Sternberg, S. H. Muggleton, *J. Chem. Inf. Model.* **2008**, *48*, 949 – 957.
[38] O. Sperandio, O. Andrieu, M. A. Miteva, M. Q. Vo, M. Souaille, F. Delfaud, B. O. Villoutreix, *J. Chem. Inf. Model.* **2007**, *47*, 1097 – 1110.
[39] G. Schneider, P. Schneider, S. Renner, *QSAR Comb. Sci.* **2006**, *25*, 1162 – 1171.

[40] Y. Tanrikulu, M. Nietert, U. Scheffer, E. Proschak, K. Grabowski, P. Schneider, M. Weidlich, M. Karas, M. Gobel, G. Schneider, *ChemBioChem* **2007**, *8*, 1932 – 1936.
[41] D. Horvath, B. Mao, *QSAR Comb. Sci.* **2003**, *22*, 498 – 509.
[42] C. Koch, G. Schneider, G. Marcou, A. Varnek, D. Horvath, *J. Computer-Aided Mol. Design* **2010**, submitted.
[43] R. Mannhold, G. I. Poda, C. Ostermann, I. V. Tetko, *J. Pharm. Sci.* **2009**, *98*, 861 – 893.
[44] Q. Li, F. S. Jørgensen, T. I. Oprea, S. Brunak, O. Taboureau, *Mol. Pharm.* **2007**, *5*, 117 – 127.
[45] M. R. Doddareddy, E. C. Klaasse, Shagufta, A. P. Ijzerman, A. Bender, *ChemMedChem* **2010**, *5*, 716 – 29.
[46] *ChemAxon Standardizer*, http://www.chemaxon.com/jchem/doc/user/standardizer.html (accessed Feb. **2009**).
[47] *ChemAxon Pmapper user guide*, http://www.chemaxon.com/jchem/doc/user/PMapper.html (accessed Feb. **2009**).
[48] *ChemAxon pKa Calculator Plugin*, http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html#pka (accessed Feb. **2009**).
[49] *ChemAxon Calculation of Partial Charge Distributions*, http://www.chemaxon.com/marvin/help/calculations/charge.html (accessed Feb. **2009**).
[50] D. Horvath, C. Jeandenans, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691 – 698.
[51] F. Bonachera, D. Horvath, *J. Chem. Inf. Model.* **2008**, *48*, 409 – 425.
[52] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Model.* **1998**, *38*, 983 – 996.
[53] D. Horvath, F. Bonachera, V. P. Solov'ev, C. Gaudin, A. Varnek, *J. Chem. Inf. Model.* **2007**, *47*, 927 – 939.
[54] *SRC PHYSPROP database*, http://www.srcinc.com/what-we-do/product.aspx?id=133&terms=Physprop (accessed Feb. **2009**).
[55] *NIH The PubChem Project*, http://pubchem.ncbi.nlm.nih.gov/.
[56] I. V. Tetko, P. Bruneau, H.-W. Mewes, D. C. Rohrer, G. I. Poda, *Drug Discovery Today* **2006**, *11*, 700 – 707.
[57] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246 – 252.
[58] T. Convard, J.-P. Dubost, H. Le Solleu, E. Kummer, *Quant. Struct.-Act. Relat.* **1994**, *13*, 34 – 37.
[59] N. Choudhury, B. Montgomery-Pettitt, *J. Am. Chem. Soc.* **2007**, *129*, 4847 – 4852.
[60] M. C. Sanguinetti, M. Tristani-Firouzi, *Nature* **2006**, *440*, 463 – 469.
[61] A. M. Brown, *Cell Calcium* **2004**, *35*, 543 – 547.

Received: August 31, 2010
Accepted: November 19, 2010
Published online: December 9, 2010

4.2.1 Neighbourhood behaviour study methodology: additional details

The original NB study to which the IPLF descriptors were subsequently added was published in the Journal of Computer-Aided Molecular Design in 2011[18].

A core subset of 2500 compounds was selected from a combinatorial library of 15,840 compounds by Dr. Lutz Weber and coworkers[19–21] [12, 21, 22] for which inhibitory concentration at 50% (IC_{50}) of enzyme activity had been determined against five serine proteases: Chymotrypsin, Factor Xa (FXA), Trypsin, Tryptase, Urokinase-type Plasminogen Activator (uPA)[21,22].

The NB study on proteases used:

- different descriptor classes: ISIDA "classical" fragments and property-labelled fragments, ChemAxon Pharmacophore Fingerprints (see 2.3.1.4), CATS 2D (see 2.3.1.5), LIQUID descriptors (see 2.3.1.6) and FPT (see 2.3.1.2) descriptors,
- three different dissimilarity metrics: the Euclidean distance, the dice coefficient-base distance and the fraction of differences (see 2.3.4)
- and two scaling strategies: normalisation and standardisation of descriptors (see 2.3.3)

A subset of descriptors selected by SQS (see 2.7.1) on a model of tryptase binding affinity on all 15,840 compounds was also added as a descriptor space and named SEL. The prediction of this model were also used as a mono-dimensional space and named DPRED.

In turn, each known active of every protease is considered to be a query compound in a SimVS against the database of remaining 2499 molecules, using the descriptors, a similarity metric and a scaling strategy. The dissimilarity cut-off is selected in order to maximise the Local Ascertained Optimality Score (LAOS) (see 2.3.5).

In order to compare descriptor spaces, a system inspired from sports' tournaments was invented. Each query of an active molecule M is considered a match between descriptor spaces (DS) which encompasses the used descriptor class, dissimilarity metric and scaling strategy. LAOSs are calculated for each descriptor spaces and a match between two descriptor spaces, DS_1 and DS_2, is won by DS_1 if the difference in LAOS is larger than a chosen threshold, the optimality relevance threshold (ΔOR): $LAOS(DS_1) - LAOS(DS_2) > \Delta OR$. DS_1 wins 3 points for winning the match. If the absolute difference between the LAOS of each DS is equal to or smaller than the optimality relevance threshold, then it is a draw and each DS earns 1 point: $|LAOS(DS_1) - LAOS(DS_2)| \leq \Delta OR$. In the end, a mean of the scores over all the matches is calculated which ranges from 0, for a DS never performing better than any other DS, and 3, for the best DS, always outperforming the others.

This mean, however, does not give an indication whether a descriptor class is proficient or not since the DS is associated with the similarity metric and the scaling method. A quality factor for the descriptor alone, the Component Merit (CM), was defined as a weighted average of all LAOSs obtained by those descriptors in conjunction with different strategies. Higher LAOSs count more (weights are taken equal to the LAOS values), to

emphasize that a descriptor set is "good" if there exists at least one metric which enables it to achieve high LAOS values. For each active query, a set of descriptor-specific CM values are thus calculated and sorted in decreasing order (bigger CM – better ranking). The sorting allows the conversion of empirical CM scores to rank indices. However, the most performing descriptor set (ranked #1) in the NB contest with respect to an active A on a target T may perform poorly on other actives A' and targets T'. The average of ranks for each challenge illustrates how often descriptor spaces outperformed competing spaces. The associated standard deviations of these ranks permit the discrimination between descriptor spaces generally performing well (low variance of rank) and those performing well on specific problems (high variance of rank).

4.2.2 Results with new nomenclature

Figure 7 from the publication has been reproduced in Figure 4.3 for the results on the NB study. Table 1, 2 and 3 from the article have been reproduced in Table 4.2, 4.3 and 4.4 for the studies on LogP and the hERG channel respectively.

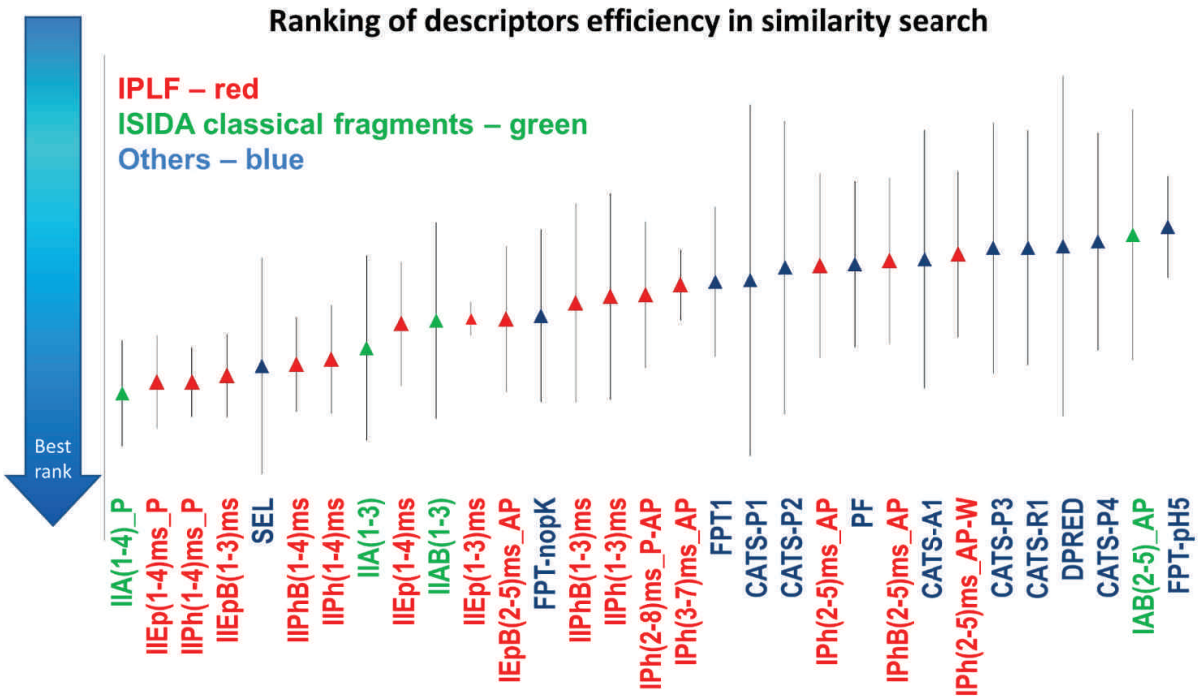


Figure 4.3: Ranking and standard deviation on the first best 30 descriptor spaces

Table 4.2: LogP SQS modelling results

Descriptors	RMSE	R^2_{test}
IIAB(1-3)	0.7509	0.8381
IIPhB(1-3)ms	0.7801	0.8109
IIA(1-3)	0.7945	0.8188
IIPh(1-4)ms	0.8032	0.8148
IAB(2-5)_AP	0.8033	0.8147
IIPh(1-3)ms	0.8116	0.8109
IIA(1-4)	0.8374	0.7987
IPh(2-5)ms_W-AP	0.8487	0.7932
IPh(2-5)ms_AP	0.8571	0.7891
IA(2-5)_AP	0.9062	0.7642
IA(2-5)_W-AP	0.9104	0.7620
IA(2-7)_AP	0.9796	0.7245
IPh(3-7)ms_AP	1.0032	0.7110
IA(2-8)_P-AP	1.0254	0.6981
IPh(2-8)ms_P-AP	1.0576	0.6788
IEpB(2-5)ms_AP	1.6888	0.1810
IIEp(1-4)ms_P	1.7538	0.1169
IIEp(1-3)ms	1.8291	0.0393
IEp(3-7)ms_AP	2.0127	0.0000
IEp(2-5)ms_W-AP	2.6091	0.0000
IEp(2-8)ms_P-AP	2.8649	0.0000
IEp(2-5)ms_AP	3.2551	0.0000

Table 4.3: Benchmarking of different LogP prediction methods

Rank	Method	RMSE	Rank	Method	RMSE
1	IIPh(1-4)ms_P	0.78	31	IPh(2-8)ms_P-AP	1.11
2	IIPh(1-3)ms	0.79	32	TlogPe	1.12
3	Consensus logP	0.80	33	VlogP	1.13
4	IIPhB(1-3)ms	0.82	34	IA(2-8)_P-AP	1.14
5	AlogPSc	0.82	35	SLIPPER-2002	1.16
6	MilogPc	0.86	36	XlogP2c	1.16
7	IIAB(1-3)	0.86	37	QuantlogP	1.17
8	SblogP	0.87	38	COSMOFragf	1.23
9	XlogP3c	0.89	39	QikProp	1.24
10	IPhB(2-5)ms_AP	0.89	40	VEGAg	1.24
11	ClogP	0.91	41	LSER	1.26
12	AlogPc	0.92	42	VlogP-NOPS	1.39
13	IIA(1-3)	0.92	43	QlogP	1.42
14	CSlogP	0.93	44	CLIPh	1.54
15	MollogP	0.93	45	MlogP (Simb)	1.56
16	IAB(2-5)_AP	0.94	46	SPARCi	1.70
17	OsirisPc	0.94	47	NCbNHET	1.71
18	IPh(2-5)ms_AP	0.95	48	GBlogP	1.75
19	IA(2-5)_AP	0.96	49	IIEpB(1-3)ms	1.77
20	IPh(3-7)ms_AP	0.97	50	IIEp(1-4)ms_P	1.89
21	IIA(1-4)ms_P	0.97	51	IIEp(1-3)ms	1.91
22	IA(2-5)_W-AP	0.97	52	IEp(3-7)ms_AP	2.09
23	AB/logPc	1.00	53	AAM	2.10
24	ACD/logP	1.00	54	IEp(2-5)ms_AP	2.11
25	AlogP98	1.00	55	IEpB(2-5)ms_AP	2.12
26	IPh(2-5)ms_W-AP	1.02	56	MlogP(Dragon)c	2.45
27	IA(2-7)_AP	1.02	57	IEp(2-5)ms_W-AP	2.51
28	ABSOLV	1.02	58	HINT	2.72
29	KowWINc	1.05	59	LSER UFZ	2.79
30	VlogP OPSd	1.07	60	IEp(2-8)ms_P-AP	3.57

Table 4.4: hERG QSAR modelling results on the external test set of 1889 molecules from PubChem

Descriptors	BA	TN	TP
IIA(1-3)	0.68	0.65	0.72
IPh(2-5)ms_AP	0.68	0.62	0.73
IIPh(1-4)ms_P	0.68	0.76	0.59
IIPh(1-4)ms	0.67	0.68	0.67
IPh(3-7)ms_AP	0.67	0.70	0.63
IA(2-5)_AP	0.67	0.64	0.69
IIA(1-4)_P	0.67	0.66	0.67
IIPh(1-3)ms	0.66	0.69	0.63
IEp(2-8)ms_P-AP	0.66	0.55	0.76
IPh(2-8)ms_P-AP	0.66	0.63	0.68
IA(3-7)_AP	0.65	0.69	0.60
IA(2-8)_P-AP	0.64	0.62	0.66
IEp(3-7)ms_AP	0.56	0.88	0.25
IEp(2-5)ms_AP	0.52	0.96	0.08
IIEp(1-4)ms_P-AP	0.52	0.95	0.09
IIEp(1-3)ms	0.49	0.95	0.04

4.3 Chromatographic Hydrophobicity Index

The interdisciplinary research project (projet interdisciplinaire de recherche in French, abbreviated as PIR) aimed at annotating the French national chemical library (Chimiothèque Nationale in French, abbreviated as CN) in respect to hydrophobicity, acidity and solubility of molecules. A subset of the CN was selected to represent its diversity and resulted in a set of 640 molecules named the essential chemical library (chimiothèque nationale essentielle, abbreviated CNE). Measurements were performed on the CNE in order to provide data to build QSAR models and annotate the whole library.

In this section, the hydrophobicity part of the PIR project is presented. Hydrophobicity was evaluated using the Chromatographic Hydrophobicity Index (CHI). The measurements done by Patrick Gizzi and the modelling study were the object of a publication in Analytical Chemistry[23] as we were able to trace back experimental errors from the outliers in the models. The publication is included in the following pages. The supporting information of the article are included in the appendices (see B). This publication was aimed at experimentalists as a show-case on how experimental errors can be detected by QSAR. Therefore, details on the final models were omitted and have been added in a section following the publication (see 4.3.1).

Furthermore, the cleaned dataset was used in a small project done on LogP increments with a student. The aim of this project was to develop hydrophobicity-based descriptors using the Ghose-Crippen approach[24,25]. To calibrate and test those, the CHI set seemed to be perfect. It is a clean set issued from the same source and CHI is related to LogP as they are both estimation of the hydrophobicity of compounds. The project in described in section 4.3.2

analytical chemistry

Article

pubs.acs.org/ac

Quantitative Structure–Property Relationship Modeling: A Valuable Support in High-Throughput Screening Quality Control

Fiorella Ruggiu,[†,‡] Patrick Gizzi,[§,▽,‡] Jean-Luc Galzi,[§,▽] Marcel Hibert,[∥,▽] Jacques Haiech,[∥,▽] Igor Baskin,[†,⊥] Dragos Horvath,[†] Gilles Marcou,[†] and Alexandre Varnek*[,†]

[†]Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France

[§]Laboratoire de Biotechnologie et Signalisation Cellulaire (Plate-forme TechMedILL), UMR 7242 CNRS, Ecole Supérieure de Biotechnologie Strasbourg, Université de Strasbourg, 67412 Illkirch Graffenstaden, France

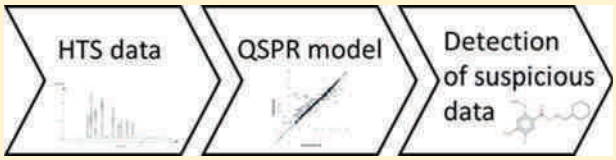
[∥]Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS, Faculté de Pharmacie, Université de Strasbourg, 74 route du Rhin, 67401 Illkirch, France

[⊥]Lomonosov Moscow State University, Moscow 119991, Russia

[▽]Laboratory of Excellence Medalis, IBMC du CNRS, Université de Strasbourg, 15 rue René Descartes, 67000 Strasbourg, France

Supporting Information

ABSTRACT: Evaluation of important pharmacokinetic properties such as hydrophobicity by high-throughput screening (HTS) methods is a major issue in drug discovery. In this paper, we present measurements of the chromatographic hydrophobicity index (CHI) on a subset of the French chemical library Chimiothèque Nationale (CN). The data were used in quantitative structure–property relationship (QSPR) modeling in order to annotate the CN. An algorithm is proposed to detect problematic molecules with large prediction errors, called outliers. In order to find an explanation for these large discrepancies between predicted and experimental values, these compounds were reanalyzed experimentally. As the first selected outliers indeed had experimental problems, including hydrolysis or sheer absence of expected structure, we herewith propose the use of QSPR as a support tool for quality control of screening data and encourage cooperation between experimental and theoretical teams to improve results. The corrected data were used to produce a model, which is freely available on our web server at http://infochim.u-strasbg.fr/webserv/VSEngine.html.



Since the advent of robotized biological testing in the 1990s, access to large, diverse, and original compound collections has become a major issue in drug discovery. However, handling of such collections raises important logistical and technical challenges, in particular because compound originality, a prerequisite for patentability, is by definition not the hallmark of standard, well-conditioned commercial collections accessible to everyone. Extensive analytical assessment of purchased compound collections is therefore a time-consuming and cost-intensive key issue, for its automation may go only as far as automated recording followed by error-prone machine interpretation of analysis results. Time and resources for in-depth structural analysis is lacking; therefore, standard purity measures are necessary but hardly sufficient.[1,2] In standard liquid chromatography/mass spectrometry (LC/MS) analysis, purity is taken as granted if an LC peak of expected mass is "predominant". However, the tacit assumptions that (a) the correct mass actually stands for the expected isomer and (b) the sensitivity of the detector is the same for the main compound and the potential impurities are virtually never checked. In practice, in-depth structural analysis is postponed to the hit reconfirmation stage, for allegedly active molecules only.

In this context, academic compound collections such as the Chimiothèque Nationale (CN), the French national chemical library regrouping original compounds issued from nationwide academic research, is a valuable asset in terms of originality and diversity but a logistical nightmare. Compounds are issued from different laboratories, conditioned according to different operating rules, and stored under variable conditions before being sent to the central repository. The CN therefore requires quality control. A "Projet Interdisciplinaire de Recherche" (PIR) has been conceived as a showcase project to illustrate the use of this collection in (high) throughput screening (HTS) tests and to highlight and fix various pitfalls due to the peculiar nature of this collection. PIR was aimed at annotating the CN with respect to hydrophobicity, solubility, and acidity by using a diverse subset of 640 molecules, named the "Chimiothèque Nationale Essentielle" (CNE), as a representative core of the CN. It was not tailored for drug design and therefore includes reactive and nondruglike molecules as well. The CNE molecules were then cherry-picked and submitted to standard

Received: November 2, 2013
Accepted: January 30, 2014
Published: January 30, 2014

quality control (QC) based on LC/MS purity check at the Integrative Chemical Biology Platform of Strasbourg (PCBIS).

Parallelized and rapid measuring of the envisaged physicochemical properties was carried out at the TechMed^{ILL} Platform in Strasbourg. Hydrophobicity—the first measured property and the one concerned in this paper—is an important property for medicinal chemists.[3] It is widely used as a criterion for acceptable drug solubility and permeability.[4] It has been shown to be related to absorption distribution metabolism excretion/toxicity (ADME/T) properties for over a century.[5] It has classically been evaluated by the octanol−water partition coefficient log $P_{o/w}$ after the proposal of Hansch and Fujita[6] and measured by the shake-flask method. However, this method is time-consuming and a modern HTS method using high-pressure liquid chromatography (HPLC) originally developed by GlaxoSmithKline researchers[7,8] has been used to assess the CNE, the chromatographic hydrophobicity index (CHI).

In reverse-phase HPLC, the partition between a hydro-organic mobile phase and a C-18 stationary phase is governed by hydrophobicity. The organic solvent percentage in mobile phase necessary for elution is referred to as the isocratic chromatographic hydrophobicity index (ICHI), which is thus a good alternative to log $P_{o/w}$ measures.[9] However, this measure requires testing several mobile phases with different organic solvent percentages and thus is time- and resource-consuming. This is why an alternative method based on a fast gradient was developed. The measured retention time in such columns are linearly correlated to ICHI[7] and to log $P_{o/w}$.[8] The method uses a linear calibration generated from the retention times obtained for a set of 10 standard compounds with known ICHI values. For any new compound, the retention time transformed by this calibration gives a number which is referred to as the CHI. This method is cost-effective and very economical in terms of compound requirement and solvent. To conclude, CHI is a measure of retention of the test compound on a fast gradient C-18 column.

It shall be noted that for compounds whose retention is not significant, a negative CHI value will be returned, meaning very low hydrophobicity. For compounds that are not easily washed off the column, a CHI value of >100 is obtained, signifying very high hydrophobicity. But a linear relationship between CHI and ICHI is observed only between 18.4 and 96.4 (the most extreme calibration values). It is important to note that this CHI range covers that of molecules that cross intestinal and brain barriers spontaneously. Molecules with CHI <0 or >100 are not useful in drug discovery programs.

Chemoinformaticians exploited the measured CHI data to build associated quantitative structure−property relationship (QSPR) models on the basis of the CNE diverse training set. The aim was to build useful models in order to annotate all the other academic molecules of the CN by their predicted properties and also to enable chemists to make predictions for novel structures, via a publicly accessible QSPR prediction web server. QSPR models are mathematical models fitted on the data that return an estimate of the expected property on the basis of molecular descriptors serving to numerically encode the features present in the chemical structure. Parameter fitting is done to ensure that, for each training compound (of known property Y), the model will return a predicted Y_{pred} very close to Y (following the classical least-squares principle). The molecular descriptors used in this study are the ISIDA property-labeled fragment counts.[10] Fitting was performed mainly by use of support vector machines (SVM),[11] because of the robustness of the produced models. Other machine learning methods were also tried out.

The main insights gained from this work come from the systematic failures observed in modeling. We define *outliers* as compounds for which their calculated property value Y_{pred} could never be brought in agreement with the observed Y, irrespectively of the employed model-building strategy. This is in line with the classical definition of an outlier as an observation that is numerically distant from the rest of the data.[12] We propose a method for their systematic annotation and then to submit them to in-depth experimental scrutiny. The observed discrepancies between Y and Y_{pred} were much higher than the expected model imprecisions, and yet independent of modeling premises it was hypothesized that this could be due to real differences in molecular structures: thus the actual molecule returning the measured Y might not correspond to the nominal structure for which Y_{pred} was estimated. We identified three periods during which a chemical alteration might have occurred: (a) since the CNE QC, during storage; (b) before the CNE QC, without being detected at that stage; or (c) during the actual hydrophobicity measurement, due to reaction with the aqueous buffer.

Systematic analysis of outliers actually revealed the above hypothesis to be basically correct. This signifies that a properly built QSPR model (with minimized modeling artifacts such as overfitting) is robust enough to highlight experimental errors. Building a QSPR model in parallel to experimental assessment of a library is not a costly undertaking and may effectively pinpoint potential experimental pitfalls, focusing the need for in-depth further analysis to the potentially "pathological" items. This could be an important first step toward the use of QSPR approaches for regulatory purposes, instead of experimental measurements, as envisaged by the REACH project (for registration, evaluation, authorization, and restriction of chemicals).[13]

This paper is organized in order to follow the chronology of the different experimental and modeling steps within the study. First the experimental protocol and results of the CHI measurements is presented, followed by an outline of the computational procedures and the outlier management section. Outlier management contains the initial building of the models, the modeling protocol for the identification of the outliers, their experimental validation, and a presentation of the results with a discussion. Finally, the consensus model, build after removal of outliers and doubtful molecules from the set is presented, followed by a conclusion section.

CHROMATOGRAPHIC HYDROPHOBICITY INDEX MEASUREMENTS

The 640 CNE compounds were received in eight microplates containing 10 mM dimethyl sulfoxide (DMSO) stock solutions. CHI measurements were done on a Gilson HPLC system with a photodiode array detector, an autosampler, and a Valco injector. Data acquisition and processing were performed with Trilution LC V2.0 software. Measurements were carried out at 20 ± 2 °C. A 5 μm Luna C18(2) column (50 × 4.6) purchased from Phenomenex was used. The mobile phase flow rate was 2 mL/min and the following program was applied for the elution: 0−0.2 min, 0% B; 0.2−2.7 min, 0−100% B; 2.7−3.2 min, 100% B; 3.2−3.4 min, 100−0% B; and 3.4−6.1 min, 0% B. Solvent A was 50 mM ammonium acetate (pH 7.4) in water, and solvent B was HPLC-grade acetonitrile (Sigma−Aldrich Chromasolv). The detection wavelengths were 254 and 230 nm.

First, a solution with 10 reference compounds with known ICHI values (see Supporting Information section 1) was injected onto the HPLC to generate a calibration line from their retention times (see Figure 1). The concentration of the

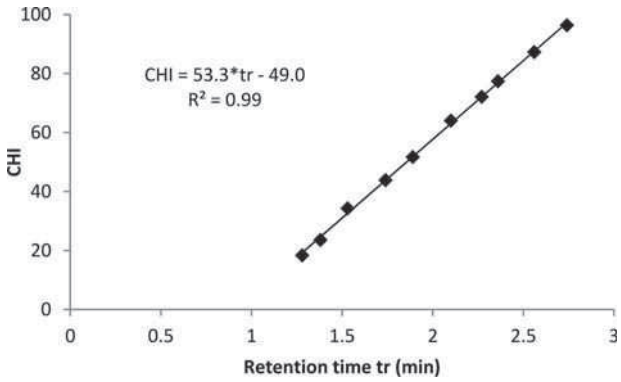


Figure 1. Calibration of the HPLC column: relationship between retention times and CHI values.

mixture was 0.2 mg/mL for each compound and the injected volume was 3 μL. A typical chromatogram of the standard solution is represented in Figure 2. The test compounds were analyzed on the same system. The 10 mM DMSO stock solutions were diluted to 200 μM in acetonitrile/50 mM ammonium acetate, pH 7.4 (1/1 v/v). The linear regression equation of the calibration line was used to convert retention time of the test compounds to CHI values (CHI 1 in Table 1).

The experimental procedure for CHI measurement was applied to all 640 molecules of CNE, and several experimental complications arose (see Figure 3). CHI values of 418 compounds were measured without any complications. The protocol is based on ultraviolet−visible (UV−vis) detection; therefore, compounds lacking chromophore moieties cannot be detected by this method, which was the case for 10% of the molecules. In addition, nothing has been detected for 4% of the molecules for unknown and probably undefined reasons (presumably compound insolubility or instability in DMSO or degradation in test buffer). Several peaks were detected for 36 compounds (6%), indicating impurity or degradation. Hence, matching a peak to the molecule drawn in the database is difficult. It was assumed that the most intense peak corresponds to it. Compounds that gave peaks with low intensity were considered but with caution, because it demonstrates a solubility problem. Finally, CHI values were measured for 545 molecules and complications were annotated in the database.

COMPUTATIONAL PROCEDURE

The computational workflow used in this work is given in Figure 4. Steps 1−5 are described in this section, whereas steps 6−8 are reported under Final Consensus Model.

Compound Standardization. The molecules were standardized by removing salts, stripping off hydrogens from the molecular graph, choosing a standard representation for groups such as nitro or imidazole, and generating major tautomer as well as major microspecies at pH = 7.4 with ChemAxon's Calculator plugin.[14]

Descriptors Calculation. ISIDA property-labeled descriptors,[10] a type of fragment count descriptors, were calculated. Sequences, extended augmented atoms, and triplets were computed on the molecular graph, which has been "colored" with one of the following properties: atomic symbols, pharmacophoric flagging, electrostatic potentials, or force field typing. The length of fragments varied for the minimum from 2 to 4 and for the maximum from 4 to 8. Further variants were then introduced for some of these, by toggling additional options: switching to "Atom pairs" mode, enabling "all path exploration", and the explicit representation of the formal charge. A total of 2772 descriptor pools were eventually generated.

Machine Learning Techniques. SVM was chosen as the reference machine learning because of its stability, mainly due to its particular error function. The Libsvm 3.12 package[11] was used for generation of ε-SVM regression models with a linear kernel, and ε was set equal to the random experimental error estimated at 2 CHI units. The cost was tested for 28 different values ranging from 0.1 to 100. Model building included both operational parameters fitting (as required by the libsvm approach) and, most important, required cross-validation

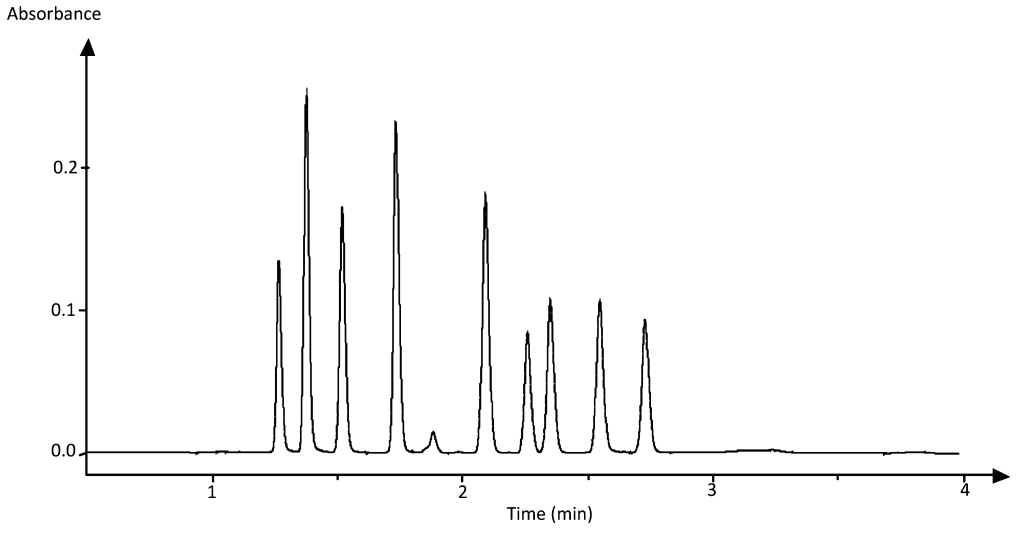


Figure 2. Typical chromatogram of the standard solution.

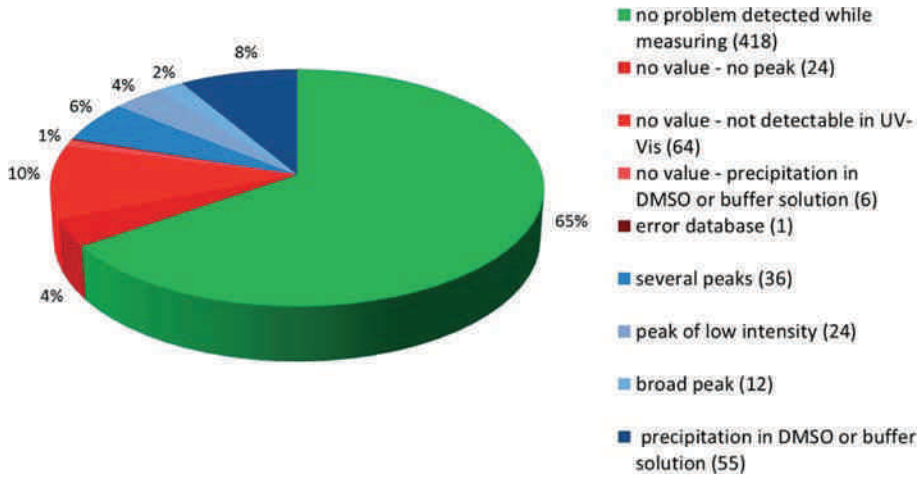


Figure 3. Experimental status of CHI measurements on 640 molecules: green, no problems detected; red, failures to determine the CHI value; and blue, measurements accompanied by observed side phenomena that may signal artifacts, all while nevertheless allowing some CHI value to be recorded.

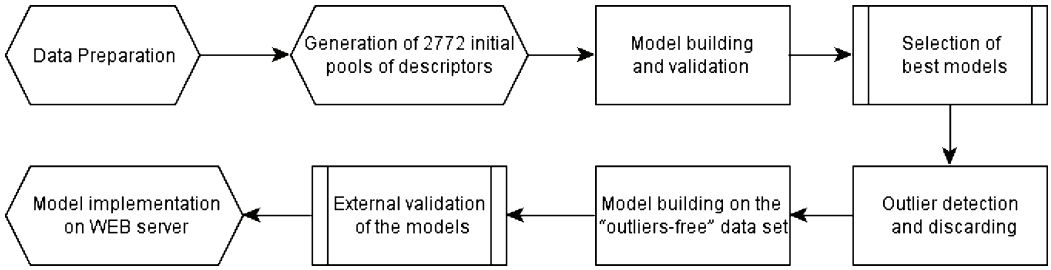


Figure 4. Computational workflow used in this work.

techniques[15] to avoid overfitting. The final model selection criterion therefore was 5-fold cross-validated root-mean-squared error (5CV-RMSE) (see Supporting Information section 2 for details on statistical parameters).

Partial least-squares (PLS) regression models[16] and stochastic quantitative structure−activity relationship (QSAR) sampler (SQS) regression models[17] issued from selected pools of descriptors were also built for comparison purposes.

Model Selection. In total, 2772 × 28 = 77 616 individual models (each corresponding to a particular descriptor pool and a particular value of cost parameter) have been obtained for a given data set. Several "best" models were selected according to 5CV-RMSE. All selected models were used for consensus predictions on the external test set: for each molecule, the CHI value was calculated as an arithmetic average of predictions made by selected individual models.

Outlier Identification Protocol. In this section we discuss the identification of recurrent outliers observed in different modeling strategies. The term "outlier" designates, in the following, a compound for which the predicted value returned by a model having used this molecule for learning strongly diverges from the experimental value.

The list of outliers, submitted to in-depth analysis in order to attempt reconfirmation of these experimental values that could not be explained by modeling, was gathered by an *eliminate-and-refit* protocol on the basis of N best models. At each step of the prediction, a given data point is considered anomalous if its calculation error at the fitting stage is higher than a threshold C_{out}. This threshold is computed as twice the highest 5CV-RMSE found in the set of N values from each SVM model: C_{out} = 2max(5CV-RMSE). The outlier list was iteratively built as follows:

(1) The molecule with the highest number of anomalous estimates is chosen, based on the current value of C_{out}. In the event of a tie, the molecule with the highest absolute mean prediction error is chosen.

(2) The corresponding compound is removed from the modeling data set and the N models are refitted. The operational parameters are not reoptimized.

(3) The experimentally measured CHI value in discrepancy with the prediction is challenged, by a thorough reanalysis of the compound (see Experimental Reassessment of Outliers).

(4) The procedure is repeated from step 1 until no more of the apparently irreconcilable experiment−prediction discrepancies can be attributed to measurement problems (cases a−c listed previously).

The choice of using fitted values is more logical than using 5CV-predicted values as model "output" to compare to the experimental value. Indeed, discrepancies between 5-CV-predicted values and experiment are more likely to occur, especially for species at the edge or outside the applicability domain.[18] If the model has already learned from a molecule, it should be able to predict it. However, if the fitted value of a molecule is in discrepancy with the measured data, this indicates that the molecule goes against what the model learned from other molecules. The stepwise manner of this protocol for picking out outliers instead of selecting several on the same model ensures that the presence of the biggest outlier does not

dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510−2520

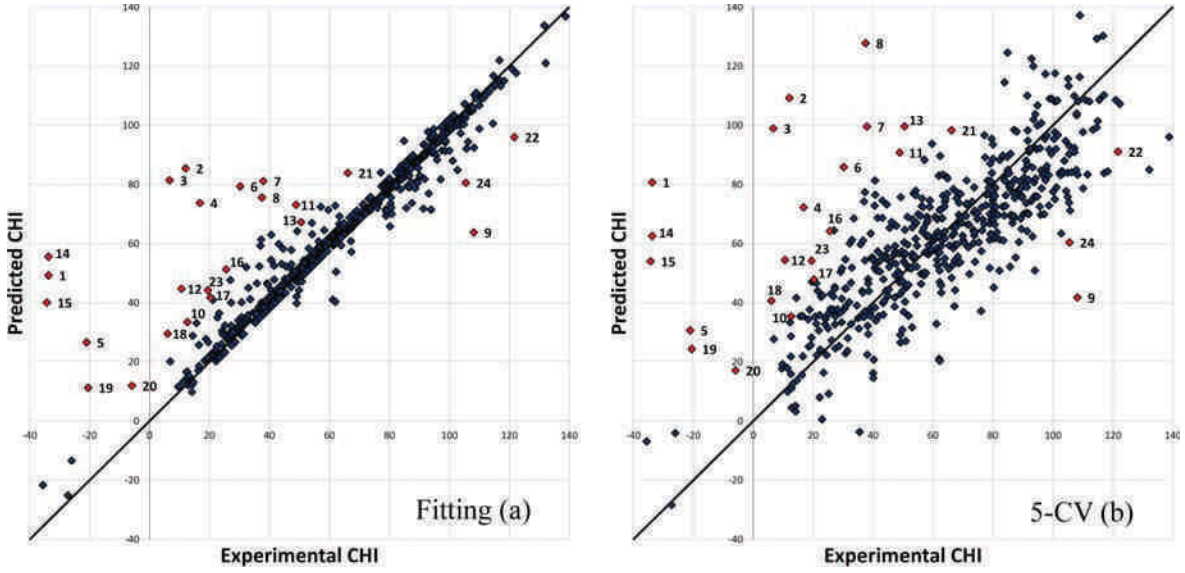


Figure 5. Experimental vs predicted CHI assessed (a) at the fitting stage and (b) in 5-fold cross-validation for the best SVM model (see Outlier Detection, Validation, and Analysis). The numbers indicate the outliers detected in the *eliminate-and-refit* protocol and listed in Table 1.

significantly skew the calculated values for other compounds. When one molecule is eliminated from the training set, the model is refitted and changes. Thus, it cannot be assumed that the molecule with the biggest error on the rebuild model is the same as the second biggest in the initial model. Besides, the fact that a compound appears as outlier for several models is a concept of paramount importance to this analysis because it permits convergence toward problematic molecules identified by different points of views.

OUTLIER DETECTION, VALIDATION, AND ANALYSIS

Outlier Detection. Ten models out of 77 616 built on the parent set of 545 compounds were selected according to 5CV-RMSE. The best of them involves atom-centric fragments colored by atomic symbols with a range of 2−4 atoms, with the use of formal charges, and with a SVM cost of 0.5. It has a train-RMSE of 11.2 and a 5CV-RMSE of 19.6. The obtained models show several recurrent outliers (see Figure 5).

The CNE set is the biggest collection of CHI values found in the literature. It is a very reliable source of data, as it was measured by the same scientist, with the same equipment, in the same conditions (room temperature, solutions used). Thus, the hypothesis that the data cannot be modeled due to multiple protocol incoherencies was discarded. A closer analysis of the structure of those molecules showed that some contained potentially reactive groups, leading us to foresee that problems may concern certain experimentally measured values, even though, in most cases, no peculiar complications were noted during these measurements.

In order to check if the relatively poor model performance is due to inclusion in the training set of molecules for which some experimental complications were detected (blue portion of the chart in Figure 3), modeling was performed on the set of 418 molecules measured without any complications (green portion of the chart in Figure 3). We did not observe any significant improvement of performance, and thus it was expected that reported experimental problems were not indicative of data limiting the quality of the models, as outliers would.

If experimental annotation were not sufficient to discard suspicious data, the question was to which extent are QSPR models able to highlight problems in a set of data issued from an HTS experiment? On the one hand, it is interesting to see how many of those with known experimental problems are perceived as outliers. Are outliers with no apparent experimental problems affected by issues that were not observable during the CHI measurement protocol?

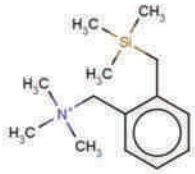
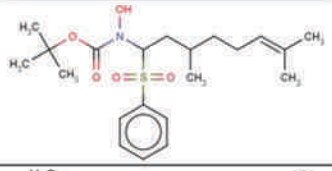
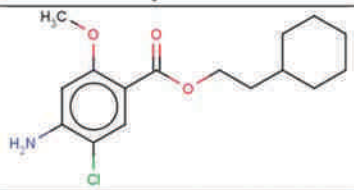
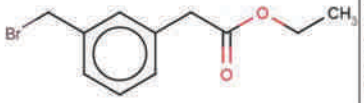
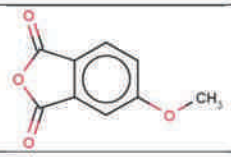
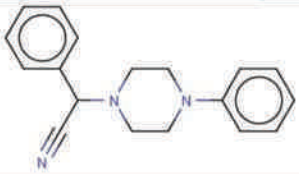
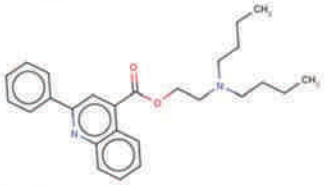
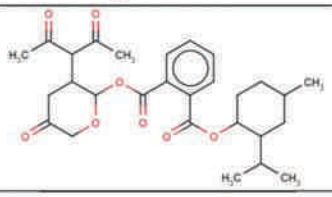
To answer these questions, the *eliminate-and-refit* protocol described under Computational Procedure has been applied for the 10 best SVM models (see the model parameters in Supporting Information section 3). This led to the detection of the 24 outliers listed in Table 1. Unsurprisingly, outliers detected at the fitting stage also behave erratically during 5CV (see Figure 5).

To ensure the outliers did not contain unique features that would make them fundamentally different from the others in the training, 1-SVM[19] using a linear kernel was applied at varying ν parameters. The outlier distribution is homogeneous within the data set. The percentage coverage within the outliers corresponds to the percentage coverage within the data set. If these outliers differed structurally from the other molecules within the set, they would never be within the dense area defined by the 1-SVM.

Experimental Reassessment of Outliers. The experimental check of compounds annotated as outliers was done by the TechMed[ILL] Plateform. CHI values of the compounds identified as outliers were measured a second time (CHI 2 in Table 1) and solutions were submitted to mass spectrometry recharacterization in order to explain differences found between experimental and predicted CHI values. Fresh DMSO stock solutions were prepared from powders except for four compounds for which powder was not available (indicated by asterisks in Table 1). The powder should contain less impurities and eventual chemical degradation is less likely to occur than in the stock solution.

First, these solutions were used to determine the CHI values again by the same procedure explained previously (see Chromatographic Hydrophobicity Index Measurements),

Table 1. Outliers List and Experimental Results[a]

Nb	Theoretical Structure	Comments	CHI 1	CHI_{pred}	CHI 2	MS
1		Desired compound presence is confirmed by MS but this product is not detected by UV. Indeed, for both CHI measurements, only one peak is detected at the void time, which corresponds to a CHI value of -34. The compound not retained by the column is not identified.	-33.7	56.4	-33.7	Y
2		The desired compound is not observed by MS. It is probably insoluble in the buffer. The UV peaks detected for CHI measurements refer to an unknown product.	12.1	88.5	9.6	N
3		The acid resulting from the hydrolysis of the ester is detected by MS. The low value obtained for CHI1 experiment is explained by this hydrolysis. The second CHI measurement with a fresh solution allows detecting the expected ester.	6.7	82.2	108.9	Y
4		The well used for CHI1 measurement contains the desired compound but at a very low concentration confirmed by a small MS response and not detectable by UV. A contaminant with a low hydrophobicity is observed by MS. CHI2 experiment allows detecting the desired compound.	16.8	76.5	86.4 and 80	Y
5*		Although the desired compound is detected by MS with a small response, the major product in the well is the diacid resulting from the hydrolysis of the anhydride.	-21.0	38.4	-24.6	Y
6		The desired compound is detected by MS but as a minor product in the solution used for CHI1 experiment. A contaminant is found. The second measurement CHI2 allows detecting the expected compound as the major product.	30.2	84.7	99.8 and smaller peak at 34.1	Y
7		The desired compound is detected by MS but as a minor product. The UV peaks detected for CHI measurements refer to an unknown product.	38.0	89.4	36.8	Y
8		The desired compound is detected by MS with a very small response. The corresponding concentration is probably not detectable by UV. Two other products are observed. The diacid resulting from the hydrolysis of the esters is detected.	37.5	87.9	33.6 and 58.7	Y

dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510−2520

Table 1. continued

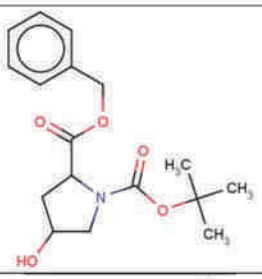
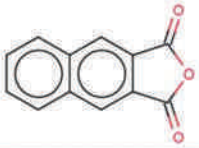
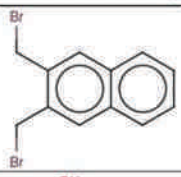
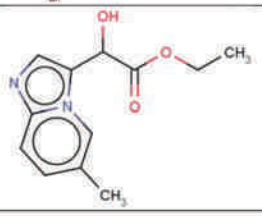
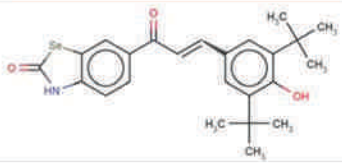
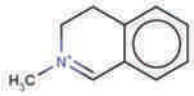
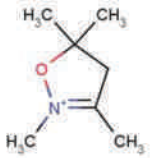
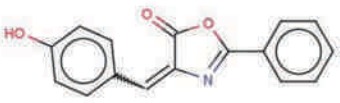
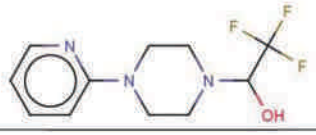
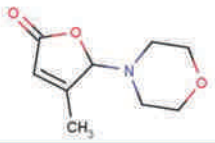
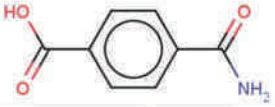
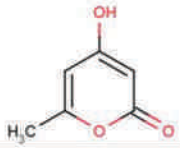
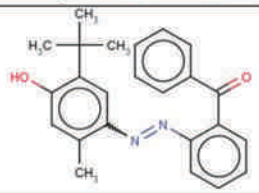
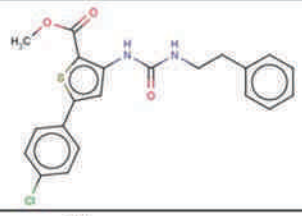
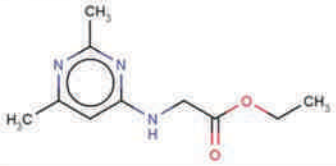
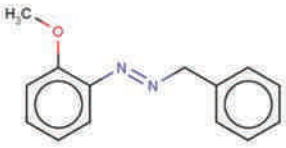
Nb	Theoretical Structure	Comments	CHI 1	CHI_{pred}	CHI 2	MS
9		The desired compound is detected by MS but as a minor product in the solution used for CHI1 experiment. A contaminant is found. The second measurement CHI2 allows detecting the expected compound as the major product.	108.2	67.5	71 and 101.9	Y
10*		Although the desired compound is detected by MS with a small response, the major product in the well is the diacid resulting from the hydrolysis of the acid anhydride.	12.6	53.4	11.2	Y
11		The desired compound is observed by MS but with a very low response. It is probably insoluble in the buffer. The UV peaks detected for CHI measurements refer to an unknown product.	48.9	91.9	49.1	Y
12		The desired compound is detected by MS but as a minor product in the solution used for CHI1 experiment. The acid resulting from the hydrolysis of the ester is detected by MS. The second measurement CHI2 allows detecting the expected compound as the major product.	10.6	47.1	10.1 and 42.1	Y
13		The desired compound's presence is confirmed by MS. The first and the second CHI measurements do not match.	50.5	89.84	114.4	Y
14		The desired compound's presence is confirmed by MS but not detected by UV. Indeed, for both CHI measurements, only one peak is detected at the void time, which corresponds to a CHI value of -34. This compound not retained by the column is not identified. CHI2 experiment allows detecting the desired compound.	-33.7	26.3	-33.7 and 20.2	Y
15*		The presence of the desired compound is confirmed by MS. As it does not contain any chromophore, it cannot be detected by UV. The peak detected at the void time for CHI measurements corresponds to a CHI value of -34. The compound not retained by the column is not identified.	-34.3	18.1	-33.7	Y
16		The desired compound is not observed and the acid resulting from the hydrolysis of the lactone is detected by MS. The low value obtained for CHI1 experiment is explained by this hydrolysis.	25.6	59.1	24.5	N

Table 1. continued

Nb	Theoretical Structure	Comments	CHI 1	CHI_{pred}	CHI 2	MS
17		The desired compound is not detected by MS while the substructure without the $C(CF_3)OH$ is observed.	20.3	49.5	27.1	N
18		The presence of the desired compound is confirmed by MS. The first and the second CHI measurements do not match.	6.1	39.7	-28.9	Y
19		The desired compound is not detected by MS. Both CHI measurements give identical results but do not correspond to the expected product.	-20.5	12.0	-24.6	N
20*		Compound's presence is confirmed by MS but as it does not contain any chromophore, it cannot be detected by UV. The UV peak detected for CHI2 measurement refers to an unknown product.	-5.9	25.1	27.7	Y
21		The presence of the desired compound is confirmed by MS. The first and the second CHI measurement do not match.	66.2	97.7	121.65	Y
22		No problem detected.	121.7	94.0	116.3	Y
23		No problem detected.	19.5	47.1	22.4	Y
24		No problem detected.	105.5	80.0	101.4	Y

[a]CHI 1 is the first CHI value, obtained with DMSO solutions in plates received from the central repository (the whole set was measured with UV−vis detection and used for the first modeling). CHI_{pred} stands for CHI average prediction and corresponds to the average prediction over the 10 best SVM models in the iterative procedure. CHI 2 is the second CHI value, obtained with fresh solutions prepared from powders (except for those marked with asterisks) and measured for the 24 outliers (with LC/UV). MS indicates whether the presence of the theoretical structure was confirmed (Y) or invalidated (N) by mass spectrometry

which permits us to check whether the stock solutions distributed by the CN had problems. Second, a LC/MS characterization was done to confirm or invalidate the presence of the expected compound (see MS column in Table 1), as described by its theoretical structure in the database. Any error in this drawn structure will induce an error in the QSPR estimate, as the descriptors calculated will not correspond to the actual measured structure. A LCMS-8030 triple-quadrupole liquid chromatograph mass spectrometer was used for these quality control measurements. Ionization of compounds was done with an electrospray source. Both single-ion monitoring and scan modes were used. The first mode was applied in order

to control whether the compounds in solution match with the given structures. The second mode allowed identification of other compounds present in the solution, such as impurities or products of degradation. As mass spectrometers do not support high flow rates and high salt concentration in mobile phase, it was impossible to reproduce the same experimental conditions of CHI measurements. Data acquisition and processing were performed with Labsolutions v5.0 software. Measurements were carried out at 25 °C. A 1.7 μm Kinetex C18 column (50 × 2.1) purchased from Phenomenex was used. The mobile phase flow rate was fixed at 0.5 mL/min and the following program was applied for the elution: 0−0.2 min, 0% B; 0.2−3 min, 0−100% B; 3−3.2 min, 100% B; 3.2−3.32 min, 100−0% B and 3.32−6 min, 0% B. Solvent A consisted of 5 mM ammonium acetate in water (pH 7.4), and solvent B was HPLC-grade acetonitrile. Injection volume was 1 μL. The nitrogen nebulizing gas flow was set at 1.5 L/min and the drying gas flow at 15 mL/min. The interface voltage was 4500 V. The temperature of the block heater was maintained at 400 °C and that of the desolvation line at 250 °C.

Table 1 summarizes the results where (i) CHI 1 is the first CHI value, obtained with DMSO solutions in plates received from the central repository (the whole set was measured with UV−vis detection and used for the first modeling); (ii) CHI_{pred} stands for CHI average prediction and corresponds to the average prediction over the 10 best SVM models in the iterative procedure; (iii) CHI 2 is the second CHI value, obtained with fresh solutions prepared from powders (except for those marked with asterisks) and measured for the 24 outliers (with LC/UV); and (iv) MS indicates whether the presence of the theoretical structure was confirmed (Y) or invalidated (N) by mass spectrometry.

Outlier Analysis. The first 21 outliers from the list (see Table 1) were experimentally confirmed to be consequences of various experimental problems and artifacts, many of which escaped direct observation at the initial high-throughput measurement stage. The reassessment was extended to three additional compounds beyond this list of 21 outliers, in order to check the proposed outlier selection criteria.

Identified problems include chemical degradation, which could be identified for six compounds: one lactone (outlier 16), two anhydrides (outliers 5 and 10), and three esters (outlier 3, 8, and 12) were hydrolyzed and the resulting degradation was found in MS. Out of the 21 compounds, only six had an experimental comment indicating eventual measurement complications: three had precipitated in the buffer or in the DMSO stock solution, one had several peaks, one had a large peak, and one had a peak of low intensity. In total, 15 compounds had experimental problems where no measurement complications had been detected.

In order to discuss the results, different compounds have been regrouped into the following categories: hydrolyzed compounds, solutions containing several products, structure not confirmed by MS, no correspondence between the different CHI measurements, and no experimental problems.

Hydrolyzed Compounds: Outliers 3, 5, 8, 10, 12, and 16. In all these cases, the MS spectrum of the hydrolyzed molecule is found, proving the chemical degradation. Such reactions are generally considered as slow[20] at pH = 7.4. However, water impurities may be contained in the DMSO stock solution due to its hygroscopic nature, and thus reaction may occur before the compound is placed in the buffer solution. For outliers 8 and 12, it seems the degradation is fast enough to occur during the second measurement, and thus two peaks are found during the second measurement of CHI. In both cases, it can be assumed that the lowest value corresponds to the acid and the higher value to the drawn structure. In the case of outliers 5 and 10, powder was not available to remake a fresh solution. It seems CHI measurements correspond in both cases to the hydrolyzed compound. In the case of outlier 3, it can be assumed that the first measured value (CHI = 6.7) corresponds to the acid. In the case of the lactone (outlier 16), the compound is not observed and only the hydrolyzed molecule is detected by MS. It can thus be assumed that the CHI values correspond to it.

Solutions Containing Several Products: Outliers 4, 6, 7, 9, 11, 14, 15, and 20. The compounds are detected by MS but with contaminants, indicating a possible degradation or impurity. Outliers 4 and 11 both have benzyl bromides, which may be hydrolyzed[21] or degraded. In the case of outlier 11, the problem is likely related to low solubility of the compound, and hence an impurity is measured in LC/UV−vis with a more intense peak. In the case of outlier 6, the theoretical structure seems to correspond to the CHI value of 99.8. In the case of outlier 15, the expected compound is confirmed by LC/MS but has no chromophore to be detected in LC/UV−vis. Thus, the measured CHI value probably corresponds to an impurity or a counterion coming out at the void time.

Theoretical Structure Not Confirmed by MS: Outliers 1, 2, 17, and 19. The compounds are not present during the experiment. It is impossible to conclude what may have happened and what is actually measured during the LC/UV experiment with the given information. Possibly, the compound was not soluble or the given powder did not contain the indicated compound due to a human error. In the case of outlier 17, a substructure of the theoretical structure is found in MS. This could have been an input or synthesis error. In the case of outlier 2, the absence may be related to the low solubility of the compound (measured as 2 μM in pH 7.4 buffer).

No Correspondence between Different CHI Measurements: Outliers 13, 18, and 21. The compounds are identified by MS, but no matching of the CHI values can be found and no other compounds are detected. Possibly some wells in the given microplates may have contained a wrong solution in the first measurement or the compounds were degraded during storage and these reactions are not fast enough to be observed during the second measurement, when the stock solutions are redone. In the case of outliers 13 and 21, the predicted values are qualitatively in better accord with the second measurements. In the case of outlier 18, it is questionable whether the compound is not hydrolyzed or degraded.

No Experimental Problems: Outliers 22−24. The compounds are detected in the expected ranges of retention times by LC/MS and both CHI measurements match. It seems these molecules are not well predicted and the discrepancy may originate from the limits of the modeling. We note that the outliers 22 and 24 are above the highest calibration value (valerophenone, CHI = 96.4).

Extreme Values of CHI. CHI is derived from the ICHI, which corresponds to the percentage of acetonitrile needed to achieve an equal distribution between the two phases. It is calibrated on a set of compounds for which the ICHI is known and the ICHI is effectively bounded between 0 and 100.

However, as the CHI is a retention time converted to an ICHI scale, it can have values outside the range 0−100.

Several outliers confirmed to have experimental problems have a negative value and it was observed that their CHI corresponds to the void time of the column; thus no actual measurement of the molecule's hydrophobicity is achieved. It can only be concluded that these have very low hydrophobicity. In the remaining molecules of the database, three such cases with values below 0 are found (structures are provided in Supporting Information section 4) and were thus discarded from the final modeling data set.

The 57 cases above 100 CHI units have been kept (excluding outlier 9), as these CHI value convey physicochemical meaningful differences between the compounds. Indeed, a retention time can be unambiguously measured: no metrological problem is expected. For this range of CHI, it can be assumed that a compound with a lower CHI than another has indeed a lower hydrophobicity. However, the assumption of a linear relationship between isocratic chromatographic hydrophobicity index and log D[8] is obviously wrong.

Outlier Dependence on the Modeling Protocol. The sensitivity of the outlier list with respect to the machine learning technique was assessed by ranking compounds according to the average errors reported by alternative PLS regression models obtained with Weka 3.7.6[16] and respective SQS[17] models. The PLS models were generated with varying number of components from 2 to 20 with a step of 2. SQS models were built on eight descriptor spaces known for their good predictive proficiency in SVM fitting. The 10 PLS models used were selected on the criteria of equivalent statistics to best model, low number of components, and different types of descriptors. The eliminate-and-refit approach was also used on PLS.

The other machine learning methods are also able to find most of these outliers, picked on the basis of SVM models. These were primarily run to cross-check whether outlier detection would be strongly impacted by the choice of machine learning protocols. This is not the case. The outlier lists obtained by use of PLS or SQS were largely consistent with the one obtained with SVM.

FINAL CONSENSUS MODEL

The compounds experimentally confirmed to have problems (21 compounds, see Table 1), compounds with CHI values below 0 (3 compounds), and all compounds with several peaks (36 compounds) were removed from the initial set. The "cleaned" data set of 485 compounds has been used to rebuild SVM models, re-exploring descriptor spaces and parameters. An external 5CV procedure was applied by splitting the initial set of molecules five times into five different folds. Best models were selected on the criterion of a 5CV RMSE better than a cutoff of 16. Only one model per descriptor space was kept. A y-randomization strategy[22] performed 20 times confirmed the significance of the selected models. In total, 81 models with 5CV-RMSE ranging from 14.5 to 16 are included in the consensus model (see Supporting Information section 7 for details).

It was observed that the best descriptor spaces were covering small fragments. The best descriptor space is an atom-centric fragmentation colored by atomic symbols with a range of 2−3 atoms and the use of formal charges. This might be related to the diversity of the molecules, which do not allow the extraction of more complex description, or to the additive character of hydrophobicity.[23]

An external test set of 195 molecules from the literature[7,8,24−26] was used to evaluate the generalization of the consensus model. Care was taken to have the most similar experimental conditions: (i) The pH varies from 7 to 7.4. (ii) A reversed-phase C18 column with a gradient of acetonitrile/buffered water was used in all cases. (iii) Calibration was slightly different in two cases;[7,26] hence, an equation was established to convert the values. (iv) Compounds were detected by UV−vis in most cases and by mass spectroscopy[25] for six molecules.

The model performs reasonably on the external test set with a RMSE of 16.4 and a determination coefficient R^2_{det} of 0.6 (see Supporting Information section 5 for details). It is not surprising to obtain worse results on the external test set than expected from cross-validation experiments. The main difference is that the former data set is issued from the literature whereas the latter is issued from the same laboratory. For data coming from literature, it is not possible to exclude some variation in the experimental setup, the least of it being that the calibration parameters of the CHI vary from one article to the other. The compounds measured by MS also notably differ from the other errors (see Supporting Information section 5 for details).

CONCLUSION

To conclude, we suggest the use of QSPR modeling to control the quality of HTS experiments. In this paper, we present the largest homogeneous data set of experimentally measured CHI values. We also propose an algorithm to list, on the basis of QSPR modeling, outliers that are likely to represent cases of severe and hidden experimental error. With this algorithm, we were able to pinpoint experimental problems for 21 compounds. These problems could not be detected during the experimental screening and they represented about 4% of the database. The final model was produced from reliable data and is publically available. The model was used to annotate the whole CN.

It is our belief that removal of outliers should not be done automatically (typical strategy in QSAR/QSPR) and outliers should bring chemists to reflect on their work. Their proper analysis demands a synergy between experimental screening teams and chemoinformatics modeling teams. The cost of a QSPR study is negligible compared to a screening campaign. The discrepancies observed between QSPR estimates and screening results are useful to detect experimental problems otherwise invisible. Such interplay could be a useful addition to regulatory tests such as those mentioned in REACH.

ASSOCIATED CONTENT

Supporting Information

Additional text, four tables, and two figures describing (1) calibration compounds and their associated ICHI values, (2) statistical parameter definitions, (3) parameters of 10 SVM models used for eliminate-and-refit protocol to detect outliers, (4) structures and CHI values of three compounds below 0 after removal of outliers, (5) experimental versus predicted value of CHI on external test set for the consensus model, (6) availability of the model for end users (http://infochim.u-strasbg.fr/webserv/VSEngine.html), and (7) descriptor spaces, ISIDA Fragmentor2012 options, libsvm options, and statistics of the 81 models used in the consensus model; and a listing of

dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510−2520

the CHI training set containing all measured values. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail varnek@unistra.fr.

Author Contributions

‡F.R. and P.G. contributed equally to the work

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the CNRS PIR project for financial support and the Centre of High-Performance Computing, Informatics Department, University of Strasbourg (France), for computational facilities.

REFERENCES

(1) Yan, B.; Fang, L.; Irving, M.; Zhang, S.; Boldi, A. M.; Woolard, F.; Johnson, C. R.; Kshirsagar, T.; Figliozzi, G. M.; Krueger, C. A.; Collins, N. *J. Comb. Chem.* **2003**, *5* (5), 547−559.

(2) Lemoff, A.; Yan, B. *J. Comb. Chem.* **2008**, *10* (5), 746−751.

(3) Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. *Bioorg. Med. Chem.* **2004**, *12*, 3391−3400.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(5) Meyer, H. *Arch. Exp. Pathol. Pharmakol.* **1899**, *42*, 109−118.

(6) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616−1626.

(7) Valkò, K.; Bevan, C.; Reynolds, D. *Anal. Chem.* **1997**, *69* (11), 2022−2029.

(8) Valkò, K.; Du, C. M.; Bevan, C.; Reynolds, D. P.; Abraham, M. H. *Curr. Med. Chem.* **2001**, *8* (9), 1137−1146.

(9) Valkò, K.; Slégel, P. *J. Chromatogr.* **1993**, *631* (1-2), 49−61.

(10) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. *Mol. Inf.* **2010**, *29* (12), 855−868.

(11) Chang, C. C.; Lin, C.-J. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 27:1−27:27.

(12) Barnett, V.; Lewis, T. *Outliers in Statistical Data*, 3rd ed.; John Wiley & Sons.: New York, 1994.

(13) Ahlers, J.; Stock, F.; Werschkun, B. *Environ. Sci. Pollut. Res.* **2008**, *15*, 565−572.

(14) ChemAxon JChem, Calculator plugin. http://www.chemaxon.com.

(15) Dietterich, T. G. *Neural Comput.* **1998**, *10* (7), 1895−1923.

(16) Hall, M.; Eibe, F.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *ACM SIGKDD Explor. Newsletter* **2009**, *11* (1), 10−18.

(17) Horvath, D.; Bonachera, F.; Solov'ev, F.; Gaudin, C.; Varnek, A. *J. Chem. Inf. Model.* **2007**, *47* (3), 927−939.

(18) Weaver, S.; Gleeson, M. P. *J. Mol. Graphics Modell.* **2008**, *26* (8), 1315−1326.

(19) Baskin, I. I.; Kireeva, N.; Varnek, A. *Mol. Inf.* **2010**, *29* (8−9), 581−587.

(20) Clayden, J.; Greeves, N.; Warren, S.; Wothers, P., *Organic Chemistry*, 1st ed.; Oxford University Press: Oxford, U.K., 2001.

(21) Vitullo, V. P.; Sridharan, S.; Johnson, L. P. *J. Am. Chem. Soc.* **1979**, *101* (9), 2320−2322.

(22) Rücker, C.; Rücker, G.; Meringer, M. *J. Chem. Inf. Model.* **2007**, *47* (6), 2345−2357.

(23) Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7* (4), 565−577.

(24) Plassa, M.; Valkò, K.; Abraham, M. H. *J. Chromatogr. A* **1998**, *803* (1−2), 51−60.

(25) Camurri, G.; Zaramella, A. *Anal. Chem.* **2001**, *73* (15), 3716−3722.

(26) Fuguet, E.; Ràfols, C.; Bosch, E.; Rosés, M. *J. Chromatogr. A* **2007**, *1173*, 110−119.

4.3.1 Final consensus model: additional details

Table 4.5: Final consensus model performances on the external test set of 195 molecules

Applicability domain	Number of molecules	RMSE	MAE	R^2_{det}	R^2_{corr}
with at least 25 models	93	14.3	11.3	0.7	0.9
without	195	16.4	12.6	0.6	0.8

Results of the final consensus model with an AD using fragment control and bounding box with at least 25 models and without the AD are presented in Table 4.5 and Figures 4.4 and 4.5. The consensus model performs well on this external test set and are in good agreement with the 5-CV results (best models with 5-CV RMSE of 14.5 and R^2_{det} of 0.7) when considering that the test set data has slightly different experimental setups. The data points in Figures 4.4 and 4.5 are coloured according to the source and method used for measurements:

- Valko1997[6] (9 compounds, squares in dark blue): CHI was measured at pH 7.3 instead of 7.4. CHI values were converted because different calibration values were used with the same compounds.
- Plass1998[26] (14 compounds, triangles in light blue): CHI values were converted with a linear regression because different calibration values were used with the same compounds. This publication measures CHI for peptides and these may have different values for L and D conformers; a mean of the values was used if the difference was lower than 3, otherwise they were left out of the test set.
- Camurri2001 with LC-UV[27] (28 compounds, triangles in green)
- Camurri2001 with LC-MS[27] (6 compounds, squares in red): The names were used to convert to structures, however, the chemical formula indicated in the publication does not correspond. Predictions for these molecules are rather unsure.
- Valko2001[7] (127 compounds, circles in yellow)
- Fuguet2007[28] (11 compounds, diamond-shaped in violet): CHI was measured at pH 7 instead of 7.4. CHI values were converted because different calibration values were used with the same compounds.

The data points measured by LC-MS from Camurri et al. are noticeably less well predicted. This could be due to the change of method but also to the bulkiness of these molecules for which effects such as forming a pellet on itself is unknown to the model because it was trained on smaller molecules. These molecules are discarded when the AD is applied.

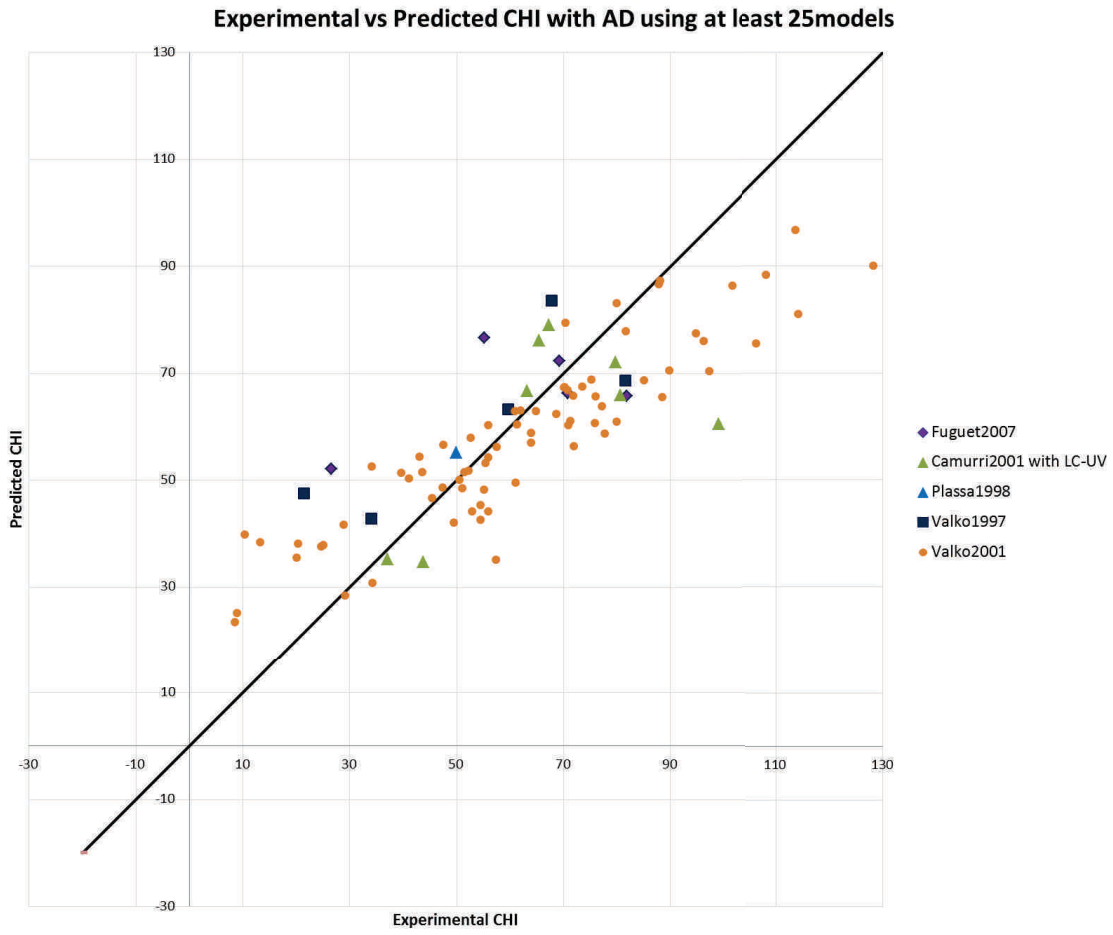


Figure 4.4: External test set prediction by the final consensus model using at least 25 models in the consensus prediction for which the molecule is within AD

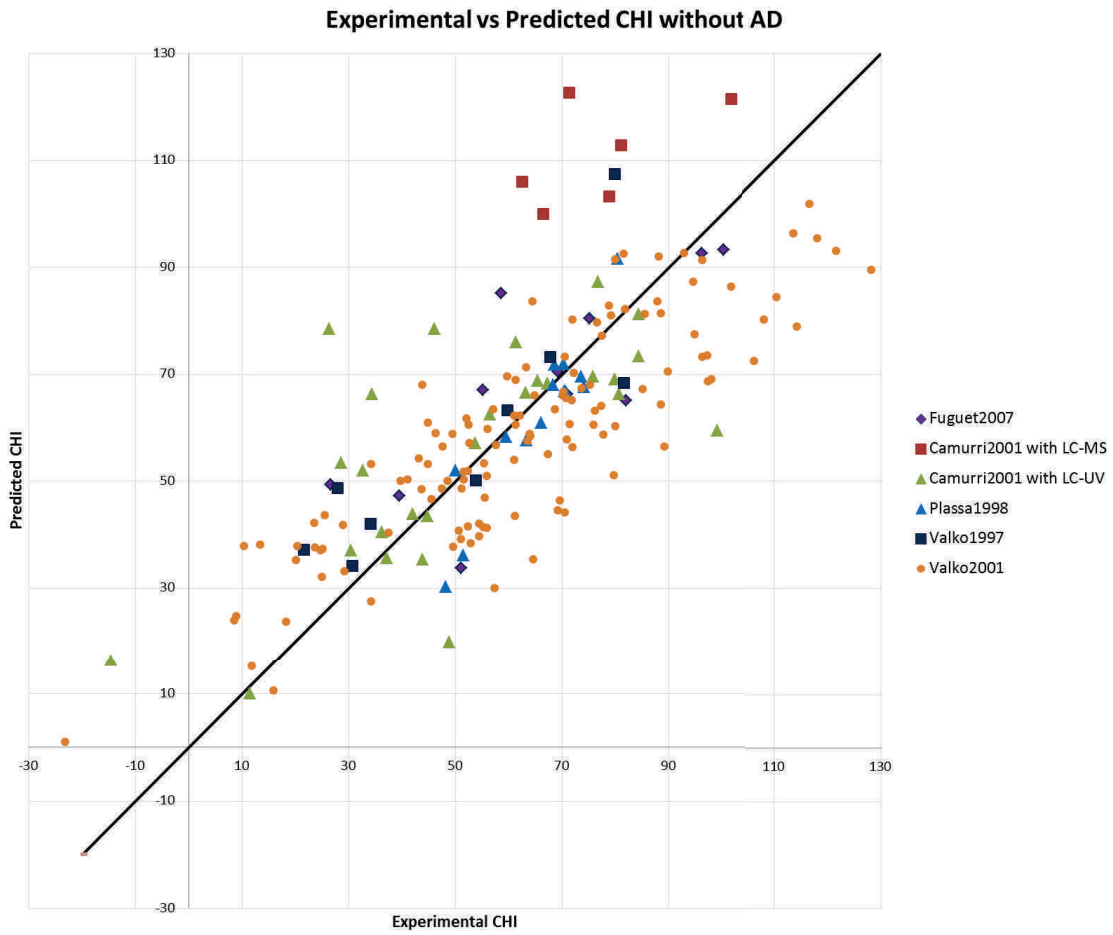


Figure 4.5: External test set prediction by the final consensus model using all 81 models without AD

4.3.2 CHI modelling with LogP increments mapping

4.3.2.1 Introduction

A students' project done with Julien Denos[29] aimed at incorporating LogP increments as a new colouration for the graph prior to the fragmentation. The challenge of this project was to establish an empirical dictionary attributing the atomic colour labels needed in descriptor build-up as a function of the atomic hydrophobicity indexes. Because the number of different colours (i.e. hydrophobicity categories) must be limited (for example "very polar", "polar", "neutral","hydrophobic", "very hydrophobic") colours will correspond to arbitrarily defined hydrophobicity index ranges. In order to pick these colouring ranges, various approaches have been explored. It was tested on the final CHI set from the study presented in the previous section. This data permits a clean benchmark because it comes from the same source and experimental conditions are known. Individual models using LogP increments-based coloured descriptors were built and compared to previously obtained models and the best were added to the consensus model to see if the performance is increased. If the latter is true, then the LogP increments-based colouration permits to obtain a highly relevant point of view of the molecules and thereby, increase the robustness of the consensus model.

4.3.2.2 Method

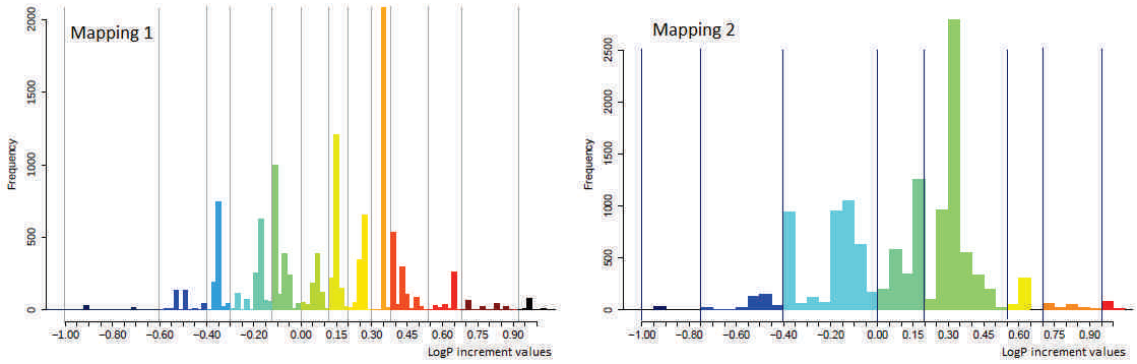


Figure 4.6: LogP increments binning schemes using histograms

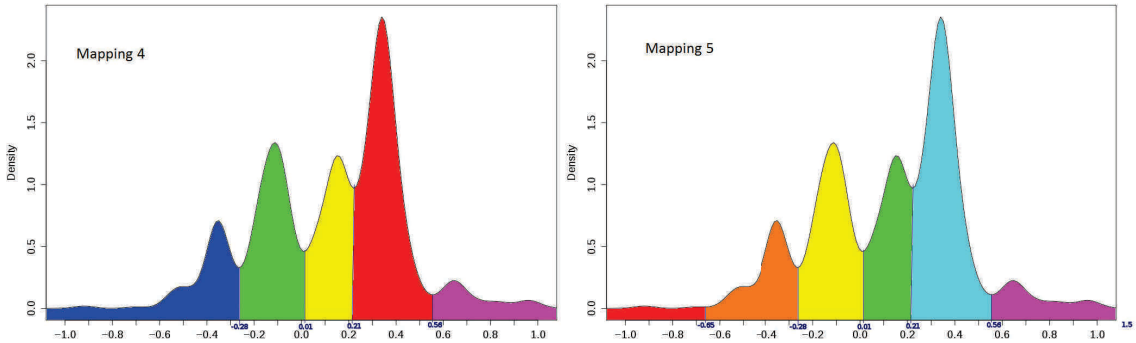


Figure 4.7: LogP increments binning schemes using kernel density estimation

Data The final training set issued from the CN and the literature test set of CHI values were standardised using our in-house script (see 7.1) which uses ChemAxon's calculation plugin[30] with major tautomers and micro-species at pH 7.4.

Descriptors The Ghose-Crippen LogP increments were calculated with ChemAxon's calculation plugin[30] on the CNE set. The distribution of the increments was visualised with histograms using R[31] and different bin widths to divide the data. A binning boundary was fixed at 0 as it indicates a change of the atom's hydrophobic character. From these histograms, two different binnings schemes were chosen to apply the mapping of the LogP increments (see Figure 4.6). The first mapping (Lp1) was done at a resolution of 1/300 of the LogP increment range length to divide the histogram and the second (Lp2) at a resolution of 1/100 of the range. The third binning scheme (Lp3) was elaborated from the second mapping. Density kernel estimation[32,33] was applied using R and the vertexes of the distribution were made to correspond to the second mapping. The curves of each binning were prolonged to intersect with the x-axis (increments values). A fuzzy binning was introduced so that the borderline increments are included into both bins. Two further mappings (Lp4 and Lp5) were defined using kernel density estimation at the local minima of between the summits as shown in Figure 4.7.

ISIDA property-labelled descriptors were calculated by varying systematically:

- the fragmentation schemes: sequences, atom-centred fragments and triplets.
- the 5 LogP increment mapping schemes mentioned above.
- the minimum length between 2 and 4 and the maximum length between the minimum and 4. Atoms counts were systematically added.
- the toogle indicating formal charges.
- the toogle for the all path exploration.

Machine Learning Models were built on each descriptor space using ϵ-SVM from the LibSVM package[34]. The same parameters used previously to build the models from the consensus model were used: a linear kernel, an ϵ set at 2 and the exploration of the cost parameter between 0.1 and 100.

Validation A 5-fold cross-validation was performed 10 times and the RMSE was chosen as a parameter for comparison to the best models obtained for each mapping schemes including pharmacophoric properties and atom symbols calculated previously. A Student t-test at a 95% confidence level was applied to verify whether the resulting predictions in 5-CV of the different mappings are significantly different from each other pairwise. Models with a 5-CV RMSE below 16 were added to the consensus model and the external test set of 195 molecules was predicted.

4.3.2.3 Results and Discussion

First of all, it should be noted that the choice of 0 as a boundary seems good when looking at the distributions as it is located in a low density area. Performances of the best models for each mapping are presented in Table 4.6 and the Student's t-test results are shown in Figure 4.8. Preferred descriptor spaces use atom-centred fragments. The model using atom symbols descriptors remains the best and no other mapping was able to equal these results. The LogP increments by Ghose-Crippen are themselves based on fragments; it seems that extracting the information directly is better than using the mapping as a mean to inductive learning transfer. It should be noted though that although the Student's t-test shows a significant difference, performances are very close (14.5 for atom symbols and 14.6 for LogP 1 and LogP 3). Models using the LogP 1, LogP 3 and pharmacophoric properties mapping schemes are shown to not be significantly different from each other and are the second best models. In fact, when looking at bins attributed to atoms, a parallel can be found between the LogP mappings and the pharmacophoric properties. The LogP binning roughly reproduces the information withheld in the pharmacophore description. LogP 1 is the mapping schemes with the most bins and LogP 3 includes the fuzziness strategy in order to make boundaries less sharply.

Table 4.6: Performances of best models for each mapping scheme

Mapping scheme	Descriptor space	Cost parameter	RMSE	R^2_{det}	R^2_{corr}
LogP 1	IILp1(2-3)_R	0.1	14.6	0.73	0.85
LogP 2	IILp2B(2-3)_R	0.1	14.9	0.72	0.85
LogP 3	IILp3(2-3)_R	0.1	14.6	0.73	0.85
LogP 4	IILp4B(2-4)_R-FC	0.1	14.9	0.71	0.84
LogP 5	IILp5B(2-4)_ms-R	0.1	15.1	0.71	0.84
Atom symbols	IIAB(2-3)_R-FC	0.8	14.5	0.74	0.87
Pharmacophoric properties	IIPhB(2-4)_ms-R-FC	0.1	14.7	0.72	0.85

Adding the models using LogP 1 and/or LogP 3 mapping schemes with a 5-CV RMSE below 16 to the consensus model made the RMSE on the predictions of the external test set without the use of the AD drop by 0.2 (from 16.4 to 16.2). This indicates that the models do contribute to the robustness of the consensus model but on a small scale. This isn't surprising as the parallel between the pharmacophoric properties and the new schemes was observed. Thus, it can be assumed that the LogP increment schemes tried out do not provide a novelty to the problem. Furthermore, the calculation of the LogP increments and then their transformation into bins is more tedious than using the pharmacophoric properties defined from the substructure directly.

It should be noted however that these are encouraging results. I believe that for problems not extensively researched for specific descriptors (inorganic chemistry, ionic liquids...), the use of such a method to obtain descriptors from a previous model could lead to more efficient descriptors.

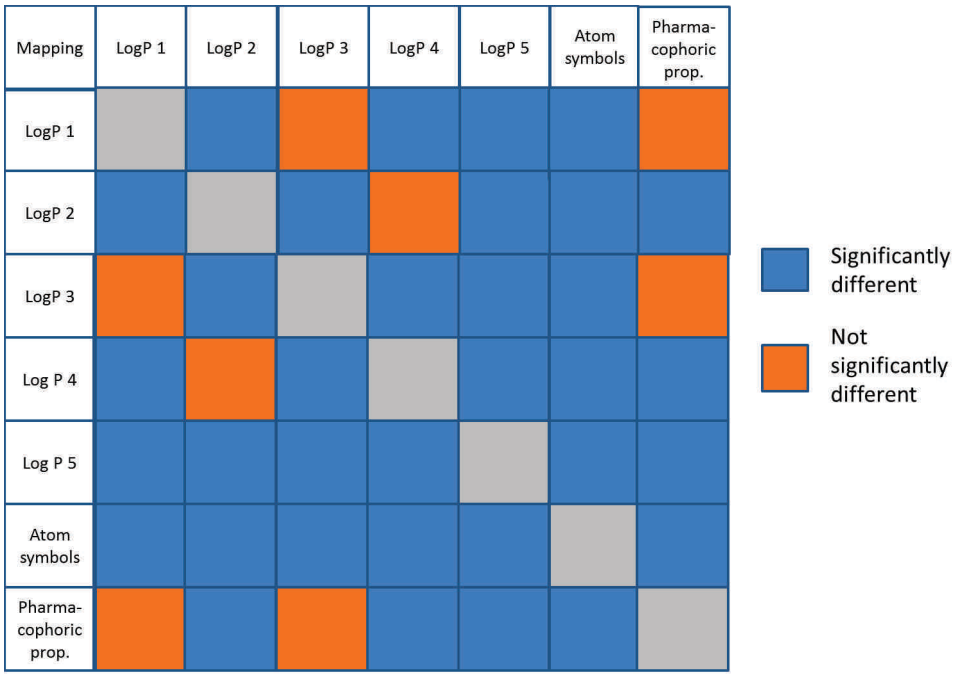


Figure 4.8: Student's t-test on the folds of the 5CV performed 10 times

4.3.3 Conclusion

To conclude, the CHI measurements done by our colleagues permitted to build QSAR models. An algorithm was proposed to detect the outliers from these models. These outliers were tested and showed that experimental errors altered their CHI value. The dataset was cleaned and a consensus model was build on this cleaned set. It was tested on an external literature set and performed with a RMSE of 14.3 with AD and a RMSE of 16.4 without AD. It is accessible freely on our web server (http://infochim.u-strasbg.fr/webserv/VSEngine.html). These results were published in Analytical Chemistry[23]. This project showed how important and profitable the synergy between experimental and theoretical teams is.

Furthermore, LogP-based descriptors were proposed and tested on the CHI data. It was shown that these descriptors achieve good results and are comparable to pharmacophore-based descriptors.

4.4 Chemogenomics-based virtual screening on GPCRs

4.4.1 Introduction

A collaborative research with the Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, University of Kyoto (Japan) aimed at identifying interesting compound descriptors amongst the IPLF descriptors for the application to the computational chemogenomics research. How much chemical diversity does a descriptor type give? The compound similarity matrices will be used to visualize -using histograms and an in-house SVM contour maps - the chemical space diversity generated by the descriptor space. This research was financed by the Japanese Society for the Promotion of Science (JSPS).

GPCRs were the main targets as they are the family of proteins with which most commercialised drugs interact[16] for this study. The project encountered several technical problems and particularly, a problem related to the memory outlined in section 4.4.4.1.

4.4.2 Method

Data Data was extracted from the GVK database[35] for the GPCRs. The database contains only interaction pairs (positives), therefore, non-interacting pairs (negatives) were assumed to be pairs absent from the database. As shown on Figure 4.9, the reference database is separated into known actives and presumed inactives. The figure also shows the different classes of compound-protein interactions (CPIs) that can be predicted with models generated from this data.

The data collection for GPCRs contains 1,402,282 CPIs comprising 232 different proteins and 523,518 different compounds. Building models on the whole data to test descriptor spaces is hardly efficient and would be too costly. In this study, 3 trials, each consisting of randomly chosen 5000 positives and 5000 negatives CPIs as training set and 5000 positives and 5000 negatives as test set were used. Test and training sets are non-overlapping and all protein targets were included in each trials.

The compound were standardised with our in-house script using ChemAxon's plugins[30] (see 7.1). The tool goes through several steps including filters to remove salts, molecules with over 100 heavy atoms and compounds containing metals. Then, the structures are standardised by stripping off hydrogen atoms, imposing a standard representation for groups such as nitro, calculating the major tautomer, calculating the major micro-species at pH 7.4 and representing aromatic bonds.

Descriptors 56 ISIDA descriptor spaces were calculated and are shown in Table 4.7. Note that due to the memory problem (see 4.4.4.1), larger descriptor spaces could not be computed during the project.

As protein descriptors, the local alignment (LA) descriptors were used with the default parameters from the first publication[36]. The scores are normalised before being used to compute the CPI matrices.

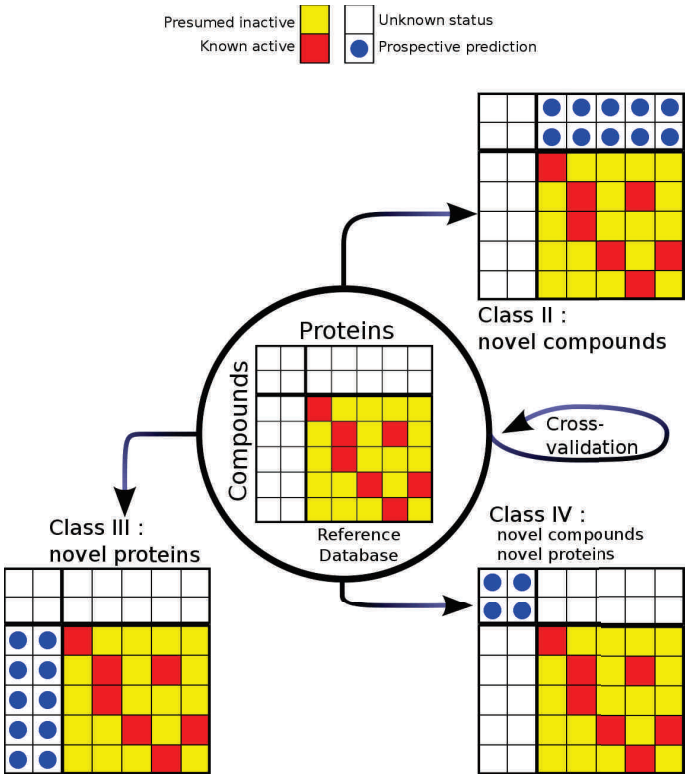


Figure 4.9: Representation of the data with the different classes of CPIs which may be predicted. This figure was reproduced with authorisation from J.B. Brown

Machine Learning The model building was done using SVM. The first step consisted of computing the similarity between the compounds and the one between the proteins. Then to compute the CPIs similarity matrices from these latter. Both of these steps were done using in-house programs by J.B. Brown (Kyoto University). Another program then enabled to bridge the CPIs similarity matrices with the LibSVM program[34] to obtain a model.

The normalized euclidean distance was used as a kernel function to compute similarity matrices. It is also known as the linear kernel and corresponds to a simple dot product between the vectors representing the molecule. Different SVM cost parameters have been tried ranging from 0.1 to 100. This model building procedure has been applied to all calculated ISIDA descriptor spaces (see Table 4.7) with the LA protein descriptor.

Validation Best models on each descriptor space were chosen by pareto optimum between the 3-CV AUC, external test accuracy and MCC.

4.4.3 Visualisation of molecular descriptor spaces

Molecular descriptor spaces were visualised using histograms of the compound similarity matrices' values. A wide spread indicates that the descriptor space is able to differentiate different classes of compound. Histograms were computed using in-house programs with 20 bins.

The SVM contour maps are directly derived from the SVM model and the instances used for training. A number of the largest support vectors are chosen and the concerned instances are used to calculate a distance matrix. Then, an algorithm known as multi-dimensional scaling[37] is applied to the distance matrix to obtain a representation of the space in 2D. The aim is to represent each pair of instances at a distance related to the one in the distance matrix. The contour is then obtained by plotting intensity surfaces, also known as heat maps, around the binders and non-binders.

4.4.4 Results and Discussion

4.4.4.1 Memory and cachesize problems

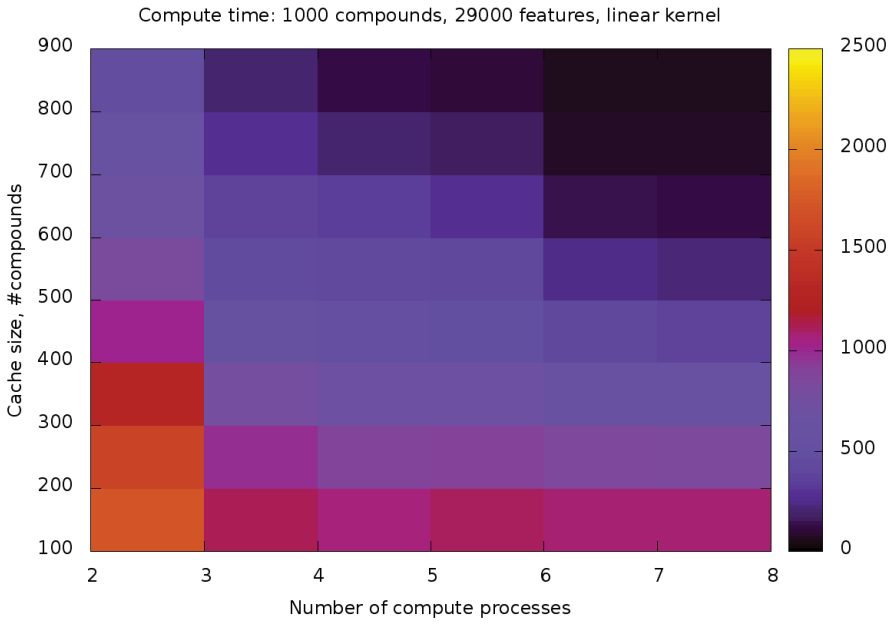


Figure 4.10: Compute time (in seconds) taken for the calculation of a compound similarity matrix with a linear kernel as a function of the number of CPUs and the cache size used on a set of 1000 compounds with a fingerprint size of 29000

The calculation of the compound similarity matrices ran into memory problems. The problem was identified to be related to the number of molecules and the length of the description which varies with the descriptor type chosen. In order to calculate effectively, the program loads the fingerprints for each molecule into the rapidly accessible memory,i.e. in the RAM. The bigger the fingerprints, the more memory is needed. Problems were first identified for fingerprints of size over 20000 with the training set (10000 compounds). ISIDA descriptors for so many compounds have a fingerprint size often above 20000 (see the “Size” column in Table 4.7).

The first measure taken to solve the problem was to calculate smaller sized descriptor spaces. Secondly, the RAM in the available machines was increased but it could not solve the issue entirely.

Finally, the compound similarity matrix program was modified to enable a selection of the cache size, which corresponds to the number of compounds' fingerprints stored into RAM. In order to speed up the calculation, the user can indicate the number of Central Processing Unit(CPU) to use for parallel calculations. A number of subprocesses are thus started on the different CPUs to calculate different parts of the similarity matrix. One subprocess brings back together the results. In the previous version of the program, each subprocess would need to allocate all compounds' fingerprints into RAM. The machines available at the time were not able to do so with even just one subprocess running because of the lack of RAM. The cache size option permits to only load a part of the needed information. When a compounds' fingerprint not loaded into RAM is needed, the least used one is replaced with the needed one. This swapping between the RAM and the disk memory consumes time. The parts of the similarity matrix calculated by each subprocess, are chosen to reduce as much as possible the swapping. The time consumption was analysed on a set of 1000 compounds with a fingerprint size of 29000. The results are illustrated in Figure 4.10. The new option permits the selection of a trade-off between number of CPUs, cache size, and total memory consumption.

Another important improvement to the tools available for CGBVS model building was the creation of a program regrouping all steps to obtain a model. This permits essentially to gain time by continuous calculation from what is stored in memory. Intermediate steps such as the compound similarity matrices and CPI matrices, do not need to be written out and reloaded any more as they use a lot of time and space. They may be written out if the user wishes to see them but they are written while the calculation for further steps is taking place. Thereby, no time is lost.

A minor improvement included the update of the program to read SVM fingerprint files as such.

Table 4.7: ISIDA descriptors, fingerprints sizes and average performances on the external test set

	Descriptors	Size	Acc	MCC	AUC		Descriptors	Size	Acc	MCC	AUC
1	IIIPh(1-5)ms	38062	0.87	0.74	0.90	29	IIAB(1-3)_R-P-FC	6303	0.84	0.69	0.88
2	IIFf(1-4)ms_P	95197	0.87	0.73	0.89	30	IIAB(1-3)_R	8823	0.84	0.69	0.88
3	IIFfB(1-3)ms_R	56969	0.86	0.72	0.89	31	IIFfB(1-2)ms	5695	0.84	0.69	0.89
4	IAB(1-7)_FC	26052	0.86	0.72	0.89	32	IIAB(1-3)_P	10260	0.84	0.69	0.88
5	IIIFf(1-5)ms	64906	0.86	0.72	0.89	33	IIFfB(1-2)ms_FC	6520	0.84	0.68	0.89
6	IFf(1-5)ms	47207	0.86	0.72	0.89	34	IIAB(1-4)	77020	0.84	0.68	0.88
7	IFfB(1-10)ms_P	16821	0.86	0.72	0.89	35	IIAB(1-3)_R-P	5020	0.84	0.68	0.88
8	IIFfB(1-3)ms_R-P	41227	0.86	0.72	0.89	36	IPh(2-5)ms	4289	0.84	0.68	0.88
9	IIFf(1-5)ms_R-P	50547	0.86	0.72	0.89	37	IAB(1-5)	3048	0.84	0.68	0.88
10	IIFf(1-3)ms_R	54388	0.86	0.72	0.89	38	IIA(1-4)_R-P	6153	0.84	0.68	0.88
11	IFfB(1-5)ms	58289	0.86	0.72	0.89	39	IPh(1-5)ms	4297	0.84	0.68	0.88
12	IIFf(1-3)ms_R-P-FC	44777	0.86	0.72	0.89	40	IAB(1-10)_P	1449	0.84	0.67	0.88
13	IPh(1-7)ms_FC	51402	0.86	0.72	0.89	41	IIA(1-4)	43365	0.83	0.67	0.86
14	IIFfB(1-3)ms_P	51886	0.86	0.72	0.89	42	IPh(1-10)ms_P	326	0.83	0.67	0.87
15	IFfB(2-5)ms	47171	0.86	0.71	0.89	43	IIPhB(1-2)ms_FC	4563	0.83	0.66	0.88
16	IFf(1-10)ms_P	6320	0.86	0.71	0.89	44	IA(1-7)_FC	5043	0.83	0.66	0.88
17	IIFf(1-3)ms	53147	0.85	0.71	0.89	45	IIA(1-3)	4179	0.83	0.66	0.88
18	IIPh(1-3)ms_R	52977	0.85	0.71	0.90	46	IIPhB(1-2)ms	3236	0.83	0.66	0.88
19	IIAB(1-4)_R	52941	0.85	0.71	0.89	47	IIA(1-3)_P-FC	4566	0.83	0.66	0.86
20	IIPhB(1-3)ms_R-P-FC	41905	0.85	0.71	0.89	48	IIIA(1-5)_FC	3306	0.83	0.66	0.88
21	IIPhB(1-3)ms_R	65627	0.85	0.70	0.89	49	IIA(1-3)_R	2658	0.83	0.65	0.87
22	IIPhB(1-3)ms_R-P	34075	0.85	0.70	0.89	50	IIA(1-3)_FC	5736	0.82	0.65	0.87
23	IIPh(1-3)ms_P	52897	0.85	0.70	0.89	51	IA(2-5)	692	0.82	0.64	0.85
24	IIPh(1-3)ms	60578	0.85	0.70	0.89	52	IIA(1-2)_FC	544	0.82	0.63	0.88
25	IIPhB(1-3)ms_P	88972	0.85	0.70	0.88	53	IIIA(1-5)	1826	0.81	0.63	0.87
26	IIIPh(1-5)ms_FC	11710	0.85	0.69	0.88	54	IIAB(1-2)	379	0.81	0.63	0.86
27	IPhB(1-5)ms	13218	0.84	0.69	0.88	55	IA(1-5)	704	0.80	0.61	0.85
28	IPhB(1-10)ms_P	2171	0.84	0.69	0.88	56	IA(1-10)_P	419	0.80	0.60	0.85

Notes: Acc stands for accuracy.

4.4.4.2 GPCRs models

Results for the 56 different ISIDA descriptors are summarised in Table 4.7. The results are ordered according to decreasing accuracy. The best descriptor space achieves a performance of 0.87 and is slightly better than previously obtained results[38] where the best result was an accuracy of 0.85 obtained with Extended-Connectivity FingerPrints (ECPFs)[39]. ECPFs are in concept very similar to the ISIDA descriptors augmented atoms so it is not surprising to find similar results. This previous study also used MACCS Public Keys[40] and Dragon descriptors[41].

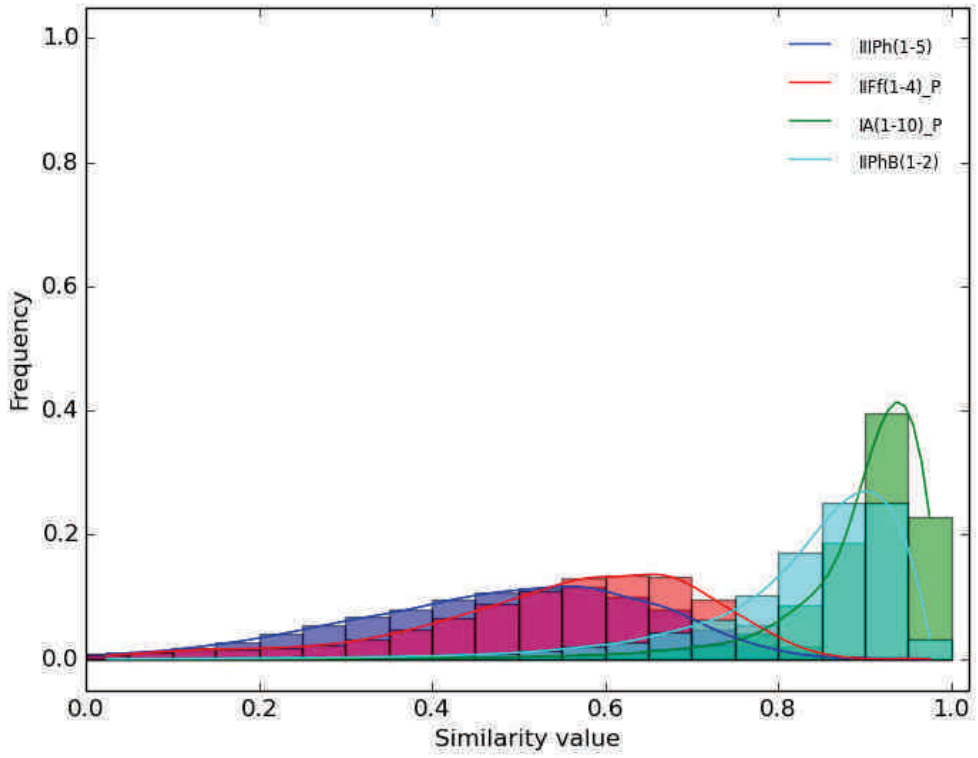


Figure 4.11: Distribution of the similarity values in different descriptor spaces. The best models correspond to a flat distribution (IIIPh(1-5)ms in dark blue and IIFf(1-4)ms_P in red), while the worst models (IA(1-10)_P in green and IIPhB(1-2)ms in light blue) see most pairs of molecules as similar.

ISIDA descriptors perform overall very well on this problem with the lowest found accuracy being 0.80 on the external test set although only small sized descriptor spaces were computed. Interestingly, the force field flag mapping (Ff) seems the most suited to this problem. This is not only visible in the accuracy ranking where the Ff mapping dominates the top but also on the histograms of the compounds' similarity matrices. Best descriptor spaces tend to have spread out distributions as illustrated in Figure 4.11. Such flat distributions were unknown to the research group before and we believe these descriptors may sample the compound space more diversely than previously used descriptors.

Other options, such as atom pairs (option P) or formal charge representation (option FC), did not seem to influence the results much. Restricted augmented atoms perform better overall than not restricted ones (option R).

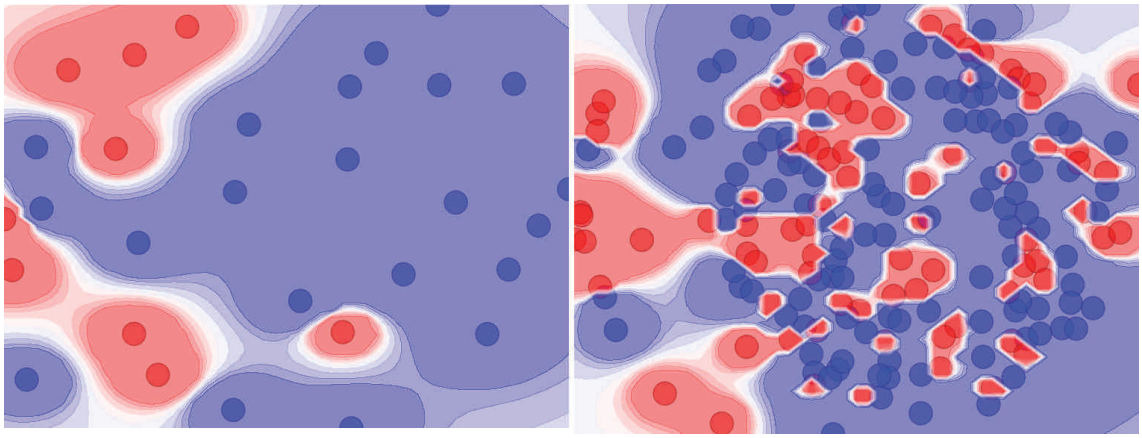


Figure 4.12: Visualisation of chemical space using multidimensional scaling. The dots represent the support vectors which are coloured by their class (red for actives and blue for inactives). These images were done on the best SVM model (IIIPh(1-5)ms) showing 30 support vectors (left) and 200 support vectors (right)

The SVM contour maps, as shown in Figure 4.12, show the complexity of the problem. They display the non-linearity of the nature of CPI interactions. We still need to do more research into modelling with the goal of deriving in methodology to give less complexity to feature space and more linearity. However, the contour maps showed that protein families or specific compounds were grouping into the same areas of space.

4.4.5 Conclusion

The project promoted the programs for the calculation of CGBVS models. The programs enables calculations with larger descriptor spaces and they can be set to fit the characteristics of the computer in terms of memory and CPU usage. One program reassembling all the different steps is available, thereby saving a lot of computational time which used to be lost in the writing processes. A small SVM visualisation tool was also done. The ISIDA tools were implemented for use in the host laboratory. The exchange enabled a closer collaboration between the two laboratories and was beneficial for both teams. First results from this project were encouraging and showed the good propensity of ISIDA descriptors, in particular force field flagging, in this specific problem. Indeed, an accuracy of 0.87 was achieved while previous results achieved 0.85.

Bibliography

[1] Bergstermann, H. *Naturwissenschaften* **38**(6), 128–132 (1951).

[2] Hansch, C., Maloney, P., Fujita, T., and Muir, R. *Nature* **194**(4824), 178–180 (1962).

[3] Braumann, T., Weber, G., and Grimme, L. H. *J Chromatogr A* **261**, 329–343 (1983).

[4] Valkó, K. and Slégel, P. *J Chromatogr A* **631**, 49–61 (1993).

[5] Lambert, W. J. *J Chromatogr A* **656**, 469–484 (1993).

[6] Valko, K., Bevan, C., and Reynolds, D. *Anal Chem* **69**(11), 2022–2029 Jun (1997).

[7] Valko, K., Du, C., Bevan, C., Reynolds, D., and Abraham, M. *Curr Med Chem* **8**(9), 1137–1146 Jul (2001).

[8] Cheng, Y. and Prusoff, W. *Biochem Pharmacol* **22**(23), 3099–3108 (1973).

[9] Keen, M., editor. *Methods in Molecular Biology, vol. 106, Receptor Binding Techniques*. Humana Press Inc, Totowa NJ, (1999).

[10] Brown, A. M. *Cell Calcium* **35**, 543–547 (2004).

[11] Sanguinetti, M. C. and Tristani-Firouzi, M. *Nature* **440**, 463–469 (2006).

[12] Taboureau, O. and Steen Jorgensen, F. *Comb Chem High T Scr* **14**(5), 375–387 (2011).

[13] Hedstrom, L. *Chem Rev* **102**(12), 4501–4524 (2002).

[14] Tymoczko, J. L., Berg, J. M., and Stryer, L. *Biochemistry. A short course*. W. H. Freeman and Company, New York, (2010).

[15] Wettschureck, N. and Offermanns, S. *Physiol Rev* **85**, 1159–1204 (2005).

[16] Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimoto, R., and Feng, C. *Nucl Acids Res* **36**(Suppl 1), D907–D912 (2008).

[17] Ruggiu, F., Marcou, G., Varnek, A., and Horvath, D. *Mol Inf* **29**(12), 855–868 Dec (2010).

[18] Horvath, D., Koch, C.and Schneider, G., Marcou, G., and A., V. *J Comput Aided Mol Des* **25**, 237–252 Jan (2011).

[19] Weber, L., Wallbaum, S., Gubernator, K., and Broger, C. *Angew Chem Int Ed Engl* **34**(20), 2280–2282 (1995).

[20] Weber, L. *Drug Discovery Today* **7**(2), 143–147 (2002).

[21] Schuller, A., Fechner, U., Renner, S., Franke, L., Weber, L., and Schneider, G. *Comb Chem High T Scr* **9**(5), 359–364 (2006).

[22] Weber, L. *Curr Med Chem* **9**(23), 2085–2093 (2002).

[23] Ruggiu, F., Gizzi, P., Galzi, J.-L., Hibert, M., Haiech, J., Baskin, I., Horvath, D., Marcou, G., and A., V. *Anal Chem* **86**, 2510–2520 (2014).

[24] Ghose, A. K. and Crippen, G. M. *J Comput Chem* **7**(4), 565–577 (1986).

[25] Viswanadhan, V. N., Ghose, A. K., Reyankar, G. R., and Robins, R. K. *J Chem Inf Comput Sci* **29**, 163–172 (1989).

[26] Plass, M., Valko, K., and Abraham, M. *J Chromatogr* **803**(1-2), 51–60 Apr (1998).

[27] Camurri, G. and Zaramella, A. *Anal Chem* **73**(15), 3716–3722 Jul (2001).

[28] Fuguet, E., Ràfols, C., Bosch, E., and Rosés, M. *J Chromatogr* **1173**(1-2), 110–119 Oct (2007).

[29] Denos, J. *Coloration des fragments ISIDA par increments LogP.* Master's thesis, (2012).

[30] *ChemAxon Calculator Plugin, JChem v.5.7.1*, (2010).

[31] R Foundation for Statistical Computing, Vienna, A. I. .-.-.-. h.-p., editor. *R Core Team.* R: A language and environment for statistical computing, (2013).

[32] Rosenblatt, M. *Ann Math Statist* **27**(3), 832–837 (1956).

[33] Parzen, E. *Ann Math Statist* **33**(3), 1065–1076 (1962).

[34] Chang, C.-C. and Lin, C.-J. *ACM Trans Intell Syst Technol* **2**(3), 1–27 (2011).

[35] (2011). GVK Biosciences Private Limited, Hyderabad India.

[36] Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. *Bioinformatics* **20**(11), 1682–1689 JUL (2004).

[37] Kruskal, J. B. *Psychometrika* **29**(1), 1–26 March (1964).

[38] Brown, J., Niijima, S., Shiraishi, A., Nakatsui, M., and Okuno, Y. In *IEEE International Conference on Bioinformatics and Biomedicine Workshops,* Gao, J., Dubitzky, W., Wu, C., Liebman, M., Alhaij, R., Ungar, L., Christianson, A., and Hu, X., editors (, Philadelphia, PA, 2012).

[39] Rogers, D. and Hahn, M. *J Chem Inf Model* **50**(5), 742–754 May (2010).

[40] Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. *J Chem Inf Comput Sci* **42**(6), 1273–1280 NOV-DEC (2002).

[41] DRAGON (Software for Molecular Descriptor Calculation) Version 6.0; Talete srl: Milano, Italy, .

Chapter 5

Applications of local descriptors

5.1 Introduction

Local descriptors have been developed for specific problems such as the acidity or the hydrogen bond strength. Indeed, such properties are associated to a specific atom and functional group of the molecule. One of the main important aspect of these problems is that molecules can be polyfunctional. Thus, a molecule has several values associated with it and it is unsuitable for machine learning to have an identical representation for these values. Using the marked atom strategy of the ISIDA descriptors to obtain local descriptors permits to add the information of the concerned functional group into the description and to differentiate the representations of the molecule associated to the different values. First, models on the acidity of molecules were developed only using the first marked atom strategy and are presented in the following section. Later on, the second and third strategies were introduced and tested on the hydrogen bond strength of acceptors (see 5.3). Both properties will be explained in their respective category.

The chapter is divided into two sections: one for the acidic dissociation constant (see 5.2) which follows this introduction and one for the hydrogen bond acceptor strength modelling (see 5.3). Three different projects were done on the acidic dissociation constant:

- Modelling of a literature set containing small organic acids and alcohols (see 5.2.3). This project was done with a student.
- Modelling of a subset of the French chemical library as part of the interdisciplinary research project (PIR) (see 5.2.4).
- Cleaning and modelling of a large database to extend our database (see 5.2.5). This project was done with a student.

5.2 Acidic dissociation constant

5.2.1 Introduction

Acidity of compounds is one of the most important physico-chemical factors in chemical reactions and interactions. For example, in drug design, the strength of the acidity

of a group determines the relative quantities of protonated/deprotonated micro-species present in solution at a given pH and thus be able to form ionic interactions with the protein. It is also closely related to the hydrophobicity and solubility of compounds which are essential in ADME/T. The ISIDA descriptors offer a good description to model the acidic dissociation constant, pK_a, by using the marked atom strategy to indicate the deprotonated atom. Unlike what is generally found in literature (see Table 5.1), the marked atom strategy enables to make use of all available data and build a global model instead of splitting the data into families. The fragment descriptors should enable the clustering of families implicitly within the model. Three small studies were done to explore this possibility. The main objective of this work was to build a model to predict the French chemical library, the "Chimiothèque Nationale" (CN) within the interdisciplinary research project (PIR).

Table 5.1: Summary of a few pK_a models found in literature

Authors	Machine Learning	Compound Family	Size	$RMSE$	SE [a]
Xing and Glen[1]	PLS	Acids	645		0.76
		Bases	384		0.86
Polanski et al.[2]	ANN/PLS	Benzoic acids	41	0.38	
		Alcanoic acids	46	0.35	
Xing et al.[3]	PLS	Acids	625		1.04
		Bases	412		1.12
Zhang et al.[4]	MLR	Aliphatic carboxylic acids	1122		0.42
		Alcohols	288		0.76
Ghasemi et al.[5]	MLR	Aromatic acids	74	0.27	
Miletti et al.[6]	PLS	Acidic nitrogen groups	421	0.41	
		N-heterocyclic bases	947	0.60	
Jelfs et al.[7]	PLS	Imines	84	0.55	
		Pyrimidines	91	0.43	
		Alcohol	202	0.58	
		Anilines	311	0.49	
		Pyridines	397	0.58	
		Carboxylic acids	681	0.34	
		Amines	1403	0.49	
Lee et al.[8]	Decision tree	Various compounds	1693	0.80	
Habibi-Yangjeh et al.[9]	GA-ANN	Various compounds	282	0.30	
Harding et al.[10]	SVM	Carboxylic acids	228	0.29	

[a] SE is the standard error which corresponds to the standard deviation.

5.2.2 Definition

Acids, as defined by Arrhenius, dissociate releasing a proton H^+ in water (see Equation 5.1). The corresponding equilibrium constant, K_a , is calculated from the activities of the compounds (see Equation 5.2). In turn, the activity depend on the concentration, temperature and ionisation strength of the solution[12]. K_a is commonly calculated by measuring the concentrations at a given temperature and ionic strength. The pK_a is then defined as the opposite logarithm of the equilibrium constant (see Equation 5.3).

$$AH + H_2O \rightleftharpoons A^- + H_3O^+ \tag{5.1}$$

$$K_a = \frac{a_{A^-} a_{H_3O^+}}{a_{AH} a_{H_2O}} = \frac{[A^-][H_3O^+]}{[AH]} \times \frac{\gamma_{A^-}\gamma_{H_3O^+}}{\gamma_{AH}} \tag{5.2}$$

$$pK_a = -log(K_a) \tag{5.3}$$

where a_X is the activity of compound X, [X] is the concentration of compound X and γ_X is the activity coefficient of compound X.

Measurements may be done with other solvents or a mixture in order to paliate certain problems such as insolubility and to extend the pKa scale below 1 and over 14 which are the minima and maxima values in water[12]. Measurements are then converted by means of an equation to the pK_a scale which is defined in water.

5.2.3 First study: Small organic acids and alcohols

The first study was done as a student project with Jacques Ehret and supervised with the help of Vitaly Solov'ev. Results were previously presented in the Master 2 project report of Jacques Ehret[13].

5.2.3.1 Method

Data Experimental values were collected from a book by Serjeant and Dempsey[14]. The dataset contained 705 small monofunctional organic molecules including carboxylic acids, phosphonic acids, thionic acids, alcohols, phenols and thiols. The molecules were standardized using our in-house tool based on ChemAxon classes (see 7.1). Chiral information was kept and all species were neutralized. After standardisation, molecules were verified one by one to find abnormal values. They were searched in literature to confirm pK_a values and their corresponding experimental conditions. pKa values were preferably chosen when measured at a temperature of 25 °C and an ionic strength of 0, which are the most common conditions. If those conditions were not found, the pKa value was taken from measurements with a temperature between 20 and 25 and an ionic strength between 0 and 0.1. Doubtful molecules were removed and it resulted in a set of 677 molecules. The first deprotonation site of the molecules was marked by hand with the help of ChemAxon's prediction[15].

Descriptors ISIDA descriptors of sequences and augmented atoms with atom symbols and electrostatic potentials were calculated with ISIDA Fragmentor2011. The descriptors were tested with and without the first marked atom strategy (MA1). Different maximum

Table 5.2: RMSE of the pK_a modelling for training and 5-CV

Descriptors	Machine Learning	Training RMSE	5-CV RMSE
Sequences - IA(2-12)_MA1	SVM	0.52	0.98
	MLR	0.91	1.11
	ASNN	0.35	1.04
Atom-centred atoms - IIA(1-3)_R-MA1	SVM	0.83	1.32
	ASNN	0.74	1.31

lengths were tried out within 3, 5, 8, 12 with a minimum length of 1 for augmented atoms, while the sequences were tested with a minimum of 1 and a maximum of 12.

Machine Learning Three methods were used:

- Consensus Multi-Linear Regression (MLR) with the ISIDA/QSPR program[16]. Note that only sequences with atom symbols descriptors could be used with this tool.
- Support Vector Machine (SVM) with the LibSVM package[17]. Epsilon regression SVM was chosen with a linear kernel. A first study was done to determine the best descriptor space by keeping the epsilon and cost parameter fixed. afterwards, the epsilon parameter was optimised between 0.001 and 1.001 with a step of 0.05 and the cost parameter was optimised between 0 and 5 with a step of 0.5.
- Neural Networks with the ASNN program[18]. The hidden layer contained 5 neurons and the seed was set to 0 (to be random). The best descriptor space obtained in the SVM study was used.

Validation The models were validated using a 5-fold cross-validation (5-CV) procedure 3 times.

5.2.3.2 Results and Discussion

The best descriptor space determined by the SVM benchmark for augmented atoms was an atom symbol based fragmentation with lengths from 1 to 3 and with the MA1 strategy (IIA(1-3)_R-MA1). Electrostatic potential mapping gave systematically worse and statistically insignificant results. Table 5.2 summarizes results for the different approaches.

It can be observed that sequences give better results than atom-centred atoms. SVM performs slightly better than the other methods. RMSE is rather high in general (about a log unit of error), in particular in comparison to literature (see Table 5.1). Most of the studies found in literature cannot be directly compared to ours because they are either focused on a particular compound family or trained on much smaller training sets. However, the study by Lee et al.[8] has various compounds and a bigger training set. It achieves a RMSE of 0.80 in 10-CV while our best approach achieves a RMSE of 0.98 in 5-CV. The approach used resembles ours by a fragment description using SMARTS. Closer analysis of the models show outliers (see 5.1 and 5.2).

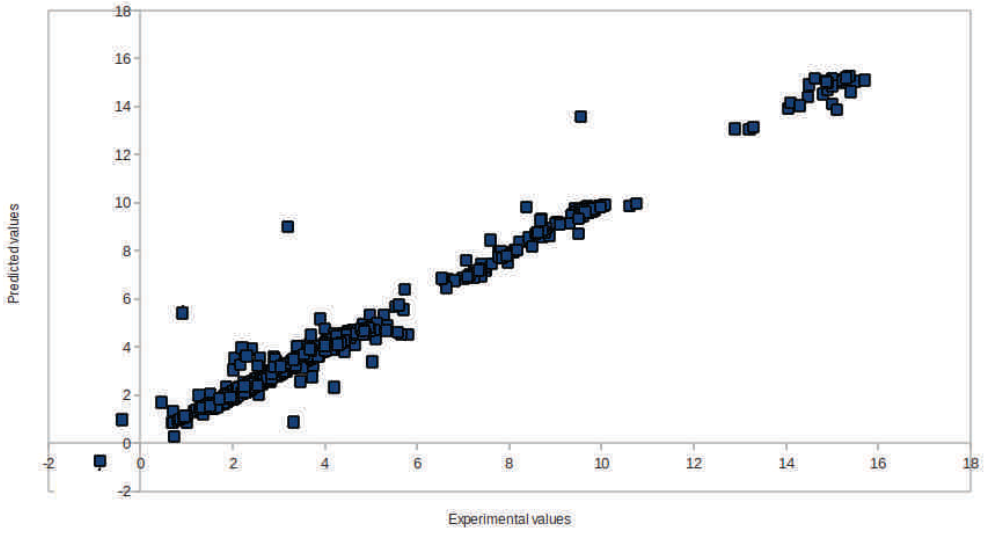


Figure 5.1: Training prediction by the SVM model with sequences IA(2-12)_MA1

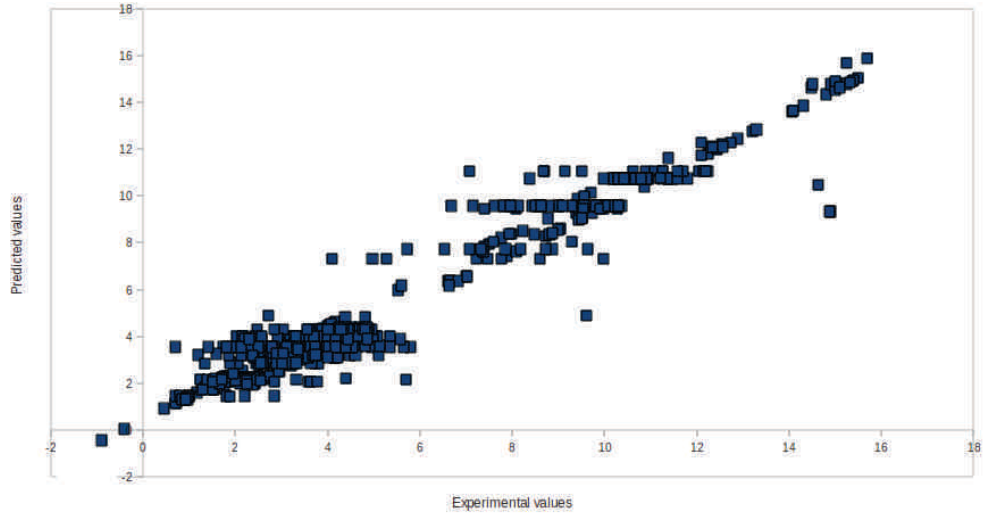


Figure 5.2: 5-fold cross-validation prediction by the SVM model with sequences IA(2-12)_MA1

It was observed that 6 alcohols were not well predicted, in particular by MLR models, although a great number of examples are available in the training set (167 molecules). It can be assumed that alcohols were not measured in water and the measurements were thus calibrated to the pK_a scale. Also, alcohols and carboxylic acids resembles each other fragment-wise and could be seen as very similar by the modelling. Thionic acids and fluorated alcohol compounds were identified with greater errors which could be explained by their relative small number of compounds in the set, 12 and 8 respectively. Certain known chemical effects were not rendered well such as the stabilisation due to a keto-enolic equilibria, the effect of ortho and para nitro group on a phenol and the conformational effect of Z, E conformers due to the hyperconjugation between the electronic acceptor atom and the hydrogen. The gauche effect can result in a pK_a difference of 1.5 log units between the conformers. The descriptors used could not account for the difference in

conformation and thus cases of Z,E conformers could not be treated. Too few examples (3) were found in the set to consider trying to model these.

5.2.3.3 Conclusion

This first study had encouraging results and showed that the marked atom descriptors were suitable for the pK_a problem. However, we found out how difficult the data is to clean. Conditions are mostly not indicated as well as protonation sites. Marking the atoms is a long and tedious work which lead us to develop a semi-automatic script (see 7.2) and test it during the following study on the subset of the French national chemical library (CN), the "Chimiothèque Nationale Essentielle" (CNE).

5.2.4 Second study: French national chemical library

5.2.4.1 Introduction

This study was part of the PIR and aimed at modelling the acidity of the molecules in the CN. The pK_a was experimentally assessed by the TechMed plateform.

Capillary electrophoresis (CE) was used to measure the pK_a values of the CNE subset. The principle of CE is based on the mobility of analytes through submillimeter capillaries exposed to a high voltage (see Figure 5.3). The analytes passing through the

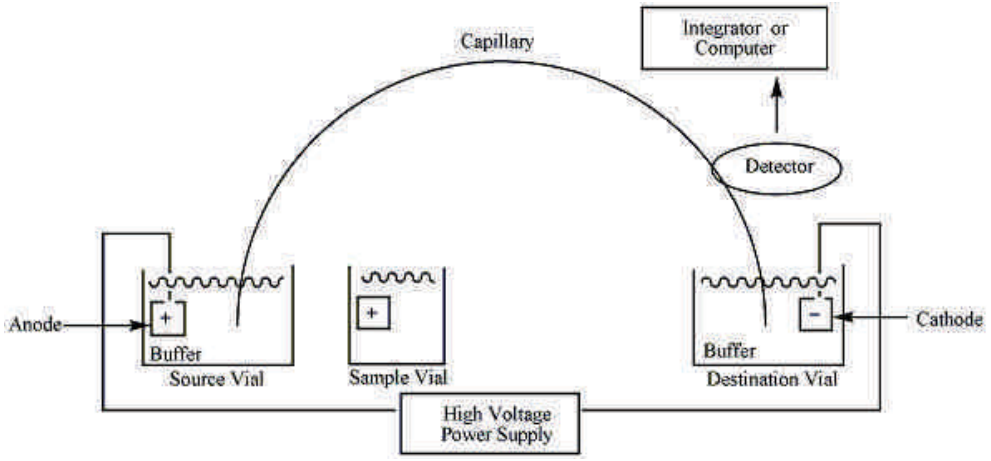


Figure 5.3: Schematic representation of capillary electrophoresis. Reproduced from wikipedia.org under the GNU Free Documentation License

capillary will be influenced by the voltage according to their ionisation state: positively charged ions will move faster than neutral species and negatively charged species will be slower than these latter. At different pHs, the proportions of the ionised micro-species will vary and thus their electrophoretic mobility vary. The effective mobility, μ_{eff}, depends on the ionisation rate, α, and the mobility of the ionized micro-species a, μ_a:

$$\mu_{eff} = \alpha\mu_a = \frac{10^{-pH}}{10^{-pK_a} - 10^{-pH}}\mu_a$$

By plotting the mobility of the ionised micro-species at different pHs, a curve is obtained (see Figure 5.4) and the pK_a is determined at the inflexion point.

CE offers various advantages including speed, lower cost and lower requirements for sample amount and purity than traditional techniques for the determination of pK_a such as titration with UV spectrometry for example[19].

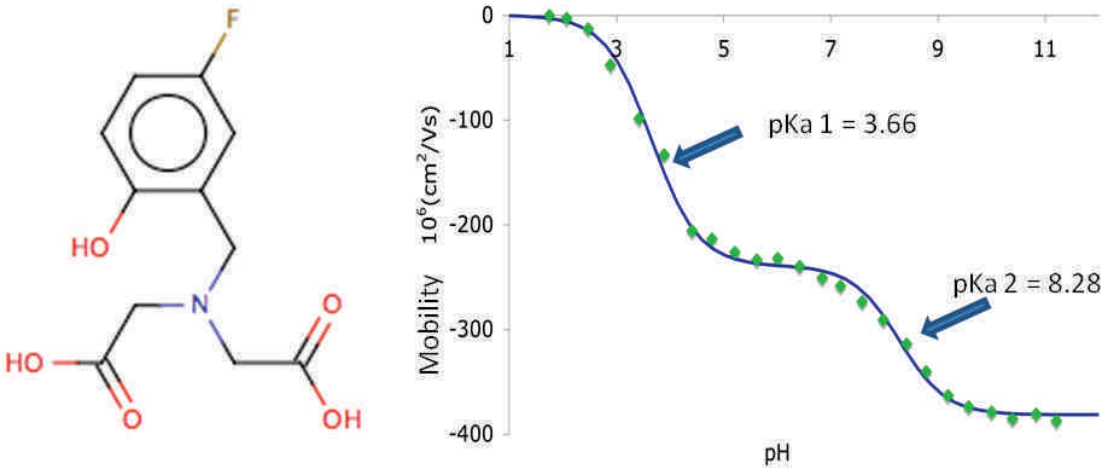


Figure 5.4: Curve obtained by measuring the mobility at different pHs with CE for the molecule on the left using the pKa PRO software

5.2.4.2 Experimental setup and results

Measurements of the pKa were carried out at the TechMed plateform in Illkirch-Graffenstaden by Patrick Gizzi according to the following protocol.

Out of the initial 640 compounds from the CNE, 381 compounds were received as powders to be measured by CE using a pKa PRO system by Advanded Analytical Technologies, Inc.

For each compound, a stock solution at 10 mM in DMSO is prepared by weighing between 1.3 to 2.0 mg of powder and adding DMSO. The stock solution is diluted for the CE analysis:

- For soluble compounds: 20% of methanol and 80% of aqueous solution (HCl at 1mM if the compound is basic or 1mM NaOH is the compound is acidic)
- For compounds with a low solubility: 60% of methanol and 40% of aqueous solution (HCl at 1mM if the compound is basic or 1mM NaOH is the compound is acidic). In this case, the cosolvent method is used.

For insoluble compounds, the cosolvent method was used by measuring the pK_a in a mixture of water and methanol at different proportions (v/v): 70/30, 60/40, 50/50 and 40/60. The measured values are then extrapolated to obtain the value for pure water (100/0) by the Yasuda-Shedlovsky method[20]. It has been demonstrated that a linear regression can be obtained by plotting the measured pKa in a cosolvent, p_sK_a, added

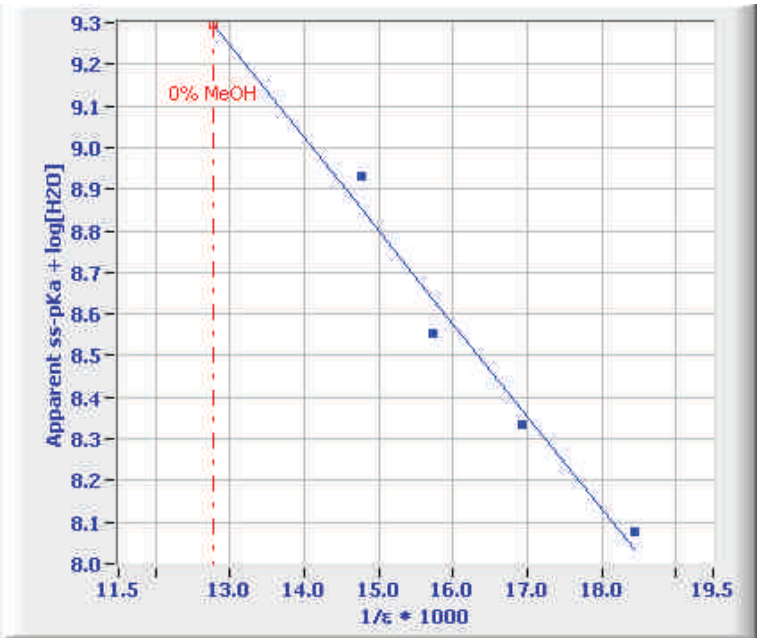


Figure 5.5: Example of cosolvent linear regression using the pKa PRO software

to the logarithm of the molar concentration of water, $log([H_2O])$ versus the dielectric constant, ϵ (see Equation 5.2.4.2).

$$p_sK_a + log([H_2O]) = \frac{A}{\epsilon} + B$$

where A is the slope and B is the intercept of the plot (see Figure 5.5).

In both cases, the stock solution is diluted to 1/50 to obtain a solution at 200 μM and 2% DMSO.

The diluted solutions are distributed on 96 wells microplates already filled with buffer solutions at different pHs ranging from 2 to 12:

- For soluble compounds: 50 μL per well and 24 wells per compounds (4 compounds can be measured on a microplate).
- For compounds with a low solubility: 50 μL per well and a microplate per compound.

A microplate is used beforehand to fill the capillaries before and during the electrophoresis experiment. The capillary are sucked up by the system during 2 seconds, then the high voltage is established (3.5kV) and compounds are detected at the end of the migration by a diode array detector at wavelength of 228 nm. The pKa PRO program is used to interpret results.

Out of the 381 compounds measured, 143 molecules had measurable pK_a values between 2 and 12. Problems encountered are presented in Figure 5.6.

5.2.4.3 Assignment of the pK_a values to polyfunctional molecules

The 143 molecules with values were analysed and attribution of the values to particular sites was done with the help of the ChemAxon pK_a plugin[15] - an example is given in

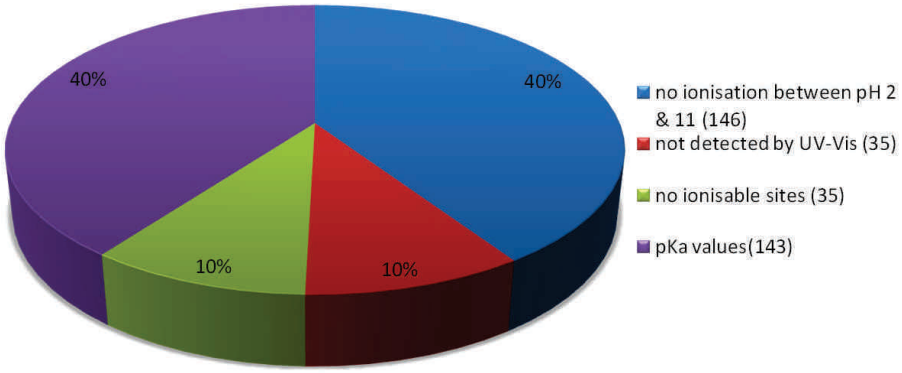


Figure 5.6: Experimental problems encountered when measuring the pK_a for 381 molecules from the CNE

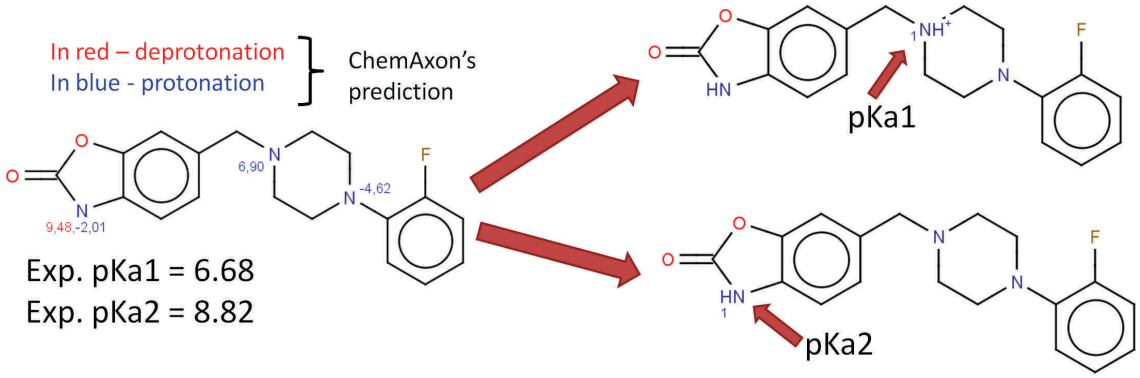


Figure 5.7: CNE molecule with ChemAxon predictions (red for acids and blue for bases) and experimental pKa values

Figure 5.7. This analysis permitted in parallel to set up the parameters for the automatic assignment in our in-house script based on ChemAxon (see 7.2). Although assignment to obtain marked atoms is tedious for small molecules with only one or two possible deprotonation sites as treated before, the problem in the case of the CNE became even more complicated: in certain cases, which site is deprotonated is unclear and also certain deprotonations seem to take place simultaneously but only yield one pK_a value.

In Figure 5.8, the dicarboxylic acid dissociate one after the other at respectively a pK_a of 2.61 and 5.57. In order to differentiate the two different deprotonations, the charge on the first deprotonated carboxylic acid needs to be included in the description of the second deprotonation. Therefore, the formal charge representation was implemented into the descriptors. It was also chosen that the molecules would be represented as the micro-species before the deprotonation occurs, i.e. in their acidic form.

Figure 5.9 exemplifies how the attribution of the values is difficult and unclear. No ChemAxon predictions are close to the actual experimental values. Errors of attribution can be made on such examples.

Figure 5.10 shows a molecule suspected of being deprotonated/protonated simultaneously. The gap between the tertiary amine's pK_a and the phenol's pK_a is probably small and

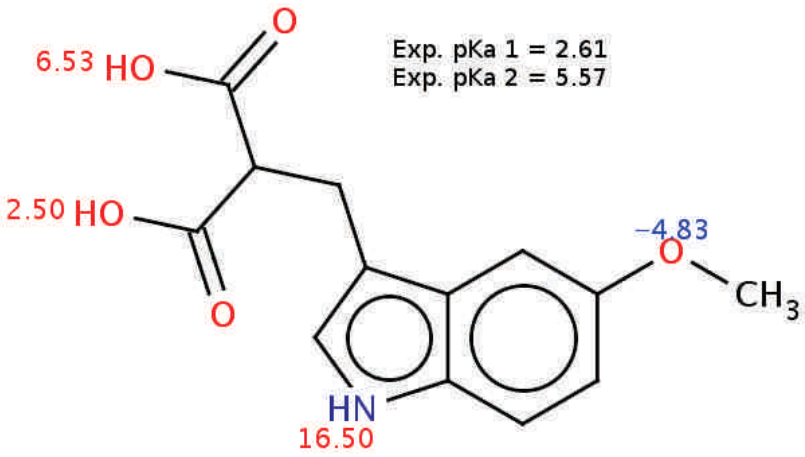


Figure 5.8: CNE molecule with ChemAxon predictions (red for acids and blue for bases) and experimental pKa values

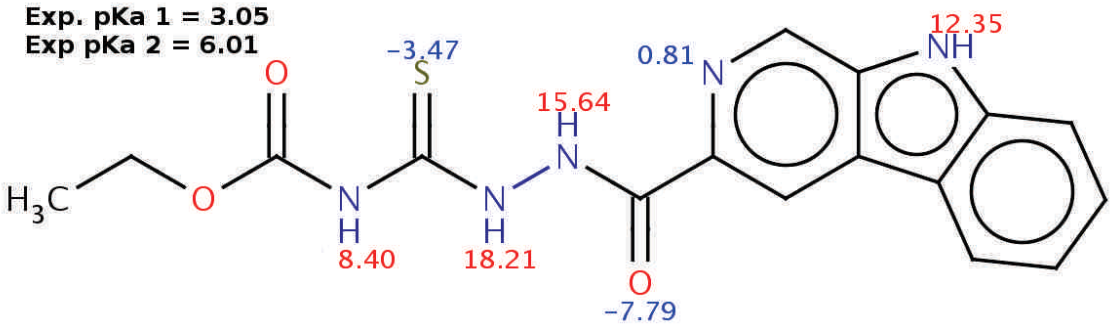


Figure 5.9: CNE molecule with ChemAxon predictions (red for acids and blue for bases) and experimental pK_a values

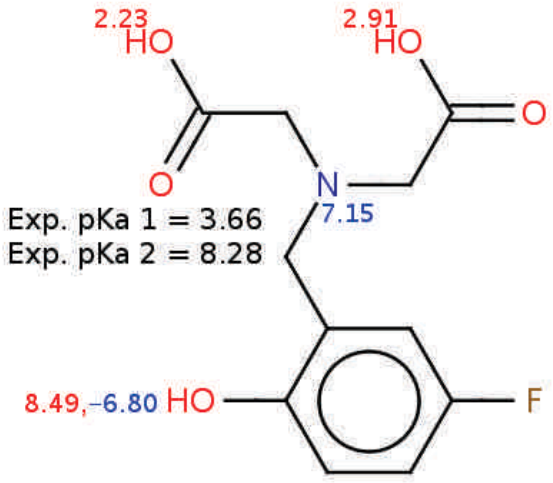


Figure 5.10: CNE molecule with ChemAxon predictions (red for acids and blue for bases) and experimental pK_a values

thus it is not visible experimentally. It can also be suspected that unlike in the first example shown in Figure 5.8, the carboxylic acids do not deprotonate consecutively but simultaneously because of the several bonds and the amine separating them. Their pK_a values are equivalent as the molecule is symmetric.

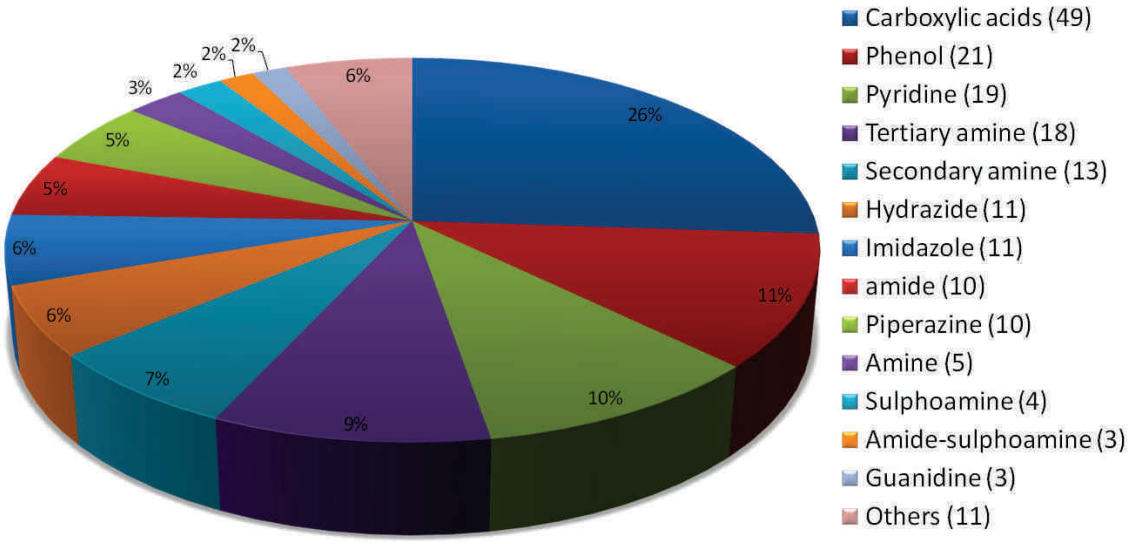


Figure 5.11: Associated group families to the 188 pK_a deprotonation sites

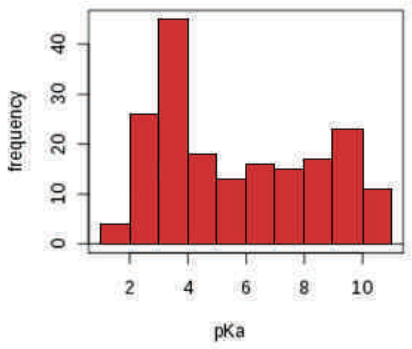


Figure 5.12: Distribution of the 188 pK_a values

In the end, 188 different deprotonation sites were assigned from the 143 molecules. It was found that the sites were very diverse as shown in the pie chart of Figure 5.11. The distribution of the pK_a values is shown in Figure 5.12. The peak around 3-4 is consistent with the fact that carboxylic acids are the most present in the set (49 examples).

5.2.4.4 QSAR method

Data The 188 depronation sites were marked and were represented in their acidic form and with the major tautomer according to ChemAxon using our in-house pK_a assignment tool (see 7.2).

Descriptors 960 ISIDA descriptor spaces were calculated by alternating:

- the use of the marked atom strategy 1, where a fragment starts with the marked atoms.
- the use of the formal charge representation
- between sequences and atom-centred fragments
- between atom symbols and pharmacophoric properties
- the minimum length between 2 and 4
- the maximum length between 2 and 10

Machine Learning and Validation Two distinct modelling were done with PLS and stepwise MLR.

The PLS was performed with Weka[21,22] using a number of component varying between 2 to 20. The models were validated using a 3-fold CV.

The stepwise linear regression was done with SQS[23]. The program internally separates randomly 1/3 of the data to be used as an external test set. It then builds a consensus model using the 3 best models build on 2/3 of the data according to a 3-fold CV.

5.2.4.5 Results and discussion

Statistical results are summarized in Table 5.3. ChemAxon's pK_a plugin[15] achieved the best results and predicted the experimental values with a RMSE of 0.97 and a determination coefficient R^2_{det} of 0.86. In general, the models build in this study performed rather poorly. The CN is known for being diverse[24] and the CNE was selected to be representative of this diversity. The number of different families present in this data (see Figure 5.11) illustrates it. This diversity and the small size of the data set make it a difficult set to model. The prediction of ChemAxon are favoured because they were used to mark the sites. However, the model behind the tool probably uses a much larger set than the available CNE set and thus yields better results.

The PLS turned out to be overfitted so models were selected with a low number of components (4). In addition, it was observed that the models were not able to differentiate the different classes of amines.

The use of the pharmacophoric mapping was tested to confirm that descriptors issued from them integrate the pH-dependent information correctly. Indeed, they perform better than the simple atom symbols. In practice, however, it does not make much sense to use them for pK_a prediction as they are based on a pK_a prediction. The formal charge representation is favoured when used in combination with atom symbols. Descriptor spaces representing atoms and bonds perform slightly better in average than those using only atoms. The fragmentation (sequences, atom-centred fragments, use of pair option) perform equally well.

Table 5.3: Statistical results of the pK_a modelling on the CNE set

Descriptors	Machine Learning	Validation	RMSE	R^2_{det}
IAB(2-6)_FC-MA1	3 MLR consensus	external	1.30	0.76
IPhB(2-7)_MA1	3 MLR consensus	external	1.23	0.78
IAB(2-8)_FC-MA1	PLS with 4components	3-CV	1.75	0.57
IPh(2-8)_P-MA1	PLS with 4components	3-CV	1.61	0.64
ChemAxon pK_a plugin			0.97	0.86

5.2.4.6 Conclusion and Perspectives

The aim of this project was to build pK_a models to annotate the CN with a model. However, models do not achieve better results than ChemAxon predictions, probably due to the lack of data and the diversity of the set. Therefore, at the moment the CN is annotated using ChemAxon's pK_a plugin since it is free for academic research. However, the relative success of the ChemAxon tool - most likely not trained on capillary elecrophoresis pK_a data, but mainly on traditional titration-based pK_as - also shows that there is no fundamental reason to separate these measurements from literature training sets, even if measurement conditions may slightly differ. In this perspective, a training set fusion of CNE and literature compounds may turn out to be a much more solid basis to build a valid pK_a model.

The benefit of using the marked atom strategy was confirmed in this study as well. It also showed that the formal charge representation is essential in conjunction with the atom symbols.

A in-house script was developed to semi-automatically assign the pK_a values to their prospective site. It also permits to standardise the micro-species into their acidic form.

5.2.5 Third study: Database cleaning

The aim of this study was to increase the number of available pK_a data in order to palliate the problems encountered in the previous study by cleaning a huge database. It was done as a student project with Guillaume Charbonnier. Results were previously reported in his master 1 project report[25].

5.2.5.1 Method

Data: The data for this project was extracted from the ChemDB database from the DISCON program[26]. The raw database contained 16 086 entries in SDF format but exporting these structures failed for 34 entries due to errors such as atoms bound to themselves. Only the structure and the pK_a values are given in this database without references, experimental conditions or indication of the deprotonated group. The database contains organic molecules and druglike compounds with sizes from 1 to 58 heavy atoms. Various functions are found including carboxylic acids, amines, thiols, etc. with a pK_a varying from -7.30 to 19.20.

The molecules were standardised using our in-house script based on ChemAxon's classes (see 7.1) and choosing to remove chirality and major tautomers. The standardised smiles

Table 5.4: Statistical results of the pK_a modelling for Training and 3-CV

Descriptors	C	Training				3-CV			
		R^2_{det}	R^2_{corr}	RMSE	MAE	R^2_{det}	R^2_{corr}	RMSE	MAE
IIAB(1-8)_R-FC-MA1	18	0.95	0.97	0.73	0.48	0.73	0.85	1.69	1.09
IAB(1-7)_AP-FC-MA1	20	0.88	0.94	1.11	0.76	0.72	0.85	1.69	1.12

were then used to identify molecules with several entries. 8856 unique molecules were identified from which 3783 had several values. The latter were difficult to handle because values do not correspond necessarily to a different deprotonation but probably to a different reference of the same deprotonation. Hence, the set of 5073 entries with one pK_a value was used and marked using our in-house script based on ChemAxon to assign protonation sites (see 7.2). Molecules are represented in their protonated state. This resulted in a set of marked molecules containing 4922 compounds.

Descriptors: ISIDA fragment descriptors were calculated with ISIDA Fragmentor2012. The marked atom strategy where fragments start with the marked atom (MA1) were calculated with the different possible fragmentations (sequences, augmented atoms, triplets) and different lengths varying from 1 to 8. The explicit formal charges as well as the all path exploration were also tried. In total, 684 descriptor spaces were generated.

Machine Learning: Partial Least Square (PLS) regression models were built with the Weka software[21,22]. The number of component (C) were tried between 2 and 20 with a step of 2. Thereby, 6840 models were created.

Validation: 3-fold cross-validation was performed 5 times to validated the model and RMSE was used as a reference statistical parameter.

5.2.5.2 Results and Discussion

The assignment of the pK_a values to a site was done automatically for 40% of the molecules. The remaining 60% were done manually with the help of the ChemAxon predictions. Several cases were ignored, particularly when too many possible sites were present (see 5.13). Results for the first two best models according to 3-CV R^2_{det} in unique descriptor spaces are given in Table 5.4. In general models topping the 3-CV R^2_{det} list have a high number of components (C=16-20) and use the formal charge option. Triplets are not found in this list, the first model is on the 684^{th} position out of 6 840 and has a 3-CV R^2_{det} of 0.55. Hence, it can be assumed that triplets are not good descriptors for pK_a predictions.

In comparison to our best model, ChemAxon predictions achieve the following results: $R^2_{det} = 0.85$, $R^2_{corr} = 0.92$, $RMSE = 1.29$ and $MAE = 0.79$ on 4860 molecules because 62 values were not predicted by the model. Literature models also seem to perform better, for example the study by Xing and Glen[1], also uses PLS regression and similar

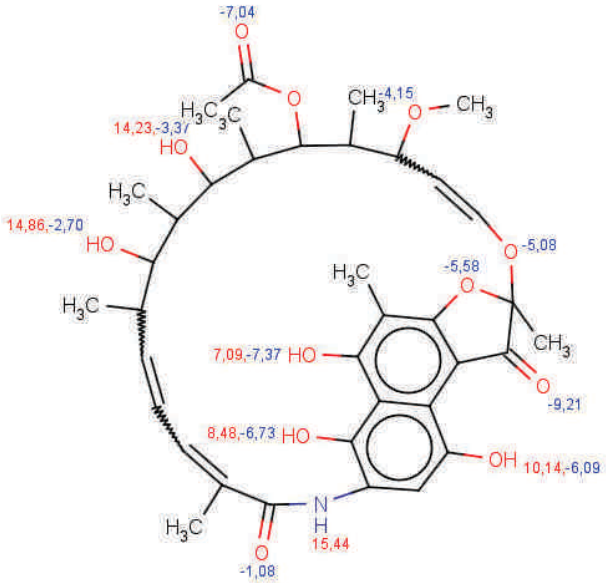


Figure 5.13: Example of a molecule for which the pK_a value could not be attributed. $pK_a = 3.00$. ChemAxon's predictions for acids are given in red and for bases in blue.

descriptors which are tree-like fragments with a Sybyl force field typing mapped on the molecular graph. They separate acids (645 molecules) and bases (384 molecules) before building their models and obtain a 10-CV R^2_{det} of 0.85 and a standard error of 0.76. We recalculated the 10-CV R^2_{det} for our best model and obtained 0.75. Our results are encouraging but do not top literature models. The data obviously still needs to be cleaned further. The models had several outliers; the first 5 are presented in Table 5.5.

5.2.5.3 Conclusion

Part of the data was cleaned and standardised in order to make pK_a models. First results are encouraging but the set needs further cleaning.

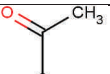
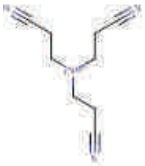
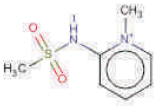
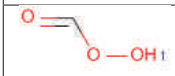
5.2.6 Conclusion and Perspectives

The pK_a project isn't finished and more data from the literature should be cleaned and converted to try new models. The quality of the data is in general mediocre because conditions are often missing from databases and even publications and therefore it is probable to find bigger variations of the error within the dataset. However, it can be assumed that most measurements are done around room temperature ($20 - 25\,^{\circ}C$) and the influence of the ionic strength is small.

The CNE data being a difficult set, it could be used as an external set to test the various models and available programs.

The three studies have confirmed the importance of using the marked atom strategy as well as formal charge representation.

Table 5.5: Most outlying molecules in the best pK_a model in 3-CV

Structure	Best model's prediction	Experimental pK_a	Comment
	4.52	15.70	The water molecule only has the marked O as a descriptor. The model seems to mistaken it for a carboxylic acid.
	10.70	19.20	The assignment was wrong. Instead of marking the oxygen, the alpha carbon should have been marked.
	6.50	1.10	Few similar compounds (15 out of 4922)
	4.50	-0.33	Unique structure
	9.50	7.10	Few similar compounds (10 out of 4922)

5.3 Hydrogen bond acceptor strength

This project has been the object of a publication[27] and is included in the following pages. An introduction to hydrogen-bond (H-bond) acceptor strength and results of our studies are described in it. This project was a collaboration with Vitaly Solov'ev from the Institute of Physical Chemistry and Electrochemistry (Moscow, Russia) and Jérôme Graton and Jean-Yves Le Questel from the University of Nantes (France). The H-bond acceptor strength data was measured and collected into a database by our colleagues from Nantes and Vitaly Solov'ev carried out the modelling using MLR.

Full Paper

www.molinf.com

molecular informatics

Special Issue STRASBOURG

DOI: 10.1002/minf.201400032

Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules

Fiorella Ruggiu,[a] Vitaly Solov'ev,[b] Gilles Marcou,[a] Dragos Horvath,[a] Jérôme Graton,[c] Jean-Yves Le Questel,[c] and Alexandre Varnek*[a]

Abstract: Here, we introduce new ISIDA fragment descriptors able to describe "local" properties related to selected atoms or molecular fragments. These descriptors have been applied for QSPR modelling of the H-bond basicity scale pK_{BHX}, measured by the 1:1 complexation constant of a series of organic acceptors (H-bond bases) with 4-fluorophenol as the reference H-bond donor in CCl_4 at 298 K. Unlike previous QSPR studies of H-bond complexation, the models based on these new descriptors are able to predict the H-bond basicity of different acceptor centres on the same polyfunctional molecule. QSPR models were obtained using support vector machine and ensemble multiple linear regression methods on a set of 537 organic compounds including 5 bifunctional molecules. They were validated with cross-validation procedures and with two external test sets. The best model displays good predictive performance on a large test set of 451 mono- and bifunctional molecules: a root-mean squared error $RMSE = 0.26$ and a determination coefficient $R^2 = 0.91$. It is implemented on our website (http://infochim.u-strasbg.fr/webserv/VSEngine.html) together with the estimation of its applicability domain and an automatic detection of potential H-bond acceptors.

Keywords: Equilibrium constants of hydrogen bonding · Hydrogen-bond acidity and basicity · Fragment descriptors · H-bond · ISIDA · pK_{BHX} · QSPR

1 Introduction

The hydrogen bond (H-bond) is one of the fundamental interactions between molecules and is of paramount importance for many properties, as well as for processes of living and abiotic nature. Many hydrogen-bonding effects are known such as density differences between ice and liquid water, joining cellulose microfibrils in wood, shaping DNA into genes and polypeptide chains into wool, hair, muscles or enzymes.[1]

The term hydrogen bond has been used for over a century but its precise definition and the nature of the interaction has been and is still stirring many debates. The IUPAC proposes the following short definition in a recent report:[2] "*The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation.*". Criteria for evidence of bond formation include the linearity and directionality of the interaction, spectroscopic evidence with a red-shift of the X–H vibrational frequency in infrared (IR) spectroscopy or the deshielding of the H in X–H in nuclear magnetic resonance spectroscopy and the thermodynamic characterization by the Gibbs free energy ΔG°.

In the case of intermolecular interactions, a hydrogen bond is formed between the molecule containing the X–H, referred as the H-bond donor (HBD), and the molecule containing the atom with which the X–H forms a bond, referred as the H-bond acceptor (HBA). These terms come from the electronic aspect of the hydrogen bond where the H-bond donor/acceptor is considered analogous to a Lewis acid/base. Hence, the term basicity is used to express the H-bond acceptors' strength which can be measured by thermodynamic quantities such as the equilibrium constant, the free energy, the enthalpy and the entropy of

[a] *F. Ruggiu, G. Marcou, D. Horvath, A. Varnek*
Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg
1, rue Blaise Pascal, 67000 Strasbourg, France
phone: + 33368851560
**e-mail: varnek@unistra.fr*

[b] *V. Solov'ev*
Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences
Leninskiy prospect, 31a, 119991, Moscow, Russian Federation

[c] *J. Graton, J.-Y. Le Questel*
Université de Nantes, UMR CNRS 6230, Chimie Et Interdisciplinarité: Synthèse, Analyse, Modélisation (CEISAM), UFR Sciences & Techniques
2, rue de la Houssinière, BP 92208, 44322 NANTES Cedex 3, France

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201400032.

the association reaction (see Equation 1) where HBA···HBD is the 1:1 formed complex.

$$HBA + HBD \rightleftharpoons HBA \cdots HBD \quad (1)$$

The H-bond is also essential in protein-drug interactions[3] and it seems that the evaluation of the strength of such an interaction is not intuitive for medicinal chemist.[4,5] Hence, a quantitative assessment of H-bond strength has for a long time been important for the chemical community. H-bond basicity scales, pK_{HB} and log K_{β},[6] have been constructed, based on the equilibrium constants of 1:1 complexation of a reference HBD, 4-fluorophenol and 4-nitrophenol, respectively, with different acceptors in CCl_4. The pK_{HB} scale was later extended to the pK_{BHX} scale by Laurence et al.[4] using Fourier transform IR (FTIR) spectroscopy. In Equation 2, the pK_{BHX} value is defined as the logarithm of the equilibrium constant, measured at 298 K, with 4-fluorophenol as the reference HBD. A strong advantage of FTIR measurements is found for the study of polyfunctional compounds, for which the significant multiple H-bond sites are observable and a measured pK_{BHX} value can be attributed to each site. To our knowledge, with around 1200 values, the pK_{BHX} database[4] constitutes the largest collection of data on H-bond basicity. Moreover, the diversity of H-bond acceptor functional groups encountered in the pK_{BHX} database enables the building of QSPR models with an expected large applicability domain.

$$pK_{BHX} = \log K = \log([HBA \cdots HBD]/[HBA][HBD]) \quad (2)$$

Earlier, the modelling of thermodynamic parameters of the 1:1 H-bond complexes has already been attempted through various approaches such as quantum chemical methods,[7–12] linear free-energy relationships (LFERs),[13–17] empirical correlations[18–24] and quantitative structure-property relationships (QSPR) using results of quantum chemical calculations as descriptors.[7–11, 25–28] The LFERs models by Raevsky[15–17] and Abraham[14] consider the free energy of H-bond complexation as a product of the acceptor and donor parameters. To our knowledge, they were not properly validated except once on a small test set including 6 reactions.[15] QSPR models by Henneman et al.[26] for pK_{HB} based on AM1-calculated descriptors were obtained considering a limited set of 42 aromatic N-heterocycles and validated on a small set of 17 compounds, resulting in a mean absolute error of 0.17 log K units. The models by Besseau et al.[7] based on density functional theory approach were trained on 59 monofunctional nitrogen bases and validated on an external test set of 142 compounds with a root mean-squared error (*RMSE*) of 0.13 (calculated from the data reported in[7]). Klamt et al.[29] used the COSMO-RS approach to assess experimental H-bond enthalpies and free energies of about 300 H-Bond complexes from the pK_{BHX} database with an accuracy of ± 2 kJ mol^{-1} (± 0.35 log K units). In another recent study, Kerdawy et al.[12] performed a series of density functional/basis set combinations and second-order Møller–Plesset calculations on the complexes of 58 simple HBAs, mainly pyridine nitrogen and carbonyl oxygen sites, from the pK_{BHX} database, with methanol as HBD. A partial optimisation of the H-bond complex leads to reasonable correlations between pK_{BHX} and the calculated interaction energies, but no validation on an external test set was reported. Green et al.[28] found reasonable linear correlations ($R^2_{corr} = 0.91$–0.97) between pK_{BHX} measured for 41 HBAs with quantum chemical topology descriptors calculated for the complexes of these compounds with 5 different HBDs (water, methanol, 4-fluorophenol, serine and methylamine). The correlation equation for methanol was successfully used to assess pK_{BHX} values for 11 bifunctional HBAs.

In our previous publication,[30] the ISIDA fragment descriptors were used to model free energy (ΔG°) and enthalpy (ΔH°) of the 1:1 complexes between organic acids and bases linked by one H-bond. In these complexes, the acids were substituted phenols, whereas the bases were represented by the large variety of chemical classes: phenols, alcohols, ethers, ketones, amides, heterocyclic compounds, phosphoryl and sulfonyl compounds. The ensemble Multiple Linear Regression (MLR) model built on a training set of 292 complexes was validated on a test set of 66 complexes. A reasonable correlation between predicted and experimental values was observed:

$$\Delta G_{pred} = 0.10 + 1.00\Delta G_{exp} \quad (R^2_{corr} = 0.92,\ RMSE = 1.64 \text{ kJ mol}^{-1}$$

i.e., 0.29 log K units).

Nowadays, mostly polyfunctional molecules (see Figure 1 as an example) are used to design new nanomaterials based on a network of H-bonds or in the drug design area. Attempts to assess simultaneously pK_{BHX} values of their different sites are still scarce. Quantum chemical models (e.g.

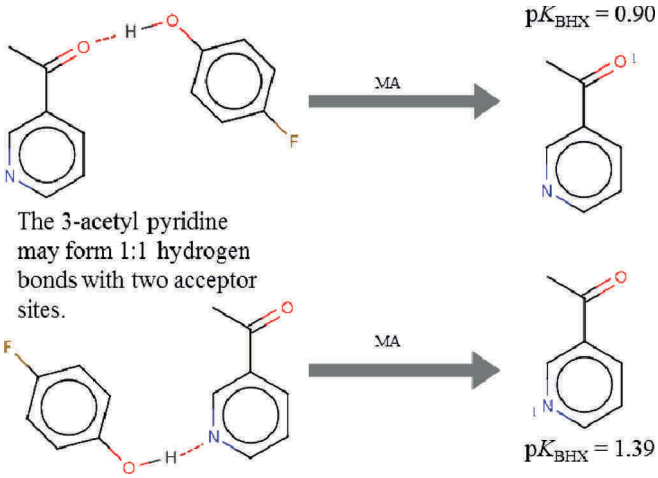


Figure 1. Example of marked atom (MA) assignment on 3-acetyl pyridine. The marking is indicated with a 1 next to the acceptor's centre.

Literature[8–13]) can potentially be used for this purpose, in support of experimental measurements, to characterise the individual H-bond basicity of acceptor sites encountered in progesterone,[31] cotinine,[32] lobeline,[33] myosmines,[34] codeine and galanthamine.[35] Performance of quantum chemical topology descriptors to treat polyfunctional molecules have also been demonstrated by Green et al.[28] At the same time, heavy and time-consuming quantum mechanics-based approaches could hardly be recommended for virtual screening of large databases frequently used in computer-aided drug- or material design. The need for fast and reliable QSPR approaches able to assess HBD or acceptor ability of different binding sites of polyfunctional molecules is thus obvious. QSPR modelling is closely related to the question of models' applicability domain. Up to now, the best performing models were built on small data sets containing very specific acceptor chemotypes which certainly limited their application to similar compounds.

In extension of our previous work,[30] we here report new molecular descriptors specifically developed to model hydrogen-bonding parameters of polyfunctional molecules. These new descriptors have been used to build QSPR models for pK_{BHX} on a large structurally diverse dataset containing 537 mono- and bi-functional compounds. The models were validated on two external test sets containing 451 and 36 HBAs and were implemented on the web for the end users.

2 Computational Procedure

The general procedure followed in this work is summarized in the workflow shown in Figure 2. First the dataset is extracted from the pK_{BHX} database and processed in order to build different QSPR models. The resulting models were validated on two external test sets.

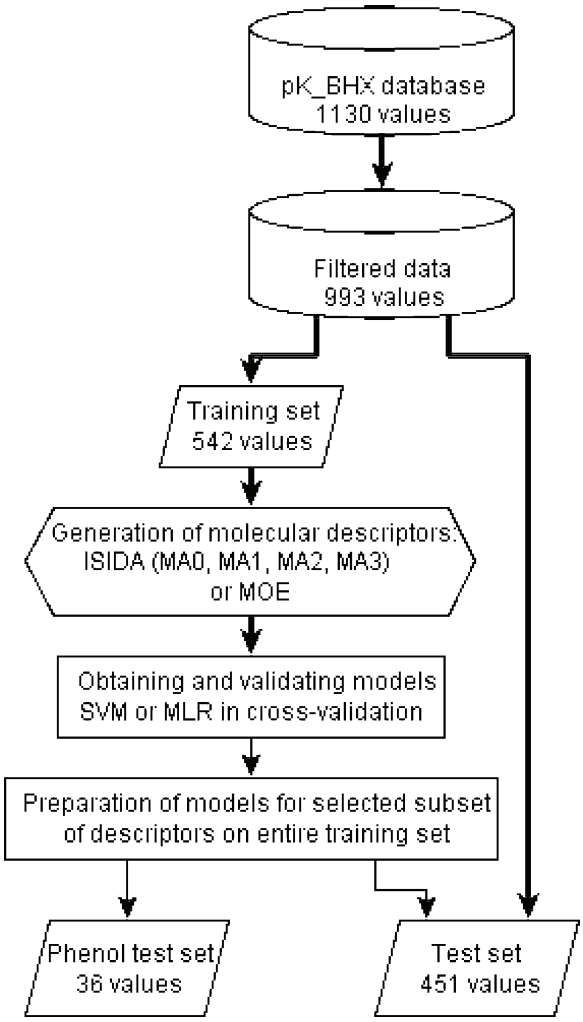


Figure 2. Workflow for the QSPR modelling of the pK_{BHX} database.

2.1 Data Preparation

Molecules from the pK_{BHX} database[4] were first filtered by removing all predicted values and by removing salts. An entry containing iron was also removed. They were then standardised by using an implicit representation of hydrogen, choosing a standard representation for groups such as nitro or imidazole, and generate major tautomer using ChemAxon's Calculator plugin.[36] The acceptors' sites had been identified experimentally during the measurement: they are indicated in the database and were thus easy to mark. If a compound contained several equivalent acceptor sites, only one of them was kept. The EdiSDF software[37] has been used to mark HBA centres.

The training set contains organic acceptors including different types of hydrogen-bonding atoms which have been categorised into 11 families:

- the carbonyl oxygen (esters, carbonates, lactones, aldehydes, ketones, amides, lactams, carbamates, ureas),
- the ether oxygen (ethers, alcohols),
- the oxygens of nitro group (nitroalkenes, nitroaromatics, nitramides),
- the oxygen of sulfinyl group (sulfites, sulfoxides),
- the oxygen of phosphoryl group (phosphoramides, phosphine oxides, phosphonates, phosphates),
- the amine nitrogen (anilines, amines),
- the imine nitrogen (amidines, pyrrolines, imines),
- the aromatic nitrogen (pyridines, azoles),
- the nitrile nitrogen (nitriles),
- the sulfur of sulfide (sulfides),
- the sulfur of thiocarbonyl group (thioamides, thioureas, thiocarbonates, thioketones, isothiocyanates).

For five bifunctional compounds (3-acetylpyridine, morpholine, *N*-methylmorpholine, thiomorpholine, and thiazolidine), the pK_{BHX} values were specified for two different acceptor sites. For the molecules in the training set, pK_{BHX}

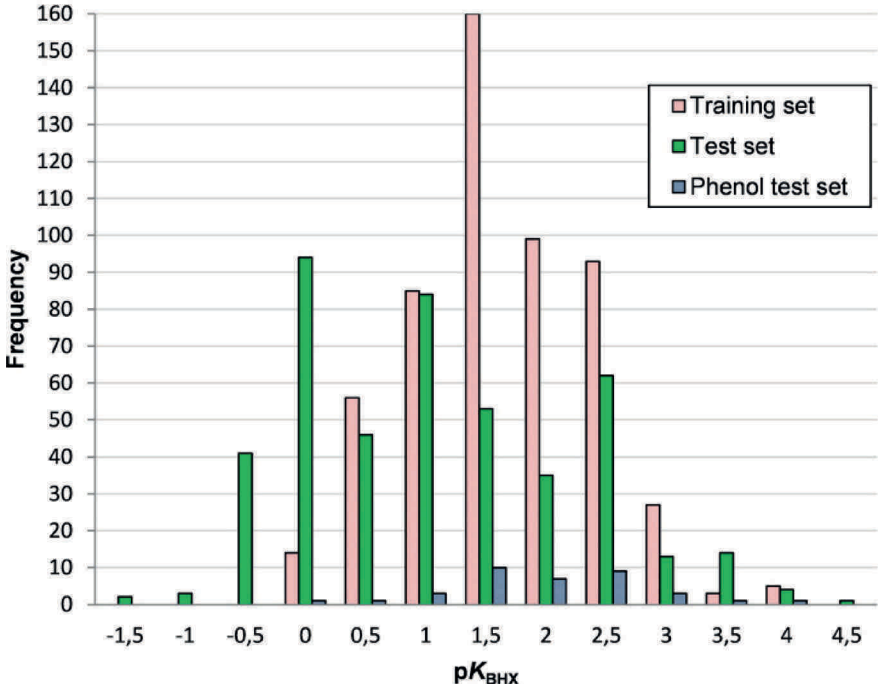


Figure 3. Experimental pK_{BHX} values distribution for the training set (red), the test set (green) and the phenol set (blue).

varies from −0.37 to 3.66 (Figure 3), whereas the size varies from 1 to 24 heavy atoms. Only direct experimental pK_{BHX} values were selected for the training set, using the exact same protocol for measurement. Thus, the training data are expected to have low internal errors, thereby ensuring the quality of data before the actual modelling. Random experimental error was evaluated by the experimentalists to be 0.04 log units.[7]

The training set is thus composed of a subset of 537 compounds (532 mono- and 5 bi-functional, in total 542 pK_{BHX} values) corresponding to direct and non-approximated experimental pK_{BHX} values. The remaining compounds, for which pK_{BHX} values were considered as less reliable, constituted an external test set (*test set*) of 451 values. The pK_{BHX} values in the latter were either corrected (when several equivalent acceptors' atoms are present) or approximated (for low soluble compounds, extreme values or values estimated from a measure of the IR shift of the donor's OH bond). Some compounds correspond to acceptor types not present in the training set (e.g. aromatic or alkene fragments, sulfates or halogens). These problematic molecules provide a good challenge to the models applicability domain. This set also includes 47 polyfunctional molecules with 2 non-equivalent acceptor sites.

One other supplementary external test set of 36 acceptors with phenol[38,39] (*phenol set*) (see Supporting information Section 3 Table SI4) was collected from the literature.

The labelling of the acceptor sites in the training and test sets have been performed manually according to database annotations. However, an automatic detection of acceptor sites has been implemented in the program to screen virtual libraries with developed models.

2.2 Descriptors

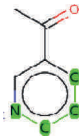
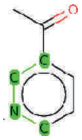
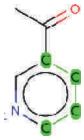
Two types of descriptors were used in this study: ISIDA fragment descriptors[30,40–42] and classical molecular descriptors generated by the Molecular Operating Environment (MOE) 2011 program.[43]

ISIDA fragment descriptors[30,40–42] represent subgraphs of a molecular graph. Each unique subgraph is considered as a descriptor, whereas its occurrence is used as the descriptor's value. The atoms of the molecular graph may be represented as their elements symbols, also called atom symbols, or forcefield typing according to the consistent valence forcefield[44] as other properties including pharmacophoric flagging and partial charge-based bins.[41]

Once, the molecular graph has been represented with the desired property, it is then segmented using a particular fragmentation scheme: sequences and atom-centred fragments of varying length. The minimum length of fragments varied from 2 to 4 and the maximum length from 2 to 15. By default, the algorithm searches for the shortest possible path between two atoms but the all path exploration option has also been tried. The atom pair option was also conversely used and consists of representing the extremities of the fragment and the length of the path between them.

Our working hypothesis is that hydrogen-bonding acceptor strength is chiefly influenced by the accepting atom,

Table 1. Examples of sequence descriptors in the different classes of different sequence paths of length 4 in 3-acetyl pyridine with the N as marked atom. If the path is not represented in the description, the field is left empty.

					
MA0 – No marked atom, all fragments	N*C*C–C	N*C*C*C	C*N*C*C	C*C*C*C	O=C–C*C
MA1 – only fragments beginning with the marked atom	N*C*C–C	N*C*C*C			
MA2 – only fragments containing the marked atom	N*C*C–C	N*C*C*C	C*N*C*C		
MA3 – all fragments with a special flag on the marked atom	N&MA&*C*C–C	N&MA&*C*C*C	C*N&MA&*C*C	C*C*C*C	O=C–C*C

the heavy donor atom and the nature of their environment. For the dataset used in this work, the structure of the HBA is the only changing factor which influences the variation in pK_{BHX}. Therefore, we prepared special types of ISIDA fragment descriptors containing marked atoms (MA) which explicitly indicate the acceptor's atom position (Figure 1). In such a way, information about both acceptor centre and its environment is encoded.

Different marked atom strategies were considered (Table 1):

- No marked atom – all fragments are generated (MA0).
- Sequences start with the marked atom or the central atom of atom-centred fragments is the marked atom (MA1).
- Only fragments containing the marked atom are generated (MA2).
- A flag (&MA&) is added to the symbol of the marked atom and all fragments are generated (MA3).

A total of 1260 descriptor families were generated for each of the 4 tested marking strategies using either the ISIDA Fragmentor[45] or ISIDA/QSPR[46] programs. Each family contains from 10 to 10000 fragment descriptors. The labelling of the acceptor sites in the training and test sets have been performed manually according to database annotations. Notice that MA1 and MA2 descriptors were already suggested in our previous work.[30] It should also to be noted that descriptors centred on selected atoms (different from those suggested in this work) were earlier applied in QSAR modelling of the dissociation constant (p*Ka*).[47–49]

2.3 MOE Descriptors

MOE descriptors were considered for the purpose of comparison. They represent a collection of 181 2D molecular descriptors, computed with the MOE 2011 program.[43] These descriptors describe physical properties, van der Waals surface area (the subdivided surface areas), the atom and bond counts (subdivided according to various criteria), the Kier and Hall chi connectivity and kappa molecular shape indices, the distance and adjacency matrices (Balaban's connectivity topological index, Wiener path number, Wiener polarity number), the pharmacophore atom types, and partial charges. All the hydrogen atoms were explicitly represented in the structures for the partial charge calculations.

2.4 Building and Validation of Models

Models were built and validated using support vector machines (SVM) with the LibSVM package[50] and MLR with the ISIDA/QSPR program.[46] Validation of models was carried out using cross-validation procedures (CV).[51] The determination coefficient (R^2), the correlation coefficient (R^2_{corr}) and the root mean squared error were used to evaluate the model ability to reproduce quantitatively the experimental data for training ($Y=Y_{calc}$) and test ($Y=Y_{pred}$) sets:

$$R^2 = 1 - \frac{\sum (Y_{exp} - Y)^2}{\sum (Y_{exp} - < Y_{exp} >)^2} \quad (1)$$

$$R^2_{corr} = \frac{\sum \{(Y_{exp} - < Y_{exp} >)(Y - < Y >)\}}{\sqrt{\sum (Y_{exp} - < Y_{exp} >)^2 \sum (Y - < Y >)^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum (Y - Y_{exp})^2}{n}} \quad (3)$$

where Y_{calc}, Y_{pred} and Y_{exp} are fitted, predicted and experimental values (here, $Y \equiv pK_{BHX}$) and summations are performed on all instances in the test set.

Ensemble modelling[52] implies the generation of many QSPR models, the selection of the most relevant ones and followed by their joint application to test compounds. For each compound from the test set, the program applies a *consensus model (CM)*, i.e., computes the property as an average of estimated values obtained with an ensemble of the models selected at the training stage. In the ISIDA/QSPR program,[46] outlying predictions of some models are excluded according to Tompson's rule. The standard deviation associated with this averaging can be used as a trustworthiness criterion: reliable predictions correspond to small standard deviations thus demonstrating that most of the models converge toward one same value.[53,54]

Any individual model entering the consensus model should be significantly better than *y*-scrambled models.[55] Therefore, for each of the employed machine learning techniques (SVM and MLR, see below) models' performances on 20 *y*-scrambling experiments were used to fit a normal distribution and a model was accepted only if its cross-validated performance was better than the 95th percentile of this distribution.

The SVM calculations were performed with epsilon support vector regression and a linear kernel using the LibSVM package.[50] Epsilon was optimised by fully exploring 10 values between 0.05 and 0.19 for the MA1 strategy. It was then set at 0.09 for the remaining of the study. The cost parameter was scanned (on a log scale) by 28 values ranging from 0.1 to 100. Each model was validated using a 5-fold cross-validation (5CV) with the inbuilt procedure of LibSVM with random splitting into learning/left-out subsets, which was reiterated 5 times in order to obtain robust average CV statistics. Models with 5CV $RMSE \leq 0.29$ pK_{BHX} units entered the CM. The *p*-value of the worse selected model compared to the *y*-scrambling performances was less than 0.001, thus far better than the minimum requirement to enter the CM. Only one SVM model (i.e. one cost parameter in this case) for each descriptor space was selected and then rebuilt on the entire training set to enter the CM.

Linear regression models were obtained with the ISIDA/QSPR software.[46] The individual MLR models were built by combining forward[56] and backward[57] stepwise variable selection techniques. In our calculations, many MLR models were generated combining different types of fragment descriptor and different variable selection algorithms. The number of generated individual models varied from 240 to 720 as a function of the descriptors used. Models with a leave-one-out CV determination coefficient $R^2 > 0.8$ were accepted for CM. All selected individual models performed significantly better than the scrambled models. Models were rebuilt on the entire training set for the CM.

Both SVM and MLR consensus models built on the entire training set were validated on two external tests sets (see Figure 2).

2.5 Applicability Domain (AD)

Generally, the AD[40] of the model defines an area of chemical space (basically, the one being densely covered by training set examples) where the model is presumably accurate. Two types of AD definitions were used simultaneously in this study: (i) *Fragment control* which consists in discarding predictions of test compounds containing fragments not occurring in the training set; (ii) *Bounding box* which considers AD as a multidimensional descriptor space confined by minimal and maximal values of occurrences of the descriptors involved in an individual model.

The applicability of a consensus model relies on the fraction of applicable individual models (i.e. the models for which AD does not discard the given molecule). If this number is lower than a threshold, the overall CM prediction is ignored.[58] In the ISIDA/QSPR software, by default, the threshold is 15% of the total number of models in the CM. For SVM, by default, the threshold is fixed at 50%.

3 Results and Discussion

3.1 Benchmarking of the Different Marked Atom Strategies

Individual SVM and MLR models were built using the different MA strategies and a Student's *t*-test was applied to the 5CV results to compare them. Average 5CV RMSE of the best models for each strategy are summarised in Table 2 and in Supporting Information Table SI 5. A *t*-test indicated that MA3 and MA2 descriptors are not significantly different at a confidence interval of 95%, while they are significantly different to MA1 and MA0 descriptors. Best models involving MA3 descriptors perform well in 5CV, both in SVM ($RMSE = 0.27$) and in MLR CM ($RMSE = 0.24$). The individual SVM models and the MLR CMs confirm the relevance of marked atom especially compared to 2D MOE descriptors ($RMSE = 0.40$, see Table 2).

The first proposed approach MA0 does not pinpoint the hydrogen-bonding site and generically describes the molecule, similarly to MOE terms. MA1 and MA2 describe the immediate surroundings of the acceptor site. The last approach, MA3, can be seen as a combination of the two different points of view mentioned previously. It encompasses the whole molecule but adds the information of the HBA so that the machine learning procedures can differentiate atoms of the same type participating or not in hydrogen bonding.

Table 2. Predictive performances of the models in 5-fold cross-validation involving the different marked atom strategies and MOE descriptors without AD.

Descriptors	Best individual SVM models		MLR CM	
	5CV-*RMSE*	5CV-R^2	5CV-*RMSE*	5CV-R^2
ISIDA MA0	0.33	0.80	0.32	0.82
ISIDA MA1	0.31	0.82	0.28	0.86
ISIDA MA2	0.28	0.86	0.27	0.87
ISIDA MA3	0.27	0.87	0.24	0.90
2D MOE	0.40	0.71	0.41	0.70

MOE descriptors perform badly, but, surprisingly, MA0 performs well in SVM. This may be due to the small size of the organic molecules and the fact that, in most cases, only one easily identifiable acceptor is found. Thus, the machine learning seems able to identify some substructures related to the surroundings of the acceptor. Eventually, since the MA3 strategy seemed to perform better in general, it was preferred for ensemble modelling in SVM and in MLR for the prediction of the external test sets.

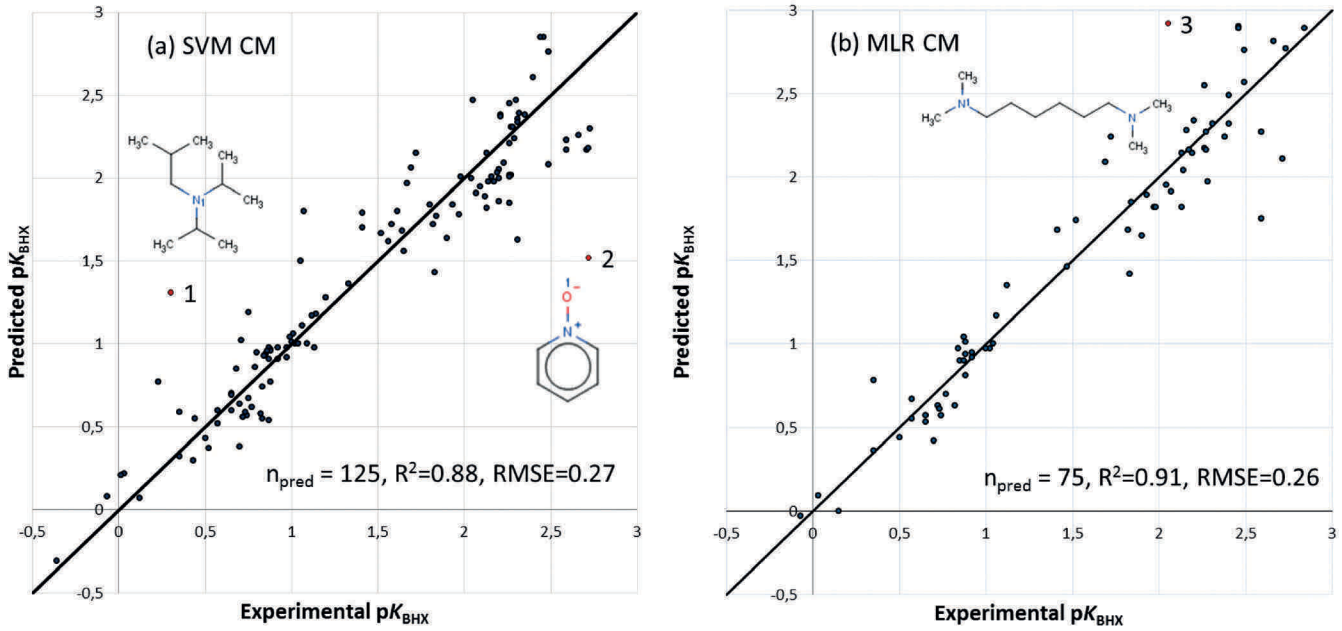


Figure 4. Predicted vs experimental pK_{BHX} values of the test set taking into account fragment control and bounding box as AD by (a) the CM SVM model where a minimum of 14 applicable models were required and (b) the CM MLR model. For the outliers 1–3, models accepted by AD greatly diverge thus showing low trustworthiness of predictions.

Notice that AD does not significantly change the models performance. Thus, MLR CM involving MA3 descriptors achieves and $RMSE = 0.22$ and $R^2 = 0.91$ on 85% of the data within AD.

3.2 Ensemble Modelling

In total, 27 models were selected for the SVM CM (see Supporting Information Table SI 6) and 306 for the MLR CM. The individual models were rebuilt on the entire training set using the sets of descriptors corresponding to the best individual models selected at the cross-validation step. Corresponding SVM and MLR CMs were applied to the external test set and the phenol set. Predicted pK_{BHX} values were assessed taking into account the fragment control and bounding box AD approaches, as well as the number of applied models.

Both SVM and MLR consensus models perform well on the test set: $RMSE = 0.29$ (SVM) and 0.26 (MLR), see Table 3 and Figure 4. These results are consistent with those obtained in cross-validation (see Table 2). The SVM model based on MOE descriptors performs poorly on the test set even if the SVM/ISIDA CM model's AD is accounted for: $RMSE = 0.56$ (Table 3).

The second external test set used was on 36 log *K* values with phenol[38,39] instead of 4-fluorophenol as HBD. According to literature,[59] such measurements should be highly correlated to pK_{BHX}. The MLR CM predictions are indeed correlated with $R^2_{corr} = 0.86$ with AD and SVM CM achieves $R^2_{corr} = 0.92$ with AD (see Supporting Information section 3 and Table SI 4–5). Thus, predicted pK_{BHX} and log *K* measured with phenol are highly correlated confirming previous observations by Raevsky et al.[16]

Table 3. Performance on the test set of the SVM and MLR consensus models (CM) based on ISIDA descriptors and the individual SVM model based on MOE descriptors using the SVM CM AD.

Method	Test set (with AD)			
	N_{mod} [a]	n_{pred} [b]	*RMSE*	R^2
SVM/ISIDA CM	14	125	0.27	0.88
	27	48	0.25	0.88
MLR/ISIDA CM	46	75	0.26	0.91
SVM/MOE	–	125	0.56	0.49

[a] N_{mod} is the minimum number of models required in CM. [b] n_{pred} is the number of acceptor sites accepted by AD.

In view of the better coverage of the phenol test set and its performance of R^2 of 0.91 on the test set, the MLR approach is considered marginally better than the SVM approach. It is interesting to note that in the associated linear correlation $\log K_{pred}$ (p-FC_6H_4OH) $= -0.33 + 1.33$ $\log K_{exp}$ (C_6H_5OH), the slope is superior to 1, thus confirming that the acidity of 4-fluorophenol is larger than that of phenol. Moreover, this linear correlation allows to predict from the phenol H-bond acidity ($pK_{AHY} = 2.06$, defined as the H-bond donor ability towards *N*-methylpyrrolidinone)[60] a value of 2.41 for 4-fluorophenol in excellent agreement with the experimental data ($pK_{AHY} = 2.38$).

In order to figure out how the CM performance depends on the number of applicable individual models, a series of SVM CM calculations have been performed on the test set. Figure 5 demonstrates variations of *RMSE* and the number

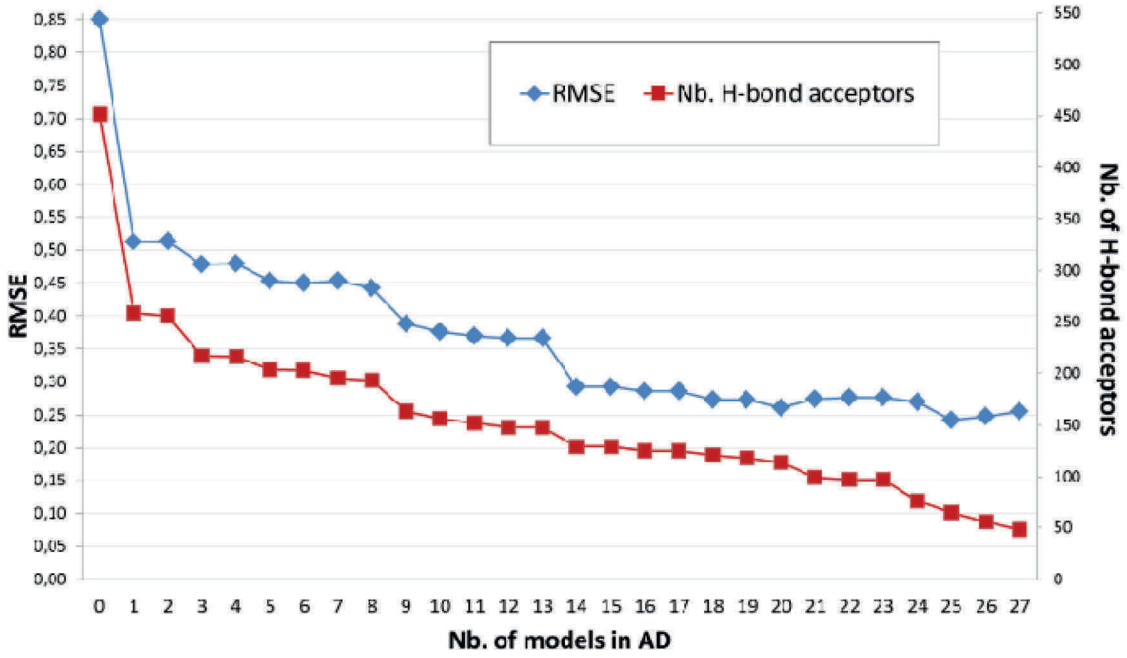


Figure 5. *RMSE* (blue diamond-shaped dots) and the number of H-bond acceptor centres (mono- and bifunctional) accepted by AD (red squares) by the SVM consensus model on the test set as a function of minimal number of applicable individual models

of accepted acceptors sites by AD as a function of the minimal number of applicable individual models (N_{mod}). At $N_{mod}=0$, the models' AD was not applied and predictions were made for all 451 acceptors sites. If, at least, one applicable model is applied ($N_{mod}=1$), the *RMSE* drastically drops because 40% of the test compounds are discarded. The increase of N_{mod} till 27 leads to further reduction of both *RMSE* (till 0.25) and the number of predicted acceptors sites (48). One can see that a good trade-off between performances and number of predictions is found at 14–15 applicable models, corresponding roughly to half of the entire ensemble of 27 models. This study shows the importance of considering AD which effectively discards the test set acceptor sites too dissimilar compared to the training set.

Table 4. Examples of bifunctional molecules predicted by the SVM and MLR models involving MA3 descriptors.

Origin	Training (5CV prediction)		Test set		Test set	
Acceptor site	1	2	1	2	1	2
Exp. pK_{BHX}	1.78	1.1	0.57	0.15	1.09	2.26
SVM CM	1.88	1.30	0.60	0.14	1.00	1.85
MLR CM	1.90	1.31	0.66	0.00	[a]	1.50

[a] H-bond acceptor site filtered out by AD.

3.3 Bifunctional Molecules

Five bifunctional molecules are found in the training set and 47 in the test dataset. For the training set compounds, performance of the models was assessed in cross-validation. Without surprise, MOE and MA0 descriptors were not able to distinguish different acceptor sites of the same molecules, and therefore, they could not represent the bifunctional cases. It should be noted that MOE models correctly predicted one of the two sites. On the other hand, models based on MA2 or MA3 descriptors did not only provide individual predictions for different binding sites, but correctly ordered them (see Table 4 and supporting information Table SI 2).

Among the 47 bifunctional molecules of the test set (94 acceptor sites), 20 molecules contain acceptor types not present in the training set: aromatic or unsaturated moieties (forming H···π-bonds) and halogens (forming H···Hal bonds). All these irrelevant sites were removed by AD. The 27 remaining bifunctional molecules were considered to evaluate the models' performances.

For these 27 bifunctional compounds, the SVM CM with AD where a minimum of 14 models is required in CM leads to only 14 predicted values out of 54. The predictive performance (*RMSE*=0.25 and $R^2=0.85$) is statistically compatible with that observed in cross-validation and the overall prediction of the test set. The poor coverage of the model was expected since there were very few bifunctional com-

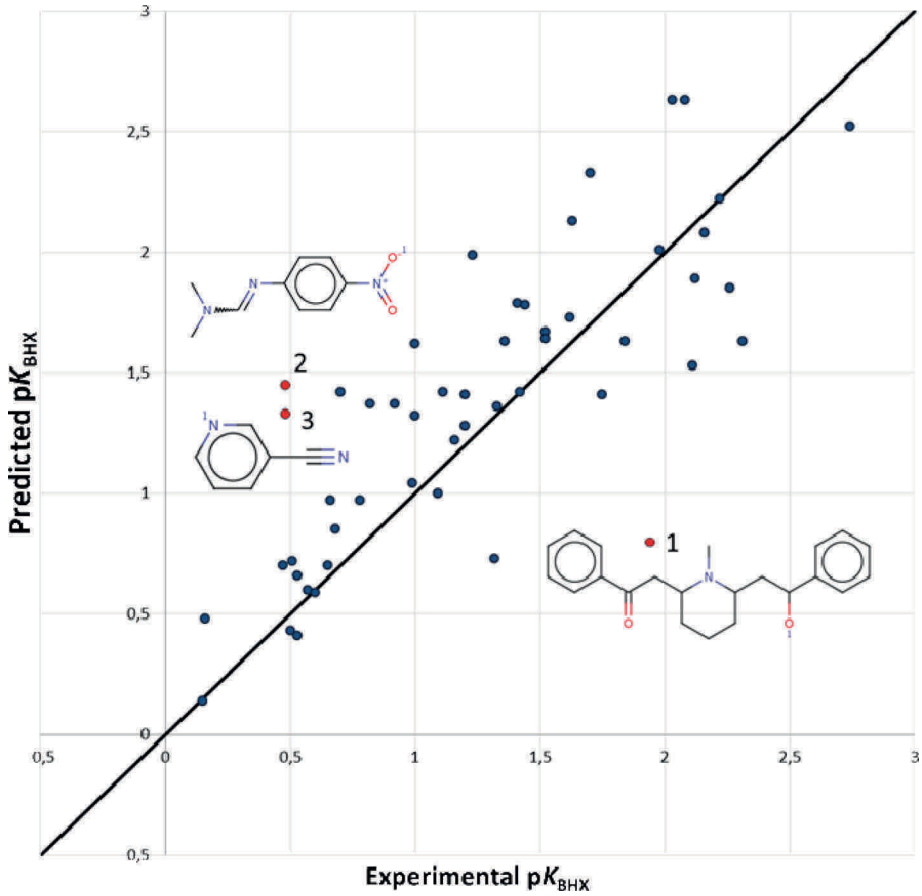


Figure 6. Prediction of 25 bifunctional molecules (51 values) by the SVM CM with the bounding box and fragment control as AD. Notice that three pK_{BHX} values for three acceptor sites were not predicted. The three structures shown correspond to the biggest prediction errors found.

pounds in the training set. The point is that, the substructural signature of most bifunctional species was never observed during training and those compounds were considered as out of AD. However, if the molecules are predicted when they are within the AD of at least one model, the *RMSE* is 0.44 which is still reasonable (see Supporting Information Section 2). Also note that these molecules were all indicated in the database as "approximated data". It can thus be assumed that the experimental error is greater on those molecules. The most pronounced outliers (see Figure 6) could be explained by either a small number of models applied for predictions or by inexact experimental data. For the most outlying molecule (indicated as 1 in Figure 6), experimental data has been measured with 4-nitrophenol[33] instead of 4-fluorophenol and then converted into pK_{BHX} by means of a LFER by Abraham et al.[59] Notice that this molecule can form an intra-molecular H-bond.

3.4 Comparison with Previously Reported Models

One can hardly compare the performance of our models with that of LFER models of Abraham and Raevsky because the latter were not properly validated on a reasonably large external test set. As far as models based on quantum mechanics calculations is concerned, the best published model[7] achieves a *RMSE* = 0.13 (calculated from the data reported in the Literature[7]). This model has been built on nitrogen HBAs restricting its application to the limited class of nitrogen compounds, despite the large panel of chemical functional groups encountered and the consideration of polyfunctional nitrogen structures. Our models perform slightly less good in cross-validation (*RMSE* = 0.25–0.27, Table 2) and on the external test sets (*RMSE* = 0.25–0.29, Table 3), but they are able to treat diverse sets of HBAs including nitrogen, oxygen, carbon, halogens, sulfur bases and the polyfunctional species with different atomic acceptor sites. They are clearly much less time-consuming because generation of fragment descriptors is a very fast procedure unlike heavy quantum mechanics calculations.

3.5 On-Line Implementation of the Models

The SVM consensus model is freely available on our web server: http://infochim.u-strasbg.fr/webserv/VSEngine.html

(for more details see Supporting Information section 1). Two different approaches are supported. One provides an automatic detection of acceptor sites which is performed following the intrinsic ChemAxon "acceptor" type in pharmacophore mapping (PMapper).[61] This allows the user to submit plain, unmarked molecular files. A synthetic trust criterion of the prediction is provided, taking into account various aspects such as the number of individual models including the compound in their AD and the standard deviation of values predicted by individual models. For example, this tool has been tested in a screening of a database containing 2470 drugs and reference compounds from the US Pharmacopeia.[62] 10215 acceptor centres were identified and ~1700 centres (some 17% of the whole database) where predicted within the fragment control AD. The relatively small prediction rate can be explained by the fact that studied pharmaceutical compounds are more complex than H-acceptor molecules present in the training set.

Alternatively, specification of expected acceptor sites may be left in charge of the user, which may submit marked molecular files to the model. This enables the user to predict hydrogen bonding acceptor strength of centres not being considered as such by ChemAxon's PMapper.

4 Conclusions

This study presents new marked atom strategies for ISIDA fragment descriptors. This development was motivated by the need to better represent the locality of the H-bonding interaction as reported in the pK_{BHX} database.

The individual and consensus models built on large and structurally diverse data set of 542 H-bond acceptors were intensively validated both in cross-validation procedure and on two external datasets. The resulting models perform well both in cross-validation ($RMSE = 0.24–0.27$) and on the external test sets ($RMSE = 0.25–0.29$). Besides previously reported results, we demonstrate that our models based on marked atom descriptors are able to predict pK_{BHX} values in bifunctional molecules containing two different H-bond atomic sites, and, in principle, could treat polyfunctional species.

The SVM consensus model is publically available on the server: http://infochim.u-strasbg.fr/webserv/VSEngine.html. Only an internet access and browser is required to execute the models; no software installation is needed.

Supporting Information

The supporting information consists of two files: one for structures and pK_{BHX} and one for modelling details.

Acknowledgements

We thank the *Centre of High-Performance Computing* of the Informatics Department of the University of Strasbourg (France) for the computational facilities.

References

[1] G. Gilli, P. Gilli, *The Nature of the Hydrogen Bond: Outline of a Comprehensive Hydrogen Bond Theory*, Oxford University Press, Oxford, **2009**.
[2] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, D. J. Nesbitt, *Pure Appl. Chem.* **2011**, *83*, 1619–1636.
[3] C. Bissantz, B. Kuhn, M. Stahl, *J. Med. Chem.* **2010**, *53*, 5061–5084.
[4] C. Laurence, K. A. Brameld, J. Graton, J.-Y. Le Questel, E. Renault, *J. Med. Chem.* **2009**, *52*, 4073–4086.
[5] J. Graton, M. Berthelot, J.-F. Gal, C. Laurence, J. Lebreton, J.-Y. Le Questel, P.-C. Maria, R. Robins, *J. Org. Chem.* **2003**, *68*, 8208–8221.
[6] M. H. Abraham, P. P. Duce, D. V. Prior, D. G. Barratt, J. J. Morris, P. J. J. Taylor, *Chem. Soc., Perkin Trans. 2* **1989**, *10*, 1355–1375.
[7] F. Besseau, J. Graton, M. Berthelot, *Chem. Eur. J.* **2008**, *14*, 10656–10669.
[8] O. Lamarche, J. A. Platts, *Chem. Eur. J.* **2002**, *8*, 457–466.
[9] Intermolecular Interactions: From Diatomics to Biopolymers (Ed: B. Pullman), Wiley, New York, **1978**.
[10] P. A. Kollman, L. C. Allen, *J. Am. Chem. Soc.* **1971**, *93*, 4991–5000.
[11] P. A. Kollman, J. McKelvey, A. Johansson, S. Rothenberg, *J. Am. Chem. Soc.* **1975**, *5*, 955–965.
[12] A. E. Kerdawy, C. S. Tautermann, T. Clark, T. Fox, *J. Chem. Inf. Model.* **2013**, *53*, 3262–3272.
[13] R. W. Taft, D. Gurka, L. Joris, P. v. R. Schleyer, J. W. Rakshys, *J. Am. Chem. Soc.* **1969**, *91*, 4801–4808.
[14] M. H. Abraham, P. L. Grellier, D. V. Prior, R. W. Taft, J. J. Morris, P. J. Taylor, C. Laurence, M. Berthelot, R. M. Doherty, M. J. Kamlet, J.-L. M. Abboud, K. Sraidi, G. Guiheneuf, *J. Am. Chem. Soc.* **1988**, *110*, 8534–8536.
[15] O. A. Raevsky, V. Y. Grigor'ev, V. P. Solov'ev, *Khim. Farm. Zh. (Russ.)* **1989**, *23*, 1294–1300.
[16] O. A. Raevsky, *J. Phys. Org. Chem.* **1997**, *10*, 405–413.
[17] O. A. Raevsky, V. Y. Grigor'ev, D. B. Kireev, N. S. Zefirov, *Quant. Struct.–Act. Relat.* **1992**, *11*, 49–63.
[18] R. S. Drago, B. B. Wayland, *J. Am. Chem. Soc.* **1965**, *87*, 3571–3577.
[19] A. V. Iogansen, *Teor. Eksp. Khim.* **1971**, *7*, 302–311.
[20] V. A. Terent'ev, *Thermodynamics of Hydrogen Bond*, Saratov University, Kuibyshev, **1973**.
[21] A. D. Sherry, K. F. Purcell, *J. Phys. Chem.* **1970**, *74*, 3535–3543.
[22] M. K. Kroeger, R. S. Drago, *J. Am. Chem. Soc.* **1981**, *103*, 3250–3262.
[23] O. A. Raevsky, V. V. Avidon, V. P. Novikov, *Khim. Farm. Zh. (Rus.)* **1982**, *16*, 968–971.
[24] O. A. Raevsky, V. Y. Grigor'ev, V. P. Solov'ev, I. V. Martynov, *Dokl. Akad. Nauk SSSR (Russ.)* **1988**, *298*, 1166–1169.
[25] J.-Y. Le Questel, M. Berthelot, C. Laurence, *J. Chem. Soc., Perkin Trans. 2* **1997**, 2711–2717.
[26] M. Hennemann, T. Clark, *J. Mol. Model.* **2002**, *8*, 95–101.

[27] A. S. Özen, F. D. Proft, V. Aviyente, P. Geerlings, *J. Phys. Chem. A* **2006**, *110*, 5860 – 5868.
[28] A. J. Green, P. L. A. Popelier, *J. Chem. Inf. Model.* **2014**, *54*, 553 – 561.
[29] A. Klamt, J. Reinisch, F. Eckert, J. Graton, J.-Y. Le Questel, *PhysChemChemPhys* **2013**, *15*, 7147 – 7154.
[30] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comput. Aid. Mol. Des.* **2005**, *19*, 693 – 703.
[31] J.-Y. Le Questel, G. Boquet, M. Berthelot, C. Laurence, *J. Phys. Chem. B* **2000**, *104*, 11816 – 11826.
[32] V. Arnaud, J.-Y. Le Questel, M. Mathé-Allainmat, J. Lebreton, M. Berthelot, *J. Phys. Chem. A* **2004**, *108*, 10740 – 10748.
[33] A. Locati, M. Berthelot, M. Evain, J. Lebreton, J.-Y. Le Questel, M. Mathé-Allainmat, A. Planchat, E. Renault, J. Graton, *J. Phys. Chem. A* **2007**, *111*, 6397 – 6405.
[34] V. Arnaud, M. Berthelot, F.-X. Felpin, J. Lebreton, J.-Y. Le Questel, J. Graton, *Eur. J. Org. Chem.* **2009**, 4939 – 4948.
[35] A. P. Atkinson, E. Baguet, N. Galland, J.-Y. Le Questel, A. Planchat, J. Graton, *Chem. Eur. J.* **2011**, 11637 – 11649.
[36] *JChem* Version 5.3.8, Calculator Plugin, ChemAxon, **2012**.
[37] V. Solov'ev, A. Varnek, *SDF manager EdiSDF* (Editor of Structure-Data Files), **2013**, http://infochim.u-strasbg.fr/spip.php?rubrique52, or, http://vpsolovev.ru/programs/.
[38] O. A. Raevsky, V. P. Solov'ev, V. Y. Grigor'ev, *VINITI Depos. N 1001-V88*, Moscow, **1988**, p. 83.
[39] O. A. Raevsky, A. F. Solotnov, V. P. Solov'ev, *Zhurnal Obshchei Khimii (Rus)* **1987**, *57*, 1240 – 1243.
[40] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, S. Gaudin, P. Vayer, V. P. Solov'ev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191 – 198.
[41] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855 – 868.
[42] V. P. Solov'ev, A. Varnek, G. Wipff, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847 – 858.
[43] *MOE*, Chemical Computing Group Inc., Montreal, **2011**.
[44] A. T. Hagler, E. Huler, S. Lifson, *J. Am. Chem. Soc.* **1974**, *96*, 5319 – 5327.
[45] *ISIDA Fragmentor2012*, Laboratoire de Chémoinformatique, UMR 7140, Université de Strasbourg, France, **2012**.
[46] V. P. Solov'ev, A. Varnek, ISIDA (*In Silico Design and Data Analysis*) *QSPR Program*, v5.76, Strasbourg – Moscow, **2012**; http://infochim.u-strasbg.fr/spip. php?rubrique53, or, http://vpsolovev.ru/programs/.
[47] L. Xing, R. C. Glen, R. D. Clark, *J. Chem. Inf. Model.* **2003**, *43*, 870 – 879.
[48] S. Jelfs, P. Ertl, P. Selzer, *J. Chem. Inf. Model.* **2007**, *47*, 450 – 459.
[49] M. Rupp, R. Körner, I. V. Tetko, *Mol. Inf.* **2010**, *29*, 731 – 740.
[50] C. C. Chang, C.-J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1 – 27.
[51] T. G. Dietterich, *Neural Comput.* **1998**, *7*, 1895 – 1923.
[52] K. M. Ali, M. J. Pazzani, *Machine Learning* **1996**, *24*, 173 – 202.
[53] D. Horvath, G. Marcou, A. Varnek, *J. Chem. Inf. Model.* **2009**, *49*, 1762 – 1776.
[54] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733 – 1746.
[55] C. Rücker, G. Rücker, M. Meringer, *J. Chem. Inf. Model.* **2007**, *47*, 2345 – 2357.
[56] A. Varnek, N. Kireeva, I. V. Tetko, I. Baskin, V. P. Solovev, *J. Chem. Inf. Model.* **2007**, *47*, 1111 – 1122.
[57] V. P. Solov'ev, A. A. Varnek, *Rus. Chem. Bull.* **2004**, *53*, 1434 – 1445.
[58] V. P. Solov'ev, A. Y. Tsivadze, A. A. Varnek, *Macroheterocycles* **2012**, *5*, 404 – 410.
[59] M. H. Abraham, P. L. Grellier, D. V. Prior, J. J. Morris, P. J. Taylor, *J. Chem. Soc. Perkin Trans. 2* **1990**, *0*, 521 – 529.
[60] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, D. J. Nesbitt, *Pure Appl. Chem.* **2011**, *83*, 1637 – 1641.
[61] *ChemAxon PMapper*, JChem v. 6.1.7 https://www.chemaxon.com/jchem/doc/user/PMapper.html, **2014.**
[62] *United States Pharmacopeia*, http://www.usp.org/, **2014**.

Received: March 12, 2014
Accepted: May 15, 2014
Published online: June 16, 2014

Bibliography

[1] Xing, L. and Glen, R. C. *J Chem Inf Comput Sci* **42**, 796–805 (2002).

[2] Polanski, J., Gieleciak, R., and Bak, A. *J Chem Inf Comput Sci* **42**(2), 184–191 (2002).

[3] Xing, L., Glen, R. C., and Clark, R. D. *J Chem Inf Model* **43**(3), 870–879 (2003).

[4] Zhang, J., Kleinöder, T., and Gasteiger, J. *J Chem Inf Model* **46**(6), 2256–2266 (2006).

[5] Ghasemi, J., Saaidpour, S., and Brown, S. D. *J Mol Struct Theochem* **805**(1-3), 27–32 (2007).

[6] Milletti, F., Storchi, L., Sforna, G., and Cruciani, G. *J Chem Inf Model* **47**(6), 2172–2181 (2007).

[7] Jelfs, S., Ertl, P., and Selzer, P. *J Chem Inf Model* **47**(2), 450–459 (2007).

[8] Lee, A. C., Yu, J. Y., and Crippen, G. M. *J Chem Inf Model* **48**(10), 2042–2053 (2008).

[9] Habibi-Yangjeh, A., Pourbasheer, E., and Danandeh-Jenagharad, M. *Monatsh Chem* **140**(1), 15–27 (2008).

[10] Harding, A. P., Wedge, D. C., and Popelier, P. L. A. *J Chem Inf Model* **49**(8), 1914–1924 (2009).

[11] Rupp, M., Körner, R., and Tetko, I. V. *Mol Inf* **29**(10), 731–740 (2010).

[12] Mussini, T., Covington, A. K., Longhi, P., and Rondinini, S. *Pure & Appl Chem* **57**(6), 865–876 (1985).

[13] Ehret, J. (2011). "QSPR prediction of acid dissociation constants, for first deprotonation of acids and alcohols", Master 2 project report.

[14] Serjeant, E. P. and Dempsey, B. *Ionization Constants of Organic Acids in Aqueous Solution*. Pergamon, Oxford, (1979).

[15] *ChemAxon pKa plugin version 5.3.8*, (2011).

[16] Solov'ev, V. P. and Varnek, A. (2011). ISIDA (In Silico Design and Data Analysis) QSPR program v5.76, Strasbourg - Moscow.

[17] Chang, C.-C. and Lin, C.-J. *ACM Trans Intell Syst Technol* **2**(3), 1–27 (2011).

[18] Tetko, I. V. (2011). ASsociate Neural Network (ASNN) program, http://www.vcclab.org/lab/asnn/.

[19] Marsh, A. and Altria, K. *Chromatographia* **64**(5-6), 327–333 (2006).

[20] Avdeef, A., Box, K., Comer, J., Gilges, M., Hadley, M., Hibbert, C., Patterson, W., and Tam, K. *J Pharm Biomed Anal* **20**, 631–641 (1999).

[21] Witten, I., Franck, E., and M.A., H. *Data Mining - Practical Machine Learning Tools and Techniques*. Elsevier, 3rd edition, (2011).

[22] *Weka v. 3.7.6, University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/*, (2012).

[23] Bonachera, F. and Horvath, D. *J Chem Inf Model* **48**(2), 409–425 Feb (2008).

[24] Krier, M., Bret, G., and Rognan, D. *J Chem Inf Model* **46**(2), 512–524 Jan (2006).

[25] Charbonnier, G. (2012). "Modélisation du pKa", Master 1 project report.

[26] ChemDB Soft (Chemistry Software Store), D. p. p., (2012). http://chemdbsoft.com/discon-pk-prediction.html.

[27] Ruggiu, F., Solov'ev, V., Marcou, G., Horvath, D., Graton, J., Le Questel, J.-Y., and Varnek, A. *Mol Inf* **33**(6-7), 477–487 (2014).

Chapter 6

Application of QSAR-based descriptors

6.1 Introduction

The idea to use increments issued from a QSAR model to mould descriptors is a form of inductive knowledge transfer, aka. transfer learning. In machine learning, it is characterised by the application of knowledge acquired while solving one problem to solve another one. In QSAR, multitask learning and features nets (FN) are two well-known forms of inductive knowledge transfer. In multitask learning, several problems are solved simultaneously and in FN, a model is build on a property and the predictions are used as descriptors. The use of increments in our case can be assimilated to a form of FN.

6.2 G Protein Coupled Receptors

In a previous study[1], GPCR affinity data had been used in a chemogenomics study using several descriptors and inductive knowledge transfer was shown to be an effective strategy to enhance prediction quality within this set. The present study is based on the data from this study.

6.2.1 Method

Data Chemogenomics data had been extracted from ChEMBL in the previous study[1] and sets of ligands for targets had been provided by Pr. Jürgen Bajorath, University of Bonn (Germany). Only the dopamine and serotonine receptors were used in this study (see Table 6.1). The Dopamine D2 receptor (T1) was used as the anchor receptor to produce the new descriptors because of its high number of ligands.

Descriptors ISIDA descriptors with atom symbols were used with the three different fragmentation:

- Sequences were calculated by alternating:

Table 6.1: List of GPCR targets with their ChEMBL ID and the number of available ligands (N_{lig})

Target(T)	Target name	ChEMBL ID	N_{lig}
1	Dopamine D2 receptor	217	1325
2	Serotonin 1a (5-HT1a) receptor	214	884
3	Serotonin 6 (5-HT6) receptor	3371	859
4	Dopamine D3 receptor	234	846
5	Serotonin 2a (5-HT2a) receptor	224	654
6	Serotonin 2c (5-HT2c) receptor	225	504
7	Dopamine D4 receptor	219	424
8	Serotonin 7 (5-HT7) receptor	3155	275
9	Dopamine D1 receptor	2056	272
10	Serotonin 2b (5-HT2b) receptor	1833	256
11	Serotonin 1d (5-HT1d) receptor	1983	139
12	Serotonin 1b (5-HT1b) receptor	1898	138
13	Dopamine D5 receptor	1850	98
14	Serotonin 5a (5-HT5a) receptor	3426	79
15	Serotonin 4 (5-HT4) receptor	1875	62

  - the use of bond information,
  - the minimum length between 2 and 4,
  - the maximum length between 2 and 10,
  - the use of atom pairs,
  - the use of all path exploration,
  - the use of formal charges

- Atom-centred fragments were calculated by alternating:
  - the use of bond information,
  - the minimum length between 2 and 3,
  - the maximum length between 2 and 4,
  - the use of atom pairs,
  - the use of formal charges

- Triplets were calculated by alternating:
  - the use of bond information,
  - the minimum length between 2 and 4,
  - the maximum length between 2 and 15,
  - the use of formal charges

In total, 445 descriptor spaces were generated for each target.

The best model issued from T1 was used to produce increments using the ISIDA ColorAtom program for all ligands from each target. Increments were visualised for each target and binning schemes were determined from those. Descriptors were calculated from the coloured graph using the same options as for atom symbols (see above). At the next stage, models for targets other than T1 were rebuild after adding the predicted T1 affinity (of the ligands of the training sets of other targets) as an additional descriptor, in order to simulate a feature net.

To summarize, three types of descriptor strategies were used:

- Atom symbols: descriptors used are based on atom symbols ISIDA descriptors
- Atom symbols + FN: The prediction by the T1 model is added to the atom symbols ISIDA descriptors
- QSAR-based: The increment-mapping issued from the T1 model is used to colour the graph. ISIDA descriptors are then calculated using that mapping.

Machine Learning and Validation Models were build using a genetic-algorithm with SVM. The genetic algorithm permits to search for good combinations of SVM parameters (kernel, cost and epsilon values) and descriptor spaces. The SVM is calculated using LibSVM[2]. For each target, the algorithm will consider all the 445 descriptor spaces. Models are validated with 3-fold CV and the corresponding determination coefficient 3-CV R^2_{det}.

6.2.1.1 Results and discussion

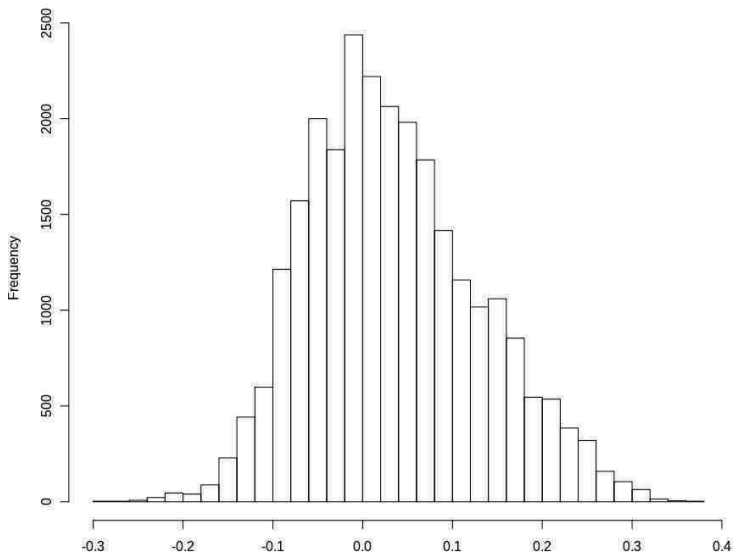


Figure 6.1: Distribution of the T1 QSAR model based increments on T2

The distributions of the increments do not present several peaks that would permit to divide them (see Figure 6.1). Therefore, the bins were made to have a central bin around

Table 6.2: Best 3-fold cross-validated determination coefficient 3-CV R^2_{det} for the atom symbols strategy on GPCRs. Corresponding target numbers are available in Table 6.1

Target	3-CV R^2_{det}
1	0.66
2	0.68
3	0.59
4	0.65
5	0.68
6	0.52
7	0.62
8	0.56
9	0.57
10	0.40
11	0.58
12	0.57
13	0.46
14	0.48
15	0.28

0 and divide the rest of the distribution symmetrically:

- Very positive > 0.1
- 0.1 $\geqslant$ Positive > 0.05
- 0.05 $\geqslant$ Slightly positive > 0.025
- 0.025 $\geqslant$ Neutral > -0.025
- -0.025 $\geqslant$ Slightly negative > -0.05
- -0.05 $\geqslant$ Negative > -0.1
- -0.1 $\geqslant$ Very Negative

Performances for the SVM models build with atom symbols descriptors are given in Table 6.2 as 3-CV R^2_{det}. No inductive knowledge transfer is observed with FN. The descriptors with the QSAR-based mapping do well in training but do not cross-validated (3-CV R^2_{det} values around 0.1-0.2).

Several reasons could explain such results. The first being the high resolution of the QSAR-based descriptors - the information might be too dissolved over the several bins. However, the inductive knowledge transfer is not observed which could be due to the target chosen being unsuitable for the task. In the initial study, all predictions from all targets had been used in FN. However, this strategy is hardly possible to integrate when producing a new colouration for the ISIDA descriptors. Another reason could be the poor performance of the used model. It does not seem to extract knowledge properly and hence, cannot help to bring better results for the other targets. Eventually, the use of atom symbols descriptors is not good. However, these were chosen as basic descriptors and

the idea was to start with descriptors not tuned for the problem (as pharmacophore-based descriptors would be) and improve on those.

6.3 Conclusion and Perspectives

The first study using QSAR-based increments to build new descriptors was a failure. However, the study was probably not well designed and is, thus, not very conclusive as to the usefulness of QSAR-based increments in the design of new descriptors. Another study of inductive knowledge transfer based on air-tissue partition coefficient[3] where the effect has been clearly established between the same properties for rats and humans could permit to have a better benchmark. The transformation from continuous data into categories to label the graph in order to calculate fragment descriptors isn't straightforward and information is probably lost within the binning process. Instead, I propose to sum up each atomic contributions over a fragment, to normalize it to the number of atoms it contains and take the result as the value of the molecular fragment descriptor. This would enable to skip the binning process which is rather unclear, particularly in the case of a smooth distribution as was the case in this study.

Bibliography

[1] Brown, J., Okuno, Y., Marcou, G., Varnek, A., and Horvath, D. *J Comput Aided Mol Des* **597**, 618 (2014).

[2] Chang, C.-C. and Lin, C.-J. *ACM Trans Intell Syst Technol* **2**(3), 1–27 (2011).

[3] Varnek, A., Gaudin, C., Marcou, G., Baskin, I., Pandey, A. K., and Tetko, I. V. *J Chem Inf Model* **49**(1), 133–144 (2009).

Chapter 7

Software Developments

7.1 Standardisation tool

The standardisation tool was done with Dragos Horvath in 2010. The script is in C-shell and uses ChemAxon's JavaClasses[1]. As advocated by Fourches et al.[2], different steps to curate the data have been integrated:

1. Calculation of the canonical SMILES[3] (a form to encode the structure of the molecules)
2. Removal of explicit hydrogen atoms
3. Dearomatisation of compounds (representation in Kekulé form)
4. Elimination of counterions for salts and solvents
5. Handling of stereochemistry: options permit to eliminate it, to keep it as is or to choose it manually.
6. Neutralisation of the groups except for quaternary ammonia
7. Standard representation for groups such as nitro
8. Calculation of the tautomers (an option permits to choose the major tautomer according to ChemAxon automatically)
9. Aromatisation of the compounds
10. Calculation of the major micro-species at a given pH (by default, the physiological pH 7.4)

Both the neutral form and the major micro-species will be outputted. It was integrated as part of our standard routine in our models on our web server[4]. The program may also be used interactively where the options to remove stereochemistry or not is left open. The program will also propose different tautomers to choose when relevant and how to aromatise the compounds when the different functions of ChemAxon give different results.

7.2 pKa assignment tool

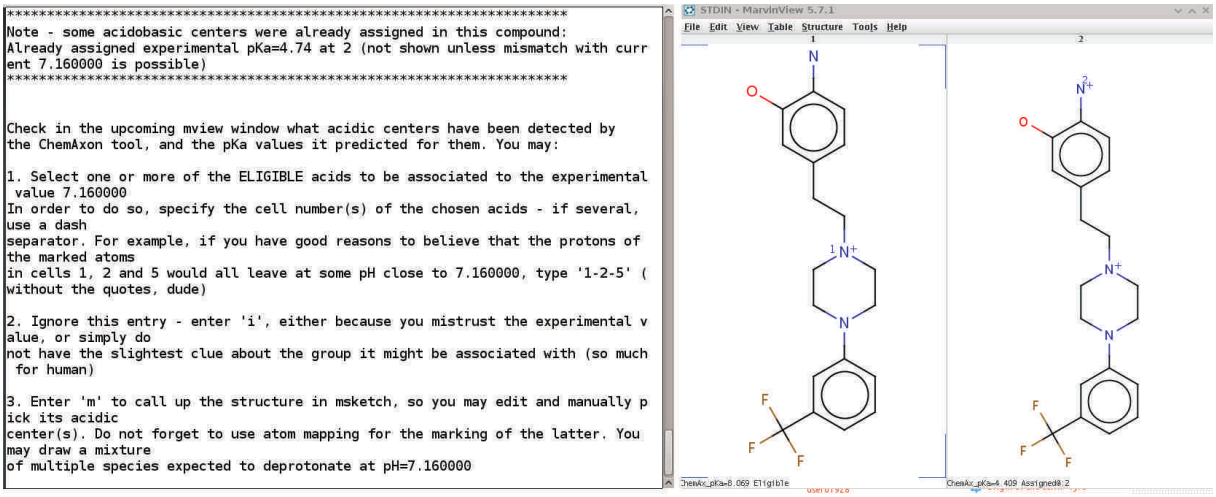


Figure 7.1: pK_a assignment script. The user is prompted to choose manually the acidic centre

The pK_a assignment script is based on ChemAxon's pK_a plugin[5] and was written with Dragos Horvath. The idea of the script is to automatically assign pK_a values to a particular functional group in the molecule based on ChemAxon predictions. In order to do so, the experimental pK_a value should be close to the predicted value. The pK_a tolerance factor in the script modulates the allowed discrepancy between the two values. The experimental pK_a value should not only have a close value to the prediction, but also no other potential protonation site should correspond to it. No other acidic centre should be found in a greater span around the experimental value. The ambiguity factor insures that potential alternative attributions are considered. The span is equal to the product of the ambiguity factor and the tolerance. The assignment is automatic if there is:

- One acidic centre within pK_a tolerance
- No other acidic centre within Ambiguity factor * tolerance

Otherwise, the script will prompt the user at the end to "resolve" these cases. A graphical menu will appear for each case were the experimental value is given and potential centres on different tautomers is proposed (see Figure 7.1. By default the pK_a tolerance factor is set at 1 and the ambiguity factor is set at 3. These parameters were found to be best to assign pK_a values on the CNE set (see 5.2.4)).

The script will seek for plausible acido-basic reactions in various tautomers or only in the major tautomer (user defined option) and start to scan from very low to very high pH values in the pK_a plugin.

7.3 Mapping script

A script based on ChemAxon's JavaClasses has been written with Dragos Horvath in order to generate the mapping/colouration of the molecular graph prior to fragmentation.

The script enables to map the pharmacophoric properties and the force field flags. The increment-based mapping are also included by calculating the increments and the bins are coded. The output can directly be given to the ISIDA Fragmentor. Instructions for installation and a manual is available in the ISIDA Fragmentor2013 manual (see Appendix A).

7.4 ISIDA descriptors software: from fragdesc to ISIDA Fragmentor2013

The fragmentation software in the laboratory has been evolving constantly. The first software was developed by Vitaly Solov'ev and was incorporated to the ISIDA/QSPR program. To test the new colourations, Dragos Horvath and myself programmed a simple fragmentation program, fragdesc, where only all paths exploration was available. Meanwhile, Gilles Marcou started a fresh fragmentor based on graph algorithms[6]. I contributed to this new fragmentor by implementing:

1. the handling of the colourations,
2. the handling of atoms with several flags,
3. the management of counts dependent on weights (for pH-dependent count),
4. a new molecule object (common to many parts of the ISIDA projects),
5. the update on the bond symbol system,
6. the dynamic bonds filters,
7. the marked atom strategies,
8. the triplet fragmentation

These different implementations were done over the years and 3 different versions were released: ISIDA Fragmentor2011, ISIDA Fragmentor2012 and ISIDA Fragmentor2013.

7.4.1 Possible extensions and perspectives

In order to be able to calculate the fuzzy pharmacophoric triplets, the fuzzy count could be implemented. It would however permit fuzziness for all types of fragmentations. A plan for this implementation and formats have already been proposed.

Another extension could be to handle user defined bond colouration and bond weights. A unit has been left open for the implementation.

As mentioned in section 6.3, the binning process to label the graph in order to calculate fragment descriptors isn't straightforward and information is probably lost. The ISIDA Fragmentor could be modified to read available increments. The descriptor's value would be equal to the normalised sum of the atomic contributions in the fragment.

Bibliography

[1] *ChemAxon Calculator Plugin, JChem v.5.7.1*, (2010).

[2] Fourches, D., Muratov, E., and Tropsha, A. *J Chem Inf Comput Sci* **50**(7), 1189–1204 JUN (2010).

[3] Weininger, D. *J Chem Inf Comput Sci* **28**(1), 31–36 Feb (1988).

[4] *Virtual Screening Plateform of the Chemoinformatics Laboratory, University of Strasbourg, France. http://infochim.u-strasbg.fr/webserv/VSEngine.html*, (accessed 2014).

[5] *ChemAxon pKa plugin version 5.3.8*, (2011).

[6] Lacomme, P., Prins, C., and Sevaux, M. *Algorithmes de edition, 2nd edition*. Eyrolles, second edition, (2003).

Chapter 8

Conclusion and Perspectives

8.1 Conclusion

1. The new ISIDA descriptors constitute a generalisation of already existing descriptors and permit the extraction of a vast range of chemical information. They allow to fill in the gaps between what were until now considered to be distinct descriptor categories ("strict" linear and circular fragment counts on one hand, "fuzzy" pair/multiplet counts on the other). They increase chances to build a good model for chemical phenomena which are by nature very complex. They can take into account certain particularities of a system through the marked atom strategy for polyfunctional molecules or by considering the pH for problems in solution. A greater number of diverse and relevant models permit to obtain better and more robust consensus models. These descriptors can adapt to a given problem by encoding the most relevant physico-chemical aspects and by tuning their numerous possibilities.

The ISIDA descriptors have on several occasions outperformed other choices of descriptors. Additionally, they are more readily interpretable as they refer to a chemical structure and models can be reduced to increments with the analysis of the fragments[1] (see 2.8). They have been successfully used in several structure-property relationships important to medicinal chemistry during my work but also in other projects by different researchers[2–5].

2. Different modelling carried out with the new descriptors show their advantages in comparison to other descriptors.

In similarity-based virtual screening, they augment considerably the number of relevant descriptor spaces. A neighbourhood behaviour benchmark[6] showed their good propensity on 5 proteases. It is notable that in comparison to the many other descriptors used, several IPLF descriptors performed always rather well. This was reflected by a low rank position and a small standard deviation of this rank overall the similarity-based virtual screening challenges performed in a sports tournament fashion.

They were also used successfully in several QSAR studies. hERG channel inhibition models were build on 562 molecules with pIC_{50} values[7] and tested on an external test set

of classification data from PubMed[8]. Models achieved a recall of 0.76 while the reference literature model[7] obtained a recall of 0.57 on the same external dataset.

Octanol-water partition coefficient (LogP) models were build on 3225 molecules issued from the Physprop database[9]. They were validated on an external test set of 9677 compounds from the same database and achieve a RMSE of 0.75, which surpasses ChemAxon's prediction (RMSE = 0.76 with respect to 9582 compounds as it fails to predict for 95 compounds). The models were compared to a previous benchmarking study[10] with 226 molecules. They slightly outperform other methods with a RMSE of 0.78 while the best method referenced obtained a RMSE of 0.80.

Within the interdisciplinary research project (PIR), the Chromatographic Hydrophobicity Index (CHI) measurements done by our colleagues permitted to build QSAR models. An algorithm was proposed to detect the outliers from these models. These outliers were tested and showed that experimental errors altered their CHI value. With this algorithm, we were able to pinpoint experimental problems for 21 compounds. These problems could not be detected during the experimental screening and they represented about 4% of the database. The dataset was cleaned and a consensus model was build on 485 compounds. It was tested on an external literature set and performed with a RMSE of 14.3 with the use of an applicability domain (AD) and a RMSE of 16.4 without AD. We suggest the use of QSAR modelling to control the quality of HTS experiments.

Within a collaborative chemogenomics project, the performances of IPLF descriptors surpass slightly those of other descriptors with an accuracy of 0.87 in comparison to 0.85 for ECPFs descriptors obtained in a previous study[11]. In particular, force field-based descriptors and certain pharmacophore-based descriptors have a better distribution of values inside the similarity matrices. The project enabled to ameliorate the available software, in particular in order to handle larger fingerprints.

By using the local descriptors, the first predictive QSAR models for the hydrogen-bond acceptor strength for polyfunctional molecules were developed. New marked atom strategies for ISIDA fragment descriptors were developed in order to better represent the locality of the hydrogen bonding interaction as reported in the pK_{BHX} database. The individual and consensus models were built using MLR and SVM on a large and structurally diverse data set of 542 hydrogen bond acceptors. The resulting models perform well both in cross-validation with a RMSE = 0.24–0.27 and on two external test sets, 451 values and 36 values, with a RMSE = 0.25–0.29. Besides previously reported results, we demonstrate that our models based on marked atom descriptors are able to predict pK_{BHX} values in bifunctional molecules containing two different hydrogen bond atomic sites, and, in principle, could treat polyfunctional species. In comparison to MOE descriptors, our descriptors performed better as they could take into account the locality of the property. The SVM consensus model is publicly available on our web server: http://infochim.u-strasbg.fr/webserv/VSEngine.html.

3. The IPLF descriptors alongside the neighbourhood behaviour study, the LogP and hERG QSAR models were published in Molecular Informatics in 2010[12]. The CHI study alongside the measurements for the PIR project and the outlier analysis were published

in Analytical Chemistry in 2014[13]. The local descriptors and the hydrogen bond acceptor strength study were published in Molecular Informatics in 2014[14].

4. Hydrophobicity and hydrogen bond acceptors strength models are freely accessible to the users through the laboratory's web server (http://infochim.u-strasbg.fr/webserv/VSEngine.html). The ISIDA descriptors were also implemented into the Mobyle portal, a French national project of the GDR Chemo/Bio-info. The ISIDA Fragmentor 2013 was implemented in the ochem platform (https://ochem.eu/) but only atom symbol based descriptors are available at the moment.

8.2 Perspectives

It remains to be seen whether other colouration schemes work on chemical problems. In particular, whether QSAR-based increments issued from the ISIDA ColorAtom program could be used for the description of molecules. In order to do so, a cleaner benchmark than the one tried could be done based on tissue-air partition coefficients data[15] where inductive knowledge transfer has been observed. The idea to replace the descriptor value by a normalised sum of the increments on the corresponding fragment could provide a solution to the difficulties encounter to bin the increments value.

The acidic dissociation constant project isn't finished. Initial results using the first marked atom strategy showed that global models could be built on them but their performances never reached those of published models or of the ChemAxon program. Our first assumption is that data was insufficient and, thus, a bigger dataset is needed. Several compounds from various sources still need to be cleaned, marked and standardised to their acidic form. A script has been developed to facilitate this task. Once this task is done, the other marked atom strategies should also be tried out on the prepared dataset. The data issued from the PIR could be used as a test set as they are a reliable source of data and challenging due to their diversity.

Bibliography

[1] Marcou, G., Horvath, D., Solov'ev, V., Arrault, A., Vayer, P., and Varnek, A. *Mol Inf* **31**, 639–642 (2012).

[2] Kireeva, N., Baskin, I. I., Gaspar, H. A., Horvath, D., Marcou, G., and Varnek, A. *Mol Inf* **31**(3-4), 301–312 (2012).

[3] Horvath, D., Marcou, G., and Varnek, A. *J Chem Inf Model* **53**(7), 1543–1562 Jul (2013).

[4] Ovchinnikova, S. I., Bykov, A. A., Tsivadze, A. Y., Dyachkov, E. P., and Kireeva, N. V. *J Cheminform* **6**, 20–38 (2014).

[5] Brown, J., Okuno, Y., Marcou, G., Varnek, A., and Horvath, D. *J Comput Aided Mol Des* **597**, 618 (2014).

[6] Horvath, D., Koch, C.and Schneider, G., Marcou, G., and A., V. *J Comput Aided Mol Des* **25**, 237–252 Jan (2011).

[7] Li, Q., Steen Jørgensen, F., Oprea, T. I., Brunak, S., and Taboureau, O. *Mol Pharm* **5**(1), 117–127 (2008).

[8] *NIH The PubChem Project, http://pubchem.ncbi.nlm.nih.gov/*, (2010).

[9] *SRC PHYSPROP database, http://www.srcinc.com/what-we-do/product.aspx?id=133&terms=Physprop*, (2009).

[10] Mannhold, R., Poda, G., Ostermann, C., and Tetko, I. *J Pharm Sci* **98**(3), 861–893 Mar (2009).

[11] Brown, J., Niijima, S., Shiraishi, A., Nakatsui, M., and Okuno, Y. In *IEEE International Conference on Bioinformatics and Biomedicine Workshops,* Gao, J., Dubitzky, W., Wu, C., Liebman, M., Alhaij, R., Ungar, L., Christianson, A., and Hu, X., editors (, Philadelphia, PA, 2012).

[12] Ruggiu, F., Marcou, G., Varnek, A., and Horvath, D. *Mol Inf* **29**(12), 855–868 Dec (2010).

[13] Ruggiu, F., Gizzi, P., Galzi, J.-L., Hibert, M., Haiech, J., Baskin, I., Horvath, D., Marcou, G., and A., V. *Anal Chem* **86**, 2510–2520 (2014).

[14] Ruggiu, F., Solov'ev, V., Marcou, G., Horvath, D., Graton, J., Le Questel, J.-Y., and Varnek, A. *Mol Inf* **33**(6-7), 477–487 (2014).

[15] Varnek, A., Gaudin, C., Marcou, G., Baskin, I., Pandey, A. K., and Tetko, I. V. *J Chem Inf Model* **49**(1), 133–144 (2009).

Abbreviations

- aka: also known as
- AD: Applicability Domain
- ADME/T: Absorption, Distribution, Metabolism, Excretion and Toxicity
- ASNN: ASsocciative Neural Network (program)
- BA: Balanced Accuracy
- CATS: Chemically Advanced Template Search (descriptors)
- CE: Capillary Electrophoresis
- CGBVS: Chemical Genomic-Based Virtual Screening
- CHI: Chromatographic Hydrophobicity Index
- CM: Consensus Model
- CN: Chimiothèque Nationale, french national chemical library
- CNE: Chimiothèque Nationale Essentielle, subset of the french national chemical library
- CPI: Compound-Protein Interaction
- CPU: Central Processing Unit
- etc: et cetera (and so on)
- FN: Feature nets
- FPT: Fuzzy Pharmacophoric Triplets
- FXA : Factor Xa (serine protease)
- GPCR: G Protein-Coupled Receptors
- hERG: human Ether-à-go-go-Related Gene
- HPLC: High Pressure Liquid Chromatography
- HTS: High-Throughput Screening
- IC_{50}: half maximal inhibitory concentration ($pIC_{50} = -log(IC_{50})$)

- i.e.: id est (that is to say)
- IPLF: ISIDA Property-Labelled Fragment
- ISIDA: In SIlico design and Data Analysis
- LAOS: Local Ascertained Optimality Score
- LFER: Linear Free-Energy Relationship
- LogP: octonal-water partitioning coefficient
- MA: Marked Atom
- MAE: Mean Absolute Error
- MOE: Molecular Operating Environment (program)
- NB: Neighbourhood behavious
- PF: (ChemAxon) Pharmacophore Fingerprints
- PIR: Projet Interdisciplinaire de Recherche, Interdisciplinary Research Project
- pKa: inverse logarithm of the acidic dissociation constant
- QSAR: Quantitative Structure-Activity Relationship
- QSPR: Quantitative Structure-Property Relationship
- RAM: Random Accessed Memory
- RMSE: Root-Mean Squared Error
- RP HPLC: Reverse-Phase High Pressure Liquid Chromatography
- SMILES: Simplified Molecular-Input Line-Entry System
- SQS: Stochastic QSAR Sampler (program)
- SVM: Support Vector Machines
- SimVS: Similarity-based Virtual Screening
- uPA : Urokinase-type Plasminogen Activator

Appendices

Appendix A

ISIDA Fragmentor2013 Manual

ISIDA Fragmentor2013 - User Manual

Fiorella Ruggiu, Gilles Marcou,
Vitaly Solov'ev, Dragos Horvath, Alexandre Varnek

Contents

Chapter 1

Introduction

The ISIDA Fragmentor2013 is a development of the Laboratoire de Chémoinformatique, Chimie de la Matière Complexe (SMS UMR 7140), Université de Strasbourg, France. This program is a part of the ISIDA project, which stands for “In SIlico Design and data Analysis“ and aims to develop tools for the calculation of descriptors, the navigation in chemical space, quantitative structure-activity modeling (QSAR) and virtual screening. The ISIDA Fragmentor2013 calculates molecular fragment count descriptors from a Structure-Data File (SDF). It is based on a series of graph algorithm from the book ”Algorithmes de graphes“ [1].

The ISIDA descriptors have been described in 6 publications:

- ISIDA Substructural Molecular Fragments (SMF)[2, 3]
- ISIDA Fuzzy Pharmacophoric Triplets (FPT) [4, 5]
- ISIDA Property-Labelled Fragments (IPLF) [6]
- Individual hydrogen-bond strength QSPR modelling with ISIDA local descriptors: a step towards polyfunctional molecules [7]

ISIDA Fragmentor2013 is able to calculate SMF and IPLF (with a ChemAxon based java program: CA_Prop_Map2011) as described in the publications. You may also read our Nomenclature document to learn about ISIDA fragment descriptors which is available on our website (http://infochim.u-strasbg.fr/spip.php?rubrique49).

The laboratory uses the ChemAxon plugins to map a property on the graph. However, one of the aims of ISIDA Fragmentor2013 is to enable the use of any combination of options and let the user as much freedom as possible to fit his needs. Therefore, the ”coloration“ of the molecular graph can be user-defined - given the input format is respected.

The next Chapter describes all the options, input and output format description, installation and usage of Fragmentor2013 and the corresponding nomenclature of ISIDA fragment descriptors. Chapter 3 is dedicated to our ChemAxon-based property mapping program.

Chapter 2

Fragmentor2013

2.1 Command line

The ISIDA Fragmentor2013 is a command line only program. You may call upon it using:
PATH/Fragmentor -i <SDFile> -o <BaseName>
[-f <string> -s <string> -h <HeaderFile>]
-t <integer> [{-l <integer> -u <integer>}
-c <SDField> -m <(0,1,2,3)> -d <(0,1,2)>
- -DoAllWays - -AtomPairs - -UseFormalCharge - -StrictFrg]

Options in squared brackets are not mandatory and those in curly brackets are linked to one another. Options are quickly explained in the next section. It is best to keep the options as they are ordered above. In any case, longer options (with - -) should always follow the short ones (with only -).

One call to ISIDA Fragmentor2013 may include several different types of fragmentation. To do so, use several -t options (indicating the type of fragmentation) with the list of corresponding options. For example if you wish to obtain sequences of atoms and bonds ranging from 1 to 4 bonds, and augmented atoms with a distance up to 1 bond, you will use the following command:
PATH/Fragmentor -i input.sdf -o output -t 3 -l 2 -u 5 -t 10 -l 2 -u 2
Note: The numbers given as lower and upper lengths correspond to the number of atoms included in the sequences. If you wish to include atom counting to the previous command then use:
PATH/Fragmentor -i input.sdf -o output -t 3 -l 2 -u 5 -t 10 -l 2 -u 2 -t 0

Certain options cannot be used together or require another option:

- Atom-centered fragments (-t 4 to 9) are always shortest path - they cannot be used with the option - -DoAllWays.

- - -StrictFrg can only be used with the -h option to indicate the header file (.hdr). The outputed svm will be limited to the descriptors indicated in that header file and keeping the same order.
- Marked Atom option (-m) can only be set to 0 or 1 for Triplets calculation (-t 10).

Make sure your input Structure-Data File (SDF) is at the V2000 format, else it might generate errors, memory leaks or wrong fragmentations. Beware that ISIDA Fragmentor2013 does not check the input file before treating it!!

2.2 List of Options

- -i : Input Structure-Data File (SDF) name.
- -o : All output files will have this name and will differ only by their extensions.
- -f : Format of the output. By default SVM - SMF, SVM and CSV are available (see output formats in 2.4.2)
- -s : A substring identifying unambiguously a field name in the SDF. The value of the field will be considered as a property to be saved along with set of descriptors of each input compound. Missing values are replaced by "?".
- -h : Name of a header file. If present, the fragmentation will reproduce the list of fragments the header contains. The output header file will match this input concatenated with new fragments discovered at the end.
- -t : Fragmentation type. See below.
- -l : Minimal length of fragments as sequences - Note: a length of 2 corresponds to a sequence with 2 atoms
- -u : Maximal length of fragments as sequences
- -c : Indicate the field name (COLOR_NAME) in the SDF of your wished coloration. Should be of format:
  > <COLOR_NAME>
  5 1:P 2:H 4: A/D
  95 1:A 2:H 4:D
  where 5 and 95 are the count to be considered for each species and the following characters are Atom number: Colouration1/Colouration2

- -m : If set to 1: All fragments must begin or end by a marked atom. A marked atom is an atom that has a label in the 7th column of the atom block in the SDF file.
  If set to 2: All fragments containing the marked atom will be generated
  If set to 3: A special flag (&MA&) will be added to the marked atom. All fragments are present. (if set to 0, all molecular fragments will be generated - same as without the option)
- -d : If set to 1:When processing Condensed Graph of Reactions (CGRs), only those fragments containing a dynamic bond are kept while the others are discarded.
  If set to 2: When processing Condensed Graph of Reactions, only those fragments containing only dynamic bonds are kept while the others are discarded. (if set to 0, all molecular fragments will be generated - same as without the option)
- - -DoAllWays : If fragments are sequences, search for all paths connecting two atoms.
- - -UseFormalCharge : Charged atoms (column 5 in the SDF file) will be indicated by adding _FC"charge_value"_
- - -AtomPairs : All constitutional details of a sequence are removed and only the number of constitutive atoms is given.
- - -StrictFrg : Only fragments included in a header file defined by a "-h" option are considered. New fragments are discarded.

Type of fragmentation (-t option)

-t 0 Count of atoms

-t 1 Sequences of atoms only

-t 2 Sequences of bonds only

-t 3 Sequences of atoms and bonds

-t 4 Atom centered fragments based on sequences of atoms

-t 5 Atom centered fragments based on sequences of bonds

-t 6 Atom centered fragments based on sequences of atoms and bonds

-t 7 Atom centered fragments based on sequences of atoms of fixed length

-t 8 Atom centered fragments based on sequences of bonds of fixed length

-t 9 Atom centered fragments based on sequences of atoms and bonds of fixed length

-t 10 Triplets

2.3 Installation

The ISIDA Fragmentor2013 project is versionned with subversion on the infochim server. A few compiled executables are available on our website http://infochim.u-strasbg.fr in the Download then Fragmentor section (http://infochim.u-strasbg.fr/spip.php?rubrique4 If you need another compiled version or wish to have access to the source code, please contact Pr. A. Varnek (varnek@unistra.fr).

2.3.1 Steps for installation

1. Acquire Fragmentor2013 project using subversion (svn):

svn checkout svn+ssh://yourlogin@infochim.u-strasbg.fr/ home/infochimie/svn/Fragmentor2013 Fragmentor2013

2. In the same directory as your Fragmentor2013 directory (cd Fragmentor2013), acquire the Molecule project using svn:

svn checkout svn+ssh://yourlogin@infochim.u-strasbg.fr/ home/infochimie/svn/Molecule Molecule

1. Compile the project using preferably Lazarus with fpc or just fpc with the following options: -MObjFPC -Scgi -O3 -g -gl -vewnhi -l -FuMolecule -Fu.

2.4 Input and output formats

2.4.1 Input: Structure-Data File (.sdf)

SDF is a format developed by MDL (now part of Accelerys). Its format should be findable on Accelerys' website and a copy of the document is given in the doc folder of the project. The V2000 format is used by ISIDA Fragmentor2013. Here is a quick description of the most important features that an SDF should contain:

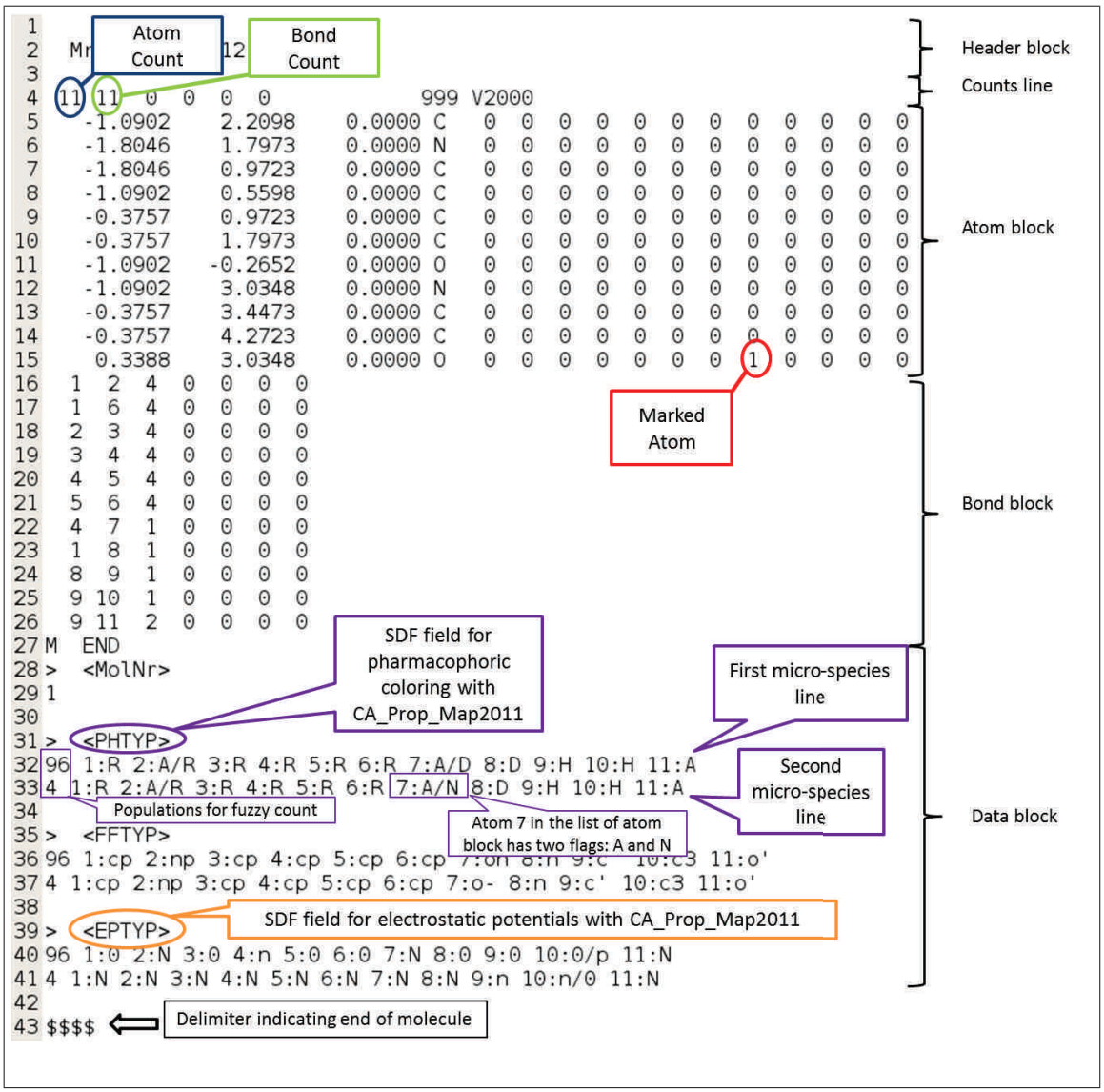


Figure 2.1: Example of SDF

Description of example SDF

- Line 1-3: Header block - contains name of molecule
- Line 4: Counts line - First 3 characters corresponds to the atom count, next 3 is the bond count.
- Line 5- 4 + atom count (15): Atom Block - each line in this block corresponds to an atom and each column corresponds to a different property of the atom. The number of lines corresponds to the atom count read in line 4.
  - Column 1-3: Spatial coordinates x,y,z
  - Column 4: Element
  - Column 6: Formal Charge (1 = +3, 2 = +2, 3 = +1, 4 = doublet radical, 5 = -1, 6 = -2, 7 = -3)
  - Column 12: Not used in MDL format. This column is used by the fragmentor to indicate marked atoms. To mark an atom, the 0 should be replaced by a 1. Like in the last atom of the atom block in 2.1 (line 15).
- Line 6 + atom count (16) - 6 + atom count + bond count (26): Bond block - each line corresponds to a bond where the two first values corresponds to the atoms involved in the bond and the third on is the bond type. ISIDA Fragmentor2013 has special bonds for CGRs outlined in the following table.
- Line 28-42: Data block - contains information separated into fields. In this example the fields generated by CA_Prop_Map2011.java are shown. The names of the fields are given as > <NAME>. The format for property mapping for ISIDA Fragmentor2013 is shown It should be of format:
  > <COLOR_NAME>
  5 1:P 2:H 4: A/D
  95 1:A 2:H 4:D
  where 5 and 95 are the count to be considered for each species and the following characters are Atom number: Colouration1/Colouration2
  COLOR_NAME should be indicated with the option -c.
- Line 43: Delimiter indicating end of molecule - the following lines will be a new molecule in the same format.

Bond Types The bond types with their respective symbols used in the generated descriptors and the integer used in the SDF to identify them. Note that column 3 corresponds to the character 7 to 9 and column 7 corresponds to character 19 to 21 found in the bond block line. The format for CGRs was modified compared to ISIDA Fragmentor2011: The symbols used in the descriptors changed and therefore ISIDA Fragmentor2012 and Fragmentor2013 are not retro-compatible with ISIDA Fragmentor2011 in the case of CGRs as well as for "Any Bonds" (column 3 = 8)' and "Special Bond" (column 3 = 9). A new format permitting the visualisation of dynamic bond with ChemAxon was implemented. However the previous format (visualisation with Edi SDF) is still readable - that is why each Dynamic bond is found twice in the following table.

Bond Type	Symbol	SDF bond column 3	SDF column 7
Simple	-	1	0 or 2
Double	=	2	0 or 2
Triple	+	3	0 or 2
Aromatic	*	4	0 or 2
Single or Double	.	5	0 or 2
Single or Aromatic	:	6	0 or 2
Double or Aromatic	"	7	0 or 2
Any bond type	?	8	0 or 2
Special bond type	_	9	0 or 2
Single bond in cycle	.	50	0 or 2
Double bond in cycle	:	60	0 or 2
Triple bond in cycle	#	70	0 or 2
Hydrogen bonds	~	80	0 or 2
Unknown bond	YY		

2.4.2 Output: Header file and SVM, SMF and CSV formats

ISIDA Fragmentor2013 will always output a header file with the extension .hdr and another file in either SVM, SMF or CSV format. By default, the SVM format is outputed and it can be changed with the option -f.

- SMF: The SMF (Substructural Molecular Fragments) format outputs 3 files: a header file .hdr, containing the index and a string representing each fragment discovered into the SDF, a sparse descriptor matrix in a .smf file and a one column file with the values of the field identified using the -s option. The sparse descripor matrix represent one molecule per line. It is read by pairs of column, the first one identifies a fragment, the second one how many times this fragment was discovered.

- SVM: The SVM (Support Vector Machine) format outputs 2 files: a header file .hdr, containing the index and a string representing each fragment discovered

Table 2.1: Visualisation of CGRs with ChemAxon

Bond Type	Symbol	SDF bond column 3	SDF column 7
Single bond creation	81	1	8
Double bond creation	82	2	4
Triple bond creation	83	3	12
Aromatic bond creation	84	4	1
Single bond cut	18	1	-1
Double bond cut	28	2	-1
Triple bond cut	38	3	-1
Aromatic bond cut	48	4	-1
Single bond to double bond	12	2	8
Single bond to triple bond	13	3	8
Single bond to aromatic bond	24	4	8
Double bond to single bond	21	1	4
Double bond to triple bond	23	3	4
Double bond to aromatic bond	24	4	4
Triple bond to single bond	31	1	12
Triple bond to double bond	32	2	12
Triple bond to aromatic bond	34	4	12
Aromatic bond to single bond	41	1	1
Aromatic bond to double bond	42	2	1
Aromatic bond to triple bond	43	3	1

into the SDF, and descriptor matrix in a file .svm following the libSVM format. The first column contains the values of the field identified using the -s option. Other columns consists in a pair of values separated by a ":". The first value identifies the fragment's index in the header file, the second one is the fragment count.

- CSV: The CSV (Comma-Separated Values) format outputs 2 files: a header file .hdr, containing the index and a string representing each fragment discovered into the SDF, and a sparse descriptor matrix in a .csv file where each value is separated by a semi-colon ";". The first value corresponds to the activity (given by the -s option), and it is then read by pairs of column, the first one identifies a fragment by its index, the second one how many times this fragment was discovered.

Table 2.2: Visualisation of CGRs with EdiSDF

Bond Type	Symbol	SDF bond column 3	SDF column 7
Single bond creation	81	81	4
Double bond creation	82	82	4
Triple bond creation	83	83	4
Aromatic bond creation	84	84	4
Single bond cut	18	18	4
Double bond cut	28	28	4
Triple bond cut	38	38	4
Aromatic bond cut	48	48	4
Single bond to double bond	12	12	8
Single bond to triple bond	13	13	8
Single bond to aromatic bond	24	24	8
Double bond to single bond	21	21	8
Double bond to triple bond	23	23	8
Double bond to aromatic bond	24	24	8
Triple bond to single bond	31	31	8
Triple bond to double bond	32	32	8
Triple bond to aromatic bond	34	34	8
Aromatic bond to single bond	41	41	8
Aromatic bond to double bond	42	42	8
Aromatic bond to triple bond	43	43	8

2.5 Nomenclature

To characterize the different fragment, they are coded according to the following:

TopologicalFragmentationColourationTypeBondInclusion (LowerLength-UpperLength)CountingType_Options

Where:

1. TopologicalFragmentation is a roman number and corresponds to the following fragmentation:

   I - Sequences (corresponds to -t 1, 2, 3)

   II - Atom-centred fragments (coressponds to -t 4, 5, 6, 7, 8, 9)

   III - Triplets (corresponds to -t 10)

2. ColourationType is a chain of letters starting with a capital and followed by only lower case letters. The following codes have been used up to now:

   - A – Atom symbol (when no special colouration is used)
   - Ph – Pharmacophoric typing (PHTYP generated by CA_Prop_Map2011.java)

- Ep – Topological electrostatic potentials (EPTYP generated by CA_Prop_Map2011.java)
- Pc – Partial Charges (PCTYP generated by CA_Prop_Map2011.java)
- Lp – LogP increments
- Ba – Benson atoms (when - -UseBenson was used)

3. BondInclusion simply indicates the inclusion of bond orders in the string with a capital B. If only bonds are used then no ColourationType will appear.

4. LowerLength and UpperLength are the number of atoms to be included at minimum and maximum respectively. Note that a LowerLength=2 and UpperLength=5 will create fragments with at minimum a topological distance of 1 and maximum a topological distance of 4.

5. CountingType corresponds to the type of weight used to count the occurrences of fragments:

   - ms – micro-species (pH dependent counting - PHTYP, EPTYP, PCTYP from CA_Prop_Map2012.java are used)

   When none is indicated then the direct count is used (weight =1).

6. Options indicate special options used during the fragmentation and are listed below:

   - P – AtomPairs (when - -AtomPairs is used)
   - R – Restricted (only for atom-centred fragments - corresponds to -t 7,8,9)
   - AP – AllPaths (when - -DoAllWays is used)
   - FC – FormalCharge representation (when - -UseFormalCharge is used)
   - MA1,MA2,MA3 – MarkedAtom with the used option number (-m 1,2 or 3) following the MA
   - SF – StrictFragmentation (when - -StrictFrg is used with a specific header in -h header.hdr)
   - AD – AllDynamic (Bonds) (when -d 2 is used)
   - OD – OneDynamic(Bond) (when -d 1 is used)

   Options are separated by a hyphen (-).

Example: IIPhB(3-5)ms_P-FC

2.5.1 A few examples of correspondance between ISIDA Fragmentor2013 options and Nomenclature of ISIDA descriptors

-t	-c	-l	-u	Other options	Nomenclature
0	/	/	/	/	No nomenclature
1	/	2	5	/	IA(2-5)
1	PHTYP	2	8	/	IPh(2-8)ms
1	PHTYP	3	5	-m 1	IPh(3-5)ms_MA
1	/	2	5	- -DoAllWays	IA(2-5)_AP
1	/	2	5	- -AtomPairs - -UseFormalCharge	IA(2-5)_P-FC
2	/	2	7	/	IB(2-7)
3	/	3	6	/	IAB(2-6)
3	EPTYP	3	6	/	IEpB(2-6)
3	PHTYP	3	6	/	IPhB(2-6)
8	/	2	4	/	IIA(2-4)
9	/	2	4	/	IIB(2-4)
10	/	2	4	/	IIAB(2-4)
11	/	2	4	/	IIA(2-4)_R
12	/	2	4	/	IIB(2-4)_R
13	/	2	4	/	IIA(2-4)_R
14	/	2	4	/	IIIA(2-4)

Chapter 3

Mapping properties using ChemAxon

3.1 Introduction

CA_Prop_Map2011 is a java program part of the Utils package based on ChemAxon's JChem classes and developed by Dragos Horvath and Fiorella Ruggiu. It requires therefore a ChemAxon license for the calculation plugin. Note that the pharmacophoric mapping is available on our Mobyle portal (http://infochim.u-strasbg.fr/spip.php?rubrique14

3.2 Usage

textbfjava Utils/CA_Prop_Map2011 -f <ChemAxon input> [-o <SDF> -min_ms_pop <double> -pH <double> -major_ms]
Options in squared brackets are not mandatory.

Options

- -f <input file> (path): the input file path and name. The input can also be piped into the program. It may be of any readable format by ChemAxon
- -o <output file> (path): the output file path and name. By default Typed.sdf. The generated SDF becomes then the input of the ISIDA Fragmentor2013.
- -min_ms_pop (double): the minimum population level of a microspecie for it to be taken into account. By default min_ms_pop=1.0
- -major_ms (toogle): if activated only the major microspecie will be considered
- -pH (double): indicate the pH at which the microspecies are calculated. By default pH=7.4
- -stdoptions (path): **DEPRECATED!!** (path to the file containing rules for the standardize)

The program does not standardize - it is recommended you standardize the file beforehand.

3.3 Installation

3.3.1 Steps for installation:

1. Download JChem from ChemAxon's website (http://www.chemaxon.com/download/jchem/jch for-java/)
2. Install JChem and its licence with the LicenseManager
3. Install a java runtime environment (JRE) and a java development kit (JDK)
4. Download the Utils package (with svn)
5. Edit your shell configuration file (.bashrc for a bash shell) to define the java CLASSPATH and eventually the path to your JRE
6. Compile CA_Prop_Map2011 with javac

3.3.2 ChemAxon JChem

To use this package you will need an installed version of JChem with licence, allowing you to use the calculation plugin. Download JChem from ChemAxon's website (http://www.chemaxon.com/download/jchem/jchem-for-java/). You will need an account on their website to do so. It is easier to use the installation with the JRE. Then install the program and run the LicenseManager to register you license. By default, it should be placed in the .chemaxon directory found in the user's home.

3.3.3 Java

To run and compile the classes, a JRE and a JDK are needed.
For linux, choose the java-sun packages. Configure your media to contain the non-free packages and updates in your mirror list (For Mandriva/GNOME, got to Administration→Configure your system → Software Management → Configure media sources). In the Software Manager, search for the following two packages and install them: java-1.6.0-sun and java-1.6.0-sun-devel. Note: If you installed JChem with a JRE, the java-1.6.0-sun will already be installed.

3.3.4 Utils package

In order to obtain the package, use subversion. To install it on linux, use the Software Manager and install the package. For Windows, use TortoiseSVN (http://tortoisesvn.net).
The deposit is on infochimie on the following path: /home/infochimie/svn/Utils. To acquire it, you will need to use the following command:
svn checkout svn+ssh://yourlogin@infochim.u-strasbg.fr/home/infochimie/svn/Utils

3.3.5 Java CLASSPATH

To compile the java programs using ChemAxon's classes, the CLASSPATH needs to contain the path to them. CA_Prop_Map2012 also requires the definition of variables to find its configuration files. You may define them just before using the program or integrate them into your shell configuration file.
Example of .bashrc:

```
CLASSPATH=/opt/chemaxon/jchem/lib/jchem.jar:/opt/scripts/JavaClasses
export CLASSPATH
STANDARD_RULES=/opt/scripts/JavaClasses/Utils/Standardize.xml
export STANDARD_RULES
SH_PHARMAFLAG_RULES=/opt/scripts/JavaClasses/Utils/shortPharmFlags.xml
export SH_PHARMAFLAG_RULES
FORCEFIELD_RULES=/opt/scripts/JavaClasses/Utils/cvffTemplates.xml
export FORCEFIELD_RULES
```

Example of .cshrc:

```
setenv CLASSPATH /opt/chemaxon/jchem/lib/jchem.jar:/opt/scripts/JavaClasses
setenv STANDARD_RULES /opt/scripts/JavaClasses/Utils/Standardize.xml
setenv SH_PHARMAFLAG_RULES /opt/scripts/JavaClasses/Utils/shortPharmFlags.xml
setenv FORCEFIELD_RULES /opt/scripts/JavaClasses/Utils/cvffTemplates.xml
```

3.3.6 Javac Compilation

Compile the program using the following command:
javac /opt/scripts/JavaClasses/Utils/CA_Prop_Map2012.java

Appendix A

Abbreviations

- CGRs: Condensed Graph of Reactions
- FPT: Fuzzy Pharmacophoric Triplets (ISIDA descriptors)
- IPLF: ISIDA Property-Labelled Fragments (descriptors)
- ISIDA: In SIlico Design and data Analysis
- JDK: Java Development Kit
- JRE: Java Runtime Environment
- QSAR: Quantitative Structure-Activity Relationship
- SDF: Structure-Data File (from MDL - now Accelerys)
- SMF: Substructural Molecular Fragments (ISIDA descriptors)

Bibliography

[1] P. Lacomme, C. Prins, and M. Sevaux, *Algorithmes de graphes*. Eyrolles, second ed., 2003.

[2] A. Varnek, D. Fourches, F. Hoonakker, and V. Solov'ev, "Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures.," *J. Computer-Aided Molecular Design*, vol. 19, pp. 693–703, Jul 2005.

[3] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. Tetko, and G. Marcou, "ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors," *Curr Comput Aided Drug Des.*, vol. 4, pp. 191–198, Sept 2008.

[4] F. Bonachera, B. Parent, F. Barbosa, N. Froloff, and D. Horvath, "Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes.," *J Chem Inf Model.*, vol. 46, pp. 2457–2477, Nov-Dec 2006.

[5] F. Bonachera and D. Horvath, "Fuzzy tricentric pharmacophore fingerprints. 2. application of topological fuzzy pharmacophore triplets in quantitative structure-activity relationships.," *J Chem Inf Model.*, vol. 48, pp. 409–425, Feb 2008.

[6] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, "ISIDA Property-Labelled Fragment Descriptors," *J Chem Inf Model.*, vol. 29, p. 855–868, Dec 2010.

[7] F. Ruggiu, V. Solov'ev, G. Marcou, D. Horvath, J. Graton, J.-Y. Le Questel, and A. Varnek, "Individual hydrogen-bond strength QSPR modelling with ISIDA local descriptors: a step towards polyfunctional molecules," *Mol Inf*, vol. tbp, p. tbp, tbp 2014.

Appendix B

Supporting Information of "Quantitative Structure-Property Relationship Modeling: A Valuable Support in High-Throughput Screening Quality Control" article

This appendix contains the supporting information for the "Quantitative Structure-Property Relationship Modeling: A Valuable Support in High-Throughput Screening Quality Control" article published in Analytic Chemistry, 2014, volume 86, pages 2510-2520. First the additional details of the modelling are given, followed by the data on the chromatographic hydrophobicity index.

Supporting Information of "QSPR modelling – a valuable support in HTS quality control"

Fiorella Ruggiu*[1], Patrick Gizzi*[2, 5], Jean-Luc Galzi[2, 5], Marcel Hibert[3, 5], Jacques Haiech[3, 5], Igor Baskin[1, 4], Dragos Horvath[1], Gilles Marcou[1], Alexandre Varnek[1]

[1] Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France

[2] Laboratoire de Biotechnologie et Signalisation Cellulaire (Plate-forme TechMed^{ILL}), UMR 7242 CNRS/Université de Strasbourg, Ecole Supérieure de Biotechnologie Strasbourg, 67412 Illkirch Graffenstaden, France

[3] Laboratoire d'Innovation Thérapeutique, UMR7200 CNRS/Université de Strasbourg, Faculté de Pharmacie, 74 route du Rhin, 67401 Illkirch, France

[4] Lomonosov Moscow State University, Moscow 119991, Russia

[5] Laboratory of Excellence Medalis

* These authors contributed equally to the work

Table of Content

1. Calibration compounds and their associated ICHI value

Reference compound	ICHI
Theophylline	18.4
Phenyltetrazole	23.6
Benzimidazole	34.3
Colchicine	43.9
Phenyltheophylline	51.7
Acetophenone	64.1
Indole	72.1
Propiophenone	77.4
Butyrophenone	87.3
Valerophenone	96.4

Table SI 1. Calibration compounds and their associated ICHI value

2. Statistical parameters definition

Root-mean squared error: $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left\{\left(Y_i - Y_{pred,i}\right)^2\right\}}{n}}$$

Mean Absolute Error: $$MAE = \frac{\sum_{i=1}^{n}|Y_i - Y_{pred,i}|}{n}$$

Determination coefficient: $$R^2_{det} = 1 - \frac{\sum_{i=1}^{n}\left\{\left(Y_i - Y_{pred,i}\right)^2\right\}}{\sum_{i=1}^{n}\{(Y_i - \bar{Y})^2\}}$$

Correlation coefficient: $$R^2_{corr} = \frac{\sum_{i=1}^{n}\left\{(Y_i - \bar{Y})\left(Y_{pred,i} - \overline{Y_{pred}}\right)\right\}}{\sqrt{\sum_{i=1}^{n}\{(Y_i - \bar{Y})^2\}\sum_{i=1}^{n}\left\{\left(Y_{pred,i} - \overline{Y_{pred}}\right)^2\right\}}}$$

Where i is the number of the concerned compound, n is the total number of compounds, $Y_{pred,i}$ is the predicted CHI value of compound i by the model, Y_i is the experimental CHI value of compound i, $\bar{Y}$ is the mean value of experimental CHI values and $\overline{Y_{pred}}$ is the mean value of predicted CHI values.

3. <u>The parameters of the 10 SVM models used for the *eliminate-and-refit* protocol to detect outliers:</u>

The descriptor software used was the ISIDA Fragmentor2012 and the SVM package was Libsvm 3.12. In the following table, a description of the descriptor space is given in the first column, then the corresponding options used in ISIDA Fragmentor2012 are given in the second column, finally the options for the Libsvm program are indicated in the third column. For more details about the descriptors, please refer to our website: http://infochim.u-strasbg.fr/spip.php?rubrique49.

Descriptor space	ISIDA Fragmentor2012 options	Libsvm 3.12 options
Restricted augmented atoms with pairs coloured by atomic symbols and including bonds and formal charges with a minimum length of 2 and a maximum length of 4	-t 9 -c Default –l 2 –u 4 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -p 2 -c 0.1
		-s 3 -k 0 -p 2 -c 0.2
Restricted augmented atoms coloured by atomic symbols and formal charges with a minimum length of 2 and a maximum length of 4	-t 9 -c Default –l 2 –u 4 --FormalCharges -t 0 -c Default	-s 3 -k 0 -p 2 -c 0.4
		-s 3 -k 0 -p 2 -c 0.5
Restricted augmented atoms with pairs coloured by atomic symbols with a minimum length of 2 and a maximum length of 4	-t 7 -c Default –l 2 –u 4 --AtomPairs -t 0 -c Default	-s 3 -k 0 -p 2 -c 0.4
Restricted augmented atoms with pairs coloured by atomic symbols and including bonds and formal charges with a minimum length of 2 and a maximum length of 3	-t 9 -c Default –l 2 –u 3 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -p 2 -c 0.2
Triplets coloured by force field including formal charges with a minimum and a maximum length of 2	-t 10 -c FFTYP –l 2 –u 2 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -p 2 -c 2
		-s 3 -k 0 -p 2 -c 5
Restricted augmented atoms coloured by force field and including bonds and formal charges with a minimum length of 2 and a maximum length of 3	-t 9 -c FFTYP –l 2 –u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -p 2 -c 0.1
Restricted augmented atoms with pairs coloured by force field and including bonds and	-t 9 -c FFTYP –l 2 –u 3 --FormalCharges --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -p 2 -c 0.1

formal charges with a minimum length of 2 and a maximum length of 3		

Table SI 2. Parameters of the 10 SVM models for the eliminate-and-refit protocol to detect outliers

4. Structures and CHI values of the 3 compounds below 0 after removing outliers

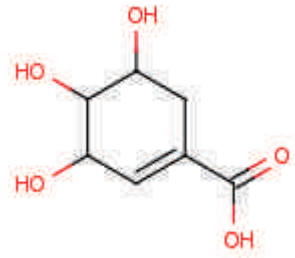
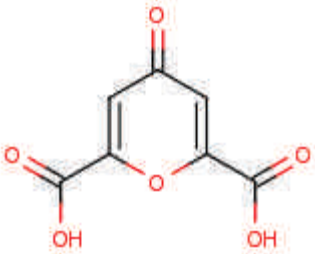
Structure	CHI 1
	-35.5
	-27.2
	-25.9

Table SI 3. The structures and CHI values of the 3 compounds below 0 removed from set

5. Experimental vs predicted value of CHI on the external test set using the consensus model

The plot below (Figure SI 1) shows the experimental CHI values against the predicted CHI values by our consensus model on the external literature test set. The model reasonably performs on the external test set with a root-mean squared error RMSE = 16.4, a mean absolute error MAE= 12.6, a determination coefficient R^2_{det} = 0.6 and a correlation coefficient R^2_{corr} = 0.8. These results are in good agreement with the 5-CV results (best models with RMSE = 14.5 and R^2_{det} = 0.7) when considering that the test set data has slightly different experimental setups.

The data points are coloured according to the source and method used for measurements:

- Valko1997[1] (squares in dark blue),
- Plassa1998[2] (triangles in light blue),
- Camurri2001[3] with LC-UV (triangles in green),
- Camurri2001[3] with LC-MS (squares in red),
- Valko2001[4] (circles in yellow) and
- Fuguet2007[5] (diamond-shaped in violet).

The data points measured by LC-MS from Camurri et al. are noticeably less well predicted. This could be due to the change of method. In general, the data points from Camurri et al. (14) are more spread than the data points from other publications.

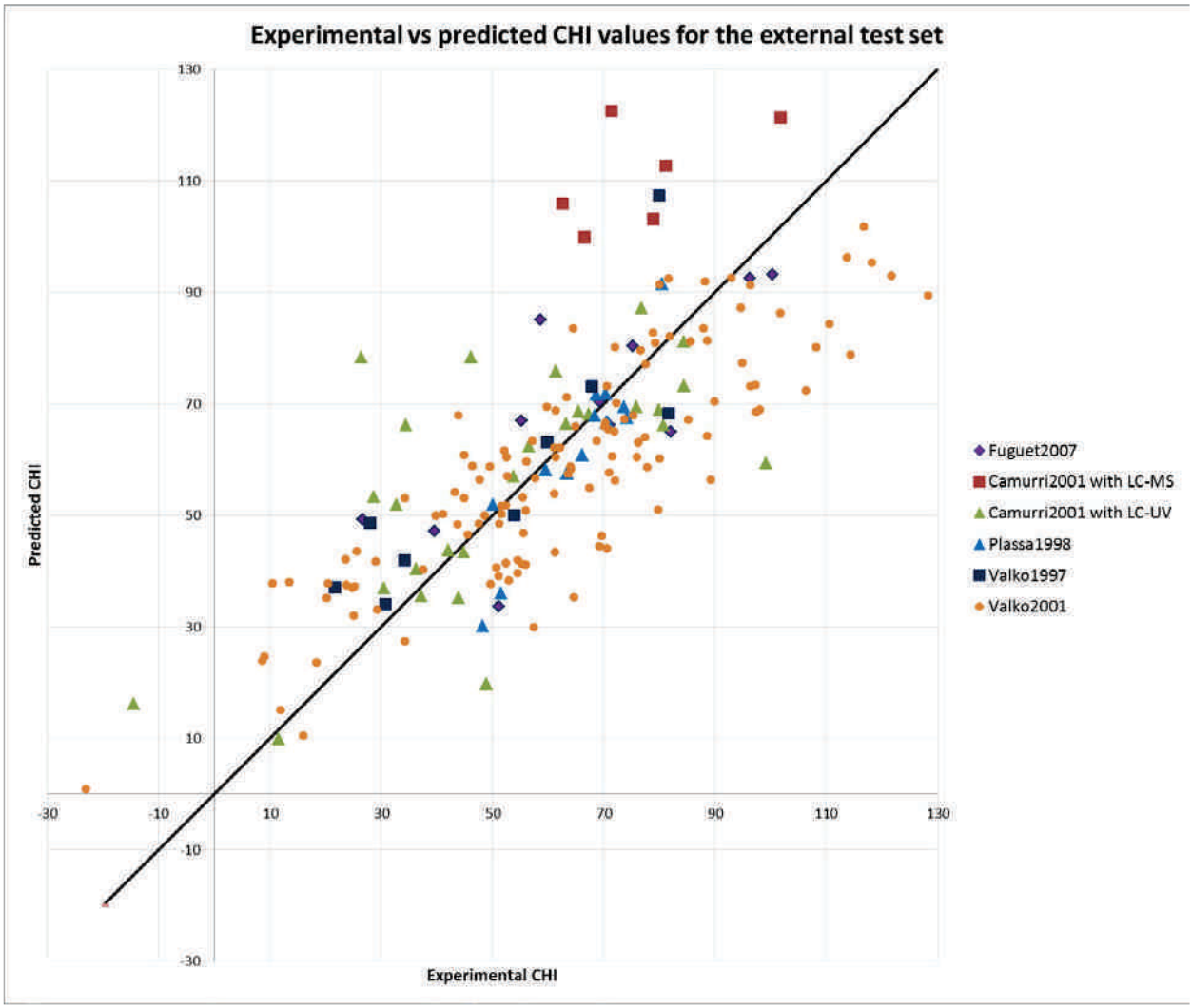


Figure SI 1. Experimental vs predicted CHI value on the external test set by the consensus model

6. Availability of the model for end users

This model was used to annotate the CN and is freely available online: http://infochim.u-strasbg.fr/webserv/VSEngine.html (see Figure SI 2). We invite all interested users to try our web server. Accounts are free and, in order to predict CHI, one needs to select the "QSAR-based Property Predictions" in the menu (indicated with a ***1*** in Figure SI 2). Then, molecules may be drawn or any format supported by ChemAxon may be uploaded. They will then appear in ***2***. Several properties are available in the menu. The consensus model for CHI described in this article is available as CHI_svm (see ***3*** in Figure SI 2). After pressing "GO!", the web server will apply the SVM models and output the most reliable prediction in the column CHI_svm (see lower part of Figure SI 2).

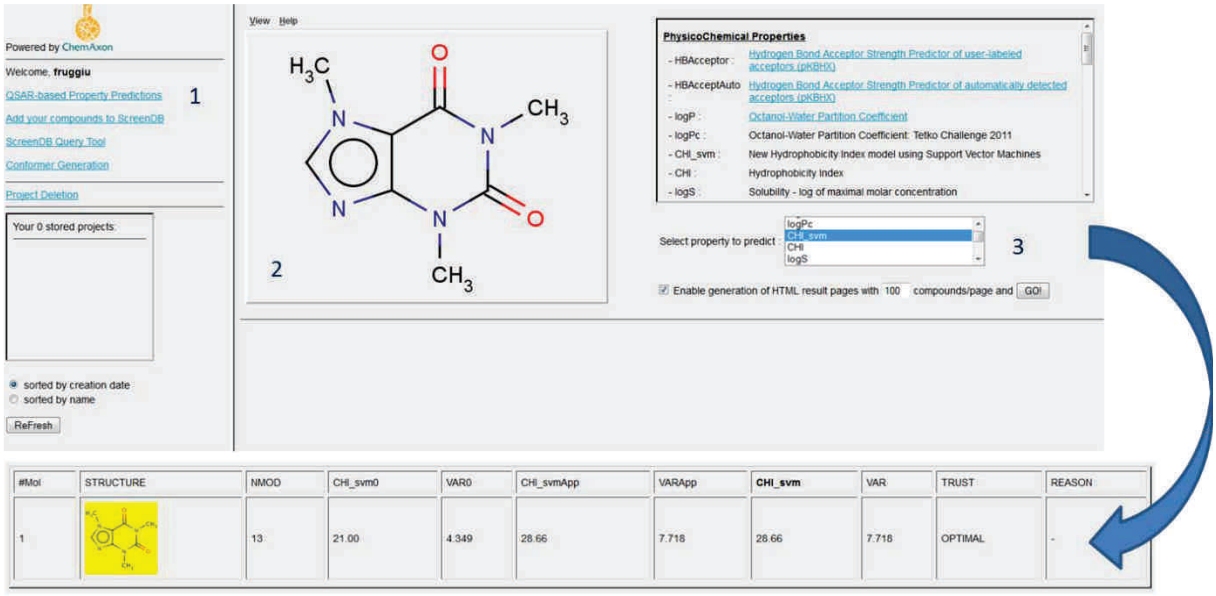


Figure SI 2. The prediction web server using caffeine as an example

7. Descriptor spaces, ISIDA Fragmentor2012 options, Libsvm options and statistics of the 81 models used in the consensus model

The software used were ISIDA Fragmentor 2012 and Libsvm 3.12. Descriptor space nomenclature and explanation of ISIDA Fragmentor 2012 options are given on our website: http://infochim.u-strasbg.fr/spip.php?rubrique49. The statistics given here are a mean of 5 iterations of the 5-fold cross-validation (5CV) procedure.

Table SI 4. Descriptors details, SVM parameters and statistical parameters for the models included in the consensus modelling

Descriptors		SVM models options	Statistics			
Descriptor Space	ISIDA Fragmentor2012 options		5CV-RMSE	5CV-R^2_{det}	5CV-R^2_{corr}	5CV-MAE
IIA2-3_R-FC	-t 7 -c Default -l 2 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,9	14,5	0,7	0,9	11,5
IIAB2-3_R-FC	-t 9 -c Default -l 2 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,8	14,5	0,7	0,9	11,4
IIA2-3_P-FC	-t 7 -c Default -l 2 -u 3 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	14,6	0,7	0,9	11,7
IIPhB2-4_ms-R	-t 9 -c PHTYP -l 2 -u 4 -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	14,6	0,7	0,9	11,5
IIAB2-3_R-P-FC	-t 9 -c Default -l 2 -u 3 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,8	14,6	0,7	0,9	11,5
IIPhB2-4_ms-R-FC	-t 9 -c PHTYP -l 2 -u 4 --FormalCharges -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	14,7	0,7	0,9	11,4
IIA2-4_R-FC	-t 7 -c Default -l 2 -u 4 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,6	14,7	0,7	0,9	11,8
IIAB2-4_R-FC	-t 9 -c Default -l 2 -u 4 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,7	14,8	0,7	0,8	11,7
IIAB2-3_FC	-t 9 -c Default -l 2 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 2	14,9	0,7	0,8	11,6
IIA2-4_R-P-FC	-t 9 -c Default -l 2 -u 4 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,6	15,0	0,7	0,8	11,7
IIFf2-3_ms-P-	-t 4 -c FFTYP -l 2 -u 3 --FormalCharges	-s 3 -k 0 -e 2 -c 0,1	15,0	0,7	0,8	11,8

FC	--AtomPairs -t 0 -c FFTYP					
IIAB2-3_P-FC	-t 6 -c Default -l 2 -u 3 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 2	15,0	0,7	0,8	11,7
IIAB2-4_R	-t 9 -c Default -l 2 -u 4 -t 0 -c Default	-s 3 -k 0 -e 2 -c 2	15,0	0,7	0,8	11,9
IIAB3-3_FC	-t 6 -c Default -l 3 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,0	0,7	0,8	11,9
IIAB3-3_R-FC	-t 9 -c Default -l 3 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,0	0,7	0,8	11,9
IIPh3-4_ms-R	-t 7 -c PHTYP -l 3 -u 4 -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,1	0,7	0,8	11,8
IIPh3-4_ms-R-FC	-t 7 -c Default -l 3 -u 4 --FormalCharges -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,1	0,7	0,8	11,8
IIPh2-4_ms-R-FC	-t 7 -c PHTYP -l 2 -u 4 --FormalCharges -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,1	0,7	0,8	11,8
IIPhB3-4_ms-R	-t 9 -c PHTYP -l 3 -u 4 -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,1	0,7	0,8	11,9
IIPhB3-4_ms-R-FC	-t 9 -c PHTYP -l 3 -u 4 --FormalCharges -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,1	0,7	0,8	11,7
IIA2-3_FC	-t 4 -c Default -l 2 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 2	15,1	0,7	0,8	11,9
IIAB2-2_R-P-FC	-t 9 -c Default -l 2 -u 2 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,1	0,7	0,8	12,0
IIFfB2-3_ms-P	-t 9 -c FFTYP -l 2 -u 3 --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,1	0,7	0,8	11,9
IIFfB2-3_ms	-t 6 -c FFTYP -l 2 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,2	0,7	0,8	12,0
IIFf2-3_ms-FC	-t 4 -c FFTYP -l 2 -u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,2	0,7	0,8	12,0
IIPh2-4_ms-R	-t 4 -c PHTYP -l 2 -u 4 -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,2	0,7	0,8	11,8
IIFfB2-3_ms-FC	-t 6 -c FFTYP -l 2 -u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,2	0,7	0,8	12,0
IIFfB2-3_ms-P-FC	-t 9 -c Default -l 2 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,1	15,2	0,7	0,8	12,0

IIA3-3_FC	-t 4 -c Default -l 3 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,2	0,7	0,8	12,1
IA2-4_P-FC	-t 1 -c Default -l 2 -u 4 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 2	15,3	0,7	0,8	12,0
IIFf2-3_ms	-t 4 -c FFTYP -l 2 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,3	0,7	0,8	12,0
IIAB2-4_R-P-FC	-t 9 -c Default -l 2 -u 4 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,7	15,3	0,7	0,8	12,1
IIA2-2_P-FC	-t 4 -c Default -l 2 -u 2 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,9	15,3	0,7	0,8	12,2
IIPhB2-3_ms-P	-t 6 -c PHTYP -l 2 -u 3 --AtomPairs -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,3	0,7	0,8	11,8
IIFfB3-3_ms-R	-t 9 -c FFTYP -l 3 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,3	0,7	0,8	12,1
IIFf2-3_ms-P	-t 4 -c FFTYP -l 2 -u 3 --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,3	0,7	0,8	12,1
IAB2-4_FC	-t 3 -c Default -l 2 -u 4 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,5	15,3	0,7	0,8	12,2
IIAB2-2_P-FC	-t 6 -c Default -l 2 -u 2 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,3	0,7	0,8	12,2
IIFfB3-3_ms	-t 6 -c FFTYP -l 3 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,4	0,7	0,8	12,1
IIFfB3-3_ms-FC	-t 6 -c FFTYP -l 3 -u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,4	0,7	0,8	12,1
IIFf3-3_ms-FC	-t 4 -c FFTYP -l 2 -u 2 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,4	0,7	0,8	12,1
IIB2-2_R-FC	-t 8 -c Default -l 2 -u 2 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 8	15,4	0,7	0,8	12,2
IIA2-2_R-FC	-t 7 -c Default -l 2 -u 2 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,9	15,4	0,7	0,8	12,2
IIFf3-3_ms-R-FC	-t 7 -c FFTYP -l 3 -u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,4	0,7	0,8	12,1

IIAB2-2_R-FC	-t 9 -c Default -l 2 -u 2 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,4	0,7	0,8	12,2
IIAB2-2_FC	-t 6 -c Default -l 2 -u 2 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,4	0,7	0,8	12,2
IIA2-2_FC	-t 4 -c Default -l 2 -u 2 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,9	15,4	0,7	0,8	12,3
IIFfB3-3_ms-R-FC	-t 9 -c FFTYP -l 3 -u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,4	0,7	0,8	12,2
IIB2-2_FC	-t 5 -c Default -l 2 -u 2 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 8	15,4	0,7	0,8	12,2
IIA3-3_R-FC	-t 7 -c Default -l 3 -u 3 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,5	0,7	0,8	12,4
IIPhB2-3_ms-FC	-t 6 -c PHTYP -l 2 -u 3 --FormalCharges -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,5	0,7	0,8	12,0
IIFfB3-4_ms-FC	-t 6 -c FFTYP -l 3 -u 4 --FormalCharges -t 0 –c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,5	0,7	0,8	12,5
IIPhB2-3_ms-P-FC	-t 6 -c PHTYP -l 2 -u 3 --FormalCharges --AtomPairs -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,5	0,7	0,8	11,9
IIFf3-3_ms-R	-t 7 -c FFTYP -l 3 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,5	0,7	0,8	12,2
IIFf3-3_ms	-t 4 -c FFTYP -l 3 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,5	0,7	0,8	12,2
IIFfB4-4_ms	-t 6 -c FFTYP -l 4 -u 4 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,5	0,7	0,8	12,5
IIA2-3_R-P-FC	-t 7 -c Default -l 2 -u 3 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,6	15,5	0,7	0,8	12,3
IIFf2-4_ms-R	-t 7 -c FFTYP -l 2 -u 4 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,6	0,7	0,8	12,3
IA2-3_P-FC	-t 1 -c Default -l 2 -u 3 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 2	15,6	0,7	0,8	12,5
IIPhB2-3_ms	-t 6 -c PHTYP -l 2 -u 3 -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	15,6	0,7	0,8	12,2
IIA2-2_R-P-FC	-t 7 -c Default -l 2 -u 2 --FormalCharges --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,9	15,6	0,7	0,8	12,5

IIFf3-3_ms-R-P-FC	-t 7 -c FFTYP -l 3 -u 3 --FormalCharges --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,6	0,7	0,8	12,2
IIAB2-3_P	-t 4 -c Default -l 2 -u 3 --AtomPairs -t 0 -c Default	-s 3 -k 0 -e 2 -c 3	15,6	0,7	0,8	12,2
IIFfB3-3_ms-R-P-FC	-t 9 -c FFTYP -l 3 -u 3 --FormalCharges --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,6	0,7	0,8	12,3
IIFfB4-4_ms-FC	-t 6 -c FFTYP -l 4 -u 4 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,6	0,7	0,8	12,6
IIFfB4-4_ms-R-FC	-t 9 -c FFTYP -l 4 -u 4 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,6	0,7	0,8	12,6
IIFf2-3_ms-R	-t 7 -c FFTYP -l 2 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,7	0,7	0,8	12,2
IAB3-4_FC	-t 3 -c Default -l 3 -u 4 --FormalCharges -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,7	15,7	0,7	0,8	12,3
IIAB2-3_	-t 6 -c Default -l 2 -u 3 -t 0 -c Default	-s 3 -k 0 -e 2 -c 7	15,7	0,7	0,8	12,3
IIFfB4-4_ms-R	-t 9 -c FFTYP -l 4 -u 4 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,7	0,7	0,8	12,7
IAB2-4_AP-FC	-t 3 -c Default -l 2 -u 4 --FormalCharges --DoAllWays -t 0 -c Default	-s 3 -k 0 -e 2 -c 0,5	15,7	0,7	0,8	12,4
IIFfB2-4_ms-R-FC	-t 9 -c FFTYP -l 2 -u 4 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,7	0,7	0,8	12,4
IIFfB3-3_ms-P-FC	-t 6 -c FFTYP -l 3 -u 3 –FormalCharges --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,7	0,7	0,8	12,4
IIFfB2-3_ms-R-FC	-t 9 -c FFTYP -l 2 -u 3 --FormalCharges -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,7	0,7	0,8	12,5
IIAB2-3_R	-t 9 -c Default -l 2 -u 3 -t 0 -c Default	-s 3 -k 0 -e 2 -c 1	15,7	0,7	0,8	12,5
IIFfB3-4_ms	-t 6 -c FFTYP -l 3 -u 4 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,8	0,7	0,8	12,7
IIFfB2-3_ms-R	-t 9 -c FFTYP -l 2 -u 3 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,8	0,7	0,8	12,4
IIFfB2-4_ms-R	-t 9 -c FFTYP -l 2 -u 4 -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,8	0,7	0,8	12,5
IIFf2-3_ms-R-	-t 9 -c FFTYP -l 2 -u 3 –FormalCharges -t 0 -c	-s 3 -k 0 -e 2 -c 0,1	15,9	0,7	0,8	12,5

FC	FFTYP					
IIFf3-3_ms-P-FC	-t 4 -c FFTYP -l 3 -u 3 --FormalCharges --AtomPairs -t 0 -c FFTYP	-s 3 -k 0 -e 2 -c 0,1	15,9	0,7	0,8	12,4
IIPh2-3_ms-FC	-t 6 -c PHTYP -l 2 -u 3 –FormalCharges -t 0 -c PHTYP	-s 3 -k 0 -e 2 -c 0,1	16,0	0,7	0,8	12,3

8. References

1. Valkò, K.; Bevan, C.; Reynolds, D., Chromatographic Hydrophobicity Index by Fast-Gradient RP-HPLC: A high-Throughput alternative to log P/log D. *Anal. Chem.* **1997,** *69*, 2022-2029.
2. Plassa, M.; Valkò, K.; Abraham, M. H., Determination of solute descriptors of tripeptide derivatives based on high-throughput gradient high-performance liquid chromatography retention data. *J. Chromatogr. A* **1998,** *803* (1-2), 51-60.
3. Camurri, G.; Zaramella, A., High-Throughput Liquid Chromatography/Mass Spectrometry Method for the Determination of the Chromatographic Hydrophobicity Index. *Anal. Chem.* **2001,** *73*, 3716-3722.
4. Valkò, K.; Du, C. M.; Bevan, C.; Reynolds, D. P.; Abraham, M. H., Rapid method for the estimation of octanol/water partition coefficient (log Poct) from gradient RP-HPLC retention and a hydrogen bond acidity term (Sigma-alphaH2). *Curr. Med. Chem.* **2001,** *8* (9), 1137-1146.
5. Fuguet, E.; Ràfols, C.; Bosch, E.; Rosés, M., Determination of the chromatographic hydrophobicity index for ionisable solutes. *J. Chromatogr. A* **2007,** *1173* 110-119.

The following CHI Training set contains all measured values. The molecules numbered from 1 to 485 correspond to the set used for our consensus modelling. The molecules numbered from 486 to 545 were considered as problematic - the outliers are given at the end of the list.

Molnb.	#SMILES	CHI
1	CC(=O)Nc1ccc(cc1)S(=O)(=O)Nc1cc(C)c2[nH]c3ccc(Cl)cc3c2c1C	79,1
2	CCC(C(=O)NCCc1cccn1C)c1ccccc1	80,7
3	CCC(C(=O)N(CC)CCN(C)C)c1ccccc1	51,1
4	OC1NC(=O)c2cc(oc2-n2cccc12)-c1ccc(Cl)cc1	74,8
5	CN(C)CCCC1C(Sc2ccccc2NC1=O)c1ccco1	50,3
6	CC1(C)Nc2cc(Cl)ccc2-n2cccc12	104,3
7	NNC(=O)c1cc2c(nc3ccccc3n2c1)-c1ccccc1	58,0
8	CC(C)=CCOC(=O)CC(NC(=O)C(F)(F)F)c1ccccc1	95,6
9	O=C1CC(c2ccccc2)n2cccc12	71,1
10	CCOC(=O)c1oc(cc1-n1cccc1C=O)-c1ccc(OC)cc1	93,6
11	Cc1c2c3cc(Cl)ccc3[nH]c2c(C)c2c(O)cc(CCCC(=O)NN)nc12	55,1
12	CCC(O)c1ccc2[nH]c3c(C)c4c(O)cc(C)nc4c(C)c3c2c1	51,6
13	COC(=O)c1c(C)nc(C)c(C(=O)OC)c1-c1cccs1	77,8
14	CN1CC(c2ccccc2)c2cccc(N)c2C1	62,6
15	COc1cc(C#N)c(cc1OC)-n1cccc1	80,0
16	NC(=O)c1cc(Cl)ccc1-n1cccc1	58,3
17	CCOC(=O)C(NC(=O)c1cc2OCOc2cc1N(=O)=O)C(=O)OCC	70,1
18	COc1ccc(CCNC(=O)C(=O)c2cccn2-c2ccc(O)cc2)cc1	76,4
19	OC(C(=O)NCCc1ccc(Cl)cc1)c1cccn1-c1ccc(cc1)C(F)(F)F	94,9
20	COc1ccc(cc1)-n1cccc1C(O)C(=O)NCCc1ccc(Cl)cc1	84,8
21	CCOC(=O)NCC(c1ccccc1)n1cccc1	83,7
22	OC1CC2N(C1)C(=O)c1ccc(cc1NC2=O)N(=O)=O	24,1
23	Cc1ccc(cc1)C(O)CC1NCc2ccsc2-n2cccc12	85,6
24	FC(F)(F)C(=O)NC1CNC(=O)c2sccc12	44,4
25	FC(F)(F)C(=O)NC1Cc2c(ccc(Cl)c2Cl)C1=O	78,9
26	OC(=O)CC(N1C(=O)c2ccccc2C1=O)c1cccc(O)c1	35,1
27	Clc1cc(Cl)cc(c1)C(=O)NC1CC(=O)c2ccccc12	87,4
28	COc1cc(cc(OC)c1OC)C(=O)OC1CCN2CCCC12	42,5
29	CCNC1CC(O)c2ccc(OC)cc12	23,3
30	Fc1ccc(cc1)C(NC(=O)C(F)(F)F)c1ccsc1-n1cccc1	100,8
31	S=C1NC=C(N1)c1ccncc1	18,6
32	OC1CCCc2c1[nH]c1cc(Cl)ccc21	85,7
33	CN1C2C(CCCC2=O)c2ccccc12	91,9
34	Cn1c2ccccc2c2ccc3cncnc3c12	100,1
35	CC1(C)CN(C1=O)c1ccc(cc1C(O)=O)N(=O)=O	29,7
36	CC(C)(CCl)C(=O)NCCc1ccc(Cl)cc1	86,0
37	CC(C)(CCl)C(=O)Nc1ccc(cc1)C(=O)NCCCCCC(O)=O	38,5
38	Clc1ccc(cc1)-c1csc(NC(=O)Nc2ccc(Br)cc2)c1	108,4
39	CC(=O)NC1(c2ccccc2-n2cccc12)c1ccccc1	77,2
40	NNC(=O)C(=O)NC1(c2ccsc2-n2cccc12)c1ccc(F)cc1	70,5
41	Clc1ccc2c(c1)N1C(=O)ON=C1C1CSCN1C2=S	87,0
42	O=C1NN=C2N1c1ccsc1-n1cccc21	50,2
43	O=N(=O)c1cc2OCOc2cc1N(=O)=O	73,6
44	O=C(Nc1ccccc1)N1CC(NC1=O)c1ccccc1	75,3
45	COC(=O)c1sc2C(C)CCCc2c1N	93,2
46	OCc1cc(Cl)ccc1-n1cccc1	80,1
47	COC(=O)c1c(C)nc(C)c(C(=O)OC)c1-c1cc2OCOc2cc1NC(C)=O	61,7
48	Cc1ccc(cc1)S(=O)(=O)NC(CC(O)=O)c1ccco1	37,4
49	COc1ccc(cc1)-c1cc(NC(=O)N(C)CCc2ccccc2)c(s1)C(O)=O	58,5

50	COc1ccc(cc1)-c1cc(NC(=O)N2CCCC2)c(s1)C(O)=O	41,9
51	O=C(NCc1ccsc1-n1cccc1)N1CCCCC1	80,1
52	CC(C)(C)CC(=O)NCc1ccsc1-n1cccc1	87,3
53	c1cc2cnc3sccc3n2c1	69,7
54	NC(=N)NC1CCC=CC1	22,5
55	O=C1CCC(CC1)NCc1ccccc1	63,1
56	Nc1cccnc1NC1CCCCCC1	66,5
57	Nc1nc(ccc1N(=O)=O)N1CCOCC1	51,9
58	O=S(Cc1ccccn1)c1nc(c[nH]1)-c1ccncc1	29,2
59	CCN(Cc1ccccc1)C(=O)Nc1c(csc1C(O)=O)-c1ccccc1	54,2
60	CC1CCCN(C1)C(=O)Nc1cc(sc1C(O)=O)-c1ccccc1	53,2
61	CCOC(=O)C1CCN(CC1)C(=O)Nc1cc(sc1C(O)=O)-c1ccc(OC)cc1	51,9
62	BrCCCCN1c2oc(cc2-n2cccc2C1=O)-c1ccccc1	107,2
63	CC(C)OC(=O)NC1CCN(CC1)c1ccc(I)cn1	40,8
64	Cc1sc(cc1N(=O)=O)N(=O)=O	76,2
65	OC(=O)c1sc(cc1NC(=O)NC1CCCC1)-c1ccccc1	51,8
66	OC(=O)c1ccsc1NC(=O)N1CCCC1	21,3
67	CN(CCc1ccccc1)C(=O)Nc1cc(sc1C(O)=O)-c1ccc(F)cc1	61,4
68	CNC(=O)Nc1sccc1C(O)=O	12,7
69	OC(=O)c1sc(cc1NC(=O)NC1CCCCCCC1)-c1ccc(F)cc1	68,7
70	OC(=O)c1sc(cc1NC(=O)NCC1CC1)-c1ccc(F)cc1	49,9
71	CCOC(=O)C1CCN(CC1)C(=O)Nc1cc(sc1C(O)=O)-c1cccs1	50,5
72	CCOC(=O)C1CCN(CC1)C(=O)Nc1sccc1C(O)=O	35,7
73	COc1ccc(cc1OC)-c1csc(C(O)=O)c1NC(=O)NCc1ccccc1Cl	45,6
74	COc1ccc(cc1)-c1csc(C(O)=O)c1NC(=O)NCC1CCCCC1	54,5
75	COc1ccc(cc1OC)-c1cc(NC(=O)NC2CCCCCC2)c(s1)C(O)=O	57,4
76	Oc1ccccc1C1NCc2ccsc2-n2cccc12	90,9
77	COC(=O)C1C(c2ccc(o2)N(=O)=O)C(C(=O)OC)=C(C)N=C1C	74,0
78	COC(=O)C1C(C(C(=O)OC)=C(C)N=C1C)c1cc(O)ccc1N(=O)=O	56,3
79	O=C1NCc2cccn2-c2ccsc12	49,7
80	COC1=NC(O)c2cccn2-c2sccc12	60,1
81	COC(=O)c1scc(c1NC(=O)N1CCCCC1)-c1ccc(Cl)cc1	101,1
82	Cc1ccc(cc1)S(=O)(=O)NC(=O)Nc1cc(sc1C(O)=O)-c1ccc(Cl)cc1	47,0
83	COC(=O)c1sc(cc1NC(=O)NCCc1ccccc1)-c1ccc(Cl)cc1	121,6
84	NC(=O)NN=C1C(=O)C(=O)c2ccccc12	38,2
85	NC(=O)NN=C1C(C(=O)c2ccccc12)c1ccccc1	57,9
86	OC1C2NCOC2=Nc2c1oc1ccccc21	49,9
87	O=C1NNCc2c1oc1ccccc21	39,5
88	NNC(=O)c1oc2ccccc2c1CN(CCO)CCO	31,5
89	O=C1NC(=Nc2nc[nH]c12)C1CCCC1	29,6
90	CCOC(=O)NC(=Nc1ccc(I)cc1)N(CC)CC	80,5
91	Fc1ccc2N=C(NC(=O)c2c1)N1CCCC1	41,0
92	OC(=O)COc1cccc(c1)C(CC(O)=O)NC(=O)C(F)(F)F	16,6
93	COC(=O)c1c(oc2ccccc12)C(N)=O	56,1
94	NC(=O)CS(=O)Cc1ccccc1	26,8
95	OC(=O)C(O)(C1CCCCC1)c1ccc(cc1)C(F)(F)F	59,1
96	NC(=N)NCc1ccc2OCOc2c1	26,3
97	CCOC(=O)NC(=S)Nc1cc(C)c2[nH]c3ccc(Br)cc3c2c1C	102,2
98	COC(=O)c1sccc1NC(=O)NS(=O)(=O)c1ccc(C)cc1	48,9
99	O=C1NC(=O)c2sccc2N1COCc1ccccc1	61,4
100	NC1=NC(=O)N(C2OC(CO)C(O)C2O)c2ccsc12	11,5
101	O=C1NC2=C(CCS2)C(=O)N1	12,9
102	O=C1NC2=C(C(CS2)c2ccccc2)C(=O)N1	40,8
103	Clc1nc2cccnc2c2cccnc12	58,5
104	CCCCNc1nc2ccccc2c2n(CC(C)C)cnc12	95,8
105	O=C1OCc2cc3OCOc3cc12	49,1
106	C1CCN2C(C1)CNc1ccccc21	80,7

189

107	C[N+](C)(CC(O)=O)Cc1cccc(I)c1	31,7
108	CC(C)c1cc(C(C)C)c(c(c1)C(C)C)S(=O)(=O)NCCc1c[nH]c2ccccc12	121,0
109	O=N(=O)c1ccc(cc1)N1CCCCC1	97,1
110	S=C1CCCCCN1	38,4
111	S=C1CCc2ccccc2N1	66,8
112	CC(=O)c1c2OC3=CC(=O)C(=C(C)NCCCCN)C(=O)C3(C)c2c(O)c(C)c1O	62,3
113	CCOC(=O)C(Cc1ccccc1)(Cc1ccccc1)C#N	101,5
114	OC(=O)C(Cc1ccccc1)Nc1nnn[nH]1	14,2
115	CC(C)(CC(N)=O)SCc1ccccc1	62,3
116	CC(C)(CC(=O)Nc1ccccc1NC(=O)CC(C)(C)S(O)(=O)=O)S(O)(=O)=O	18,4
117	CC1C(=O)CCC(C)(C)C1=C1OCC(C)(C)CO1	86,6
118	CC(C)(C)OC(=O)C(Cc1ccc(O)cc1)NCc1ccccc1	89,5
119	CCN(CC)CCCNc1nc2ncnc2c(NCc2ccccc2)[nH]1	38,2
120	CCCCCC(O[Si](c1ccccc1)(c1ccccc1)C(C)(C)C)C=O	109,0
121	Brc1ccc(cc1)S(=O)(=O)Nc1ccnn1C1CCCCC1	63,9
122	OCCCOc1ccc(C(=O)c2cccs2)c(Cl)c1Cl	79,6
123	OS(=O)(=O)c1ccc(cc1)C#N	16,6
124	OCCCOc1ccccc1CN1CCc2sccc2C1	67,9
125	O=C1C(=O)c2ccccc2C=C1N(=O)=O	32,3
126	O=C(NN(C(=O)c1ccccc1)c1ccccc1)c1ccccc1	82,3
127	Oc1c(Cl)c(NCC=O)c(O)c2ccccc12	54,8
128	CC(=O)Oc1c(C)c2CC(C)(C)Oc2c(C)c1C	105,8
129	Cc1c(C)c2OC(C)(CO)CCc2c(C)c1O	62,8
130	OCCCCN1C(=O)c2ccccc2C1=O	47,3
131	FC(F)(F)CCN1C(=O)c2ccccc2C1=O	80,2
132	Brc1ccc(cc1)S(=O)(=O)Nc1ccnn1-c1ccccc1	53,2
133	Nc1ccc(cc1)S(=O)(=O)N(CCCCC(O)=O)c1ccnn1-c1ccccc1	40,1
134	CCOC(=O)C(O)(Cc1ccccc1N(=O)=O)C(=O)OCC	81,8
135	CCCCS(=O)(=O)Cc1nc2C(=O)c3ccccc3C(=O)c2nc1CS(=O)(=O)CCCC	85,9
136	ClC(Cl)(Cl)c1nc2ccc(cc2s1)N(=O)=O	105,4
137	COc1ccc2NC(C3Cc4ccccc4C3c2c1)C(F)(F)F	105,1
138	NC1(COc2cccc(O)c2C1)C(O)=O	11,5
139	CCS(=O)CC1CCC(O1)N1C=C(C)C(=O)NC1=O	21,6
140	CS(=O)(=O)OC1C(CCl)OC(C1OS(C)(=O)=O)n1cnc2c(N)ncnc12	50,3
141	CCOC(COc1ccc2OC(=COc2c1)C(O)=O)OCC	41,9
142	CC1CCc2c(Br)c(F)c(Br)cc2N1C=O	99,5
143	CC1OC(C)(C)CCC1OC(=O)c1cc(cc(c1)N(=O)=O)N(=O)=O	103,0
144	OC(=O)C#Cc1ccc2OCOc2c1	27,5
145	CC(C)(C)Nc1nc(nc2ncccc12)-c1ccccc1	87,0
146	CNc1ccc(Cl)cc1C(=O)NCCCN1CCCCC1	51,0
147	CC(=O)C1C(C)=C(C(C)=C1C(C)=O)c1ccccc1	84,3
148	CCOC(=O)CNc1cc(C)nc(C)n1	19,5
149	CCN(CC)CCCC#CC(O)(c1ccccc1)c1ccccc1	56,4
150	CCCC(=O)NCCC1CCOc2ccc(OC)nc12	46,3
151	CCS(=O)c1ccccc1	40,6
152	COc1cccc2OCC3(CCCN3)Cc12	53,2
153	FC(F)(F)C(Cc1ccccc1)N1CCOCC1	98,7
154	CN(C)c1ccc(cc1)C(O)c1ccc(I)cc1	97,1
155	COc1c(O)cc2CCN(C3Cc4ccccc4-c1c23)C(=O)C(F)(F)F	93,1
156	COc1cc2CCNC(Cc3ccc(O)cc3)c2cc1O	29,7
157	CN1CCC(COC(=O)c2cccc3ccccc23)CC1	55,9
158	FC(F)(F)Oc1ccccc1NC(=O)OCCN1CCCCC1	61,2
159	CN(C)S(=O)(=O)N1CCN(CC1)c1ccnc(n1)C(N)=O	35,6
160	CC(NC(C)=O)c1cccc2ccccc12	62,9
161	CCC(=O)NCC(c1ccccc1)c1ccccc1	74,9
162	N#CC(=Cc1ccc(OC2CCCCO2)cc1)C#N	98,0
163	COc1cnc2n(c(cc2c1)[Sn](C)(C)C)S(=O)(=O)c1ccccc1	118,3

164	OCC(C(Cc1ccccc1)Cc1ccccc1)C(O)=O	44,6
165	Cc1ccc(cc1)S(=O)(=O)OC1CC(OCc2ccccc2)C(O)C(C1)OS(=O)(=O)c1ccc(C)cc1	77,7
		191
166	CN1C2CCC1CC(C2)OC(=O)C(CO)c1ccccc1	33,7
167	CC(C(O)c1ccccc1)N(C)C	27,3
168	COc1ccccc1N1CCN(CCC(O)c2ccccc2)CC1	66,8
169	COC1OC2COC(OC2C([N-][N+]#N)C1OS(=O)(=O)c1ccc(C)cc1)c1ccccc1	107,9
170	OC(C1CC2CCN1CC2C=C)c1ccnc2ccccc12	40,6
171	CCCC(=O)Nc1cccc(c1)-c1csc(NC)n1	68,1
172	O=C(C1CCCCC1)c1ccc2CCNc2n1	91,4
173	CN1C(N)=NC(=Cc2c[nH]c3ccccc23)C1=O	43,7
174	C[Si](C)(C)C(=Cc1ccccc1)C(O)=O	52,9
175	CC1(C)OCC(CSc2nc3ccccc3o2)O1	91,1
176	CC(NCC(Cc1ccccc1)NS(=O)(=O)c1ccccc1C(F)(F)F)c1cccc2ccccc12	114,7
177	COC1OC2COC(OC2CC1C=C)c1ccccc1	97,3
178	CCCC(=O)NCCCc1cc(OC)cc(OC)c1	67,3
179	CC1(C)OCC2OC3(COCC=C)NC(=S)OC3C2O1	69,2
180	Cc1ccc(cc1)S(=O)(=O)N1CC1(C)C#N	76,3
181	CC(=O)OCC1CC(OC(C)=O)C(C1)n1nnc(C(N)=O)c1I	50,9
182	C=Cc1cnc2NC(=O)Cc2c1	38,9
183	CCC1C(Cc2c[nH]c[n+]2C)COC1=O	37,4
184	CC(C)[Si](C(C)C)(C(C)C)n1ccc2c(Cl)ccnc12	62,4
185	CN1C(=O)C(C)(C)c2cc(cnc12)C(C)=O	50,7
186	CC(=O)OCC12C(CC=CC1=O)C(O)CC2=O	49,1
187	OCc1ccc(Cl)nc1	31,0
188	CCOc1nc(NC(C)(C)C)c2cccnc2n1	64,0
189	OC(=O)c1cc(O)c2C(=O)c3c(O)cccc3C(=O)c2c1	40,1
190	CC1=C(C(=O)N(N1)c1ccc(cc1)N(=O)=O)N(=O)=O	27,0
191	OCC1OC(OC2OC=C3C(CCNC3=O)C2C=C)C(O)C(O)C1O	23,0
192	COc1c2OCOc2cc2CCN(C)C(O)c12	27,3
193	CN(C)CCc1ccc(O)cc1	15,7
194	COc1ccc(C(=O)Cc2c(CCN(C)C)cc3OCOc3c2OC)c(C(O)=O)c1OC	36,6
195	CCC1=C(C)C(=O)N(CCN(C)C)N=C1	29,3
196	O=C1NN=C(C=C1Cc1ccccc1)c1ccccc1	78,7
197	CN1N=C(C=CC1=C(C#N)C#N)c1ccccc1	91,2
198	COc1ccc2C(O)C=C(Oc2c1)c1ccccc1	79,0
199	OC(=O)C1CCN(CC1)c1[nH]cnc2ncnc12	12,9
200	Cc1cccc(C=Cc2cccc(c2)C(F)(F)F)n1	108,0
201	CN(C)CCCN1N=C(c2ccccc2)c2ccccc2CC1=O	51,5
202	CN1N=C(Cc2ccccc2C1=O)c1ccccc1	81,1
203	O=C1CCCC(=NN1CCN1CCOCC1)c1cccc(c1)N(=O)=O	60,8
204	Clc1ccc(cc1)C1=NNC(=S)CCC1	79,2
205	CC1=C2CCCCC2(O)OC1=O	43,5
206	CC(C)C1=CC(=NNC1=O)c1ccc2OCOc2c1	67,2
207	OC(=O)C1CCCN(C1)C(=O)c1ccccc1	24,2
208	O=C1Nc2ccccc2C=C1N1CCCCC1	71,8
209	OC(C1CCCCC1)c1cnc2ccccn12	60,5
210	COc1ccc(CNc2nccs2)cc1OCCO	45,3
211	COc1ccc(C=CC(=O)N2CCOCC2)cc1OC	48,8
212	CCCCOc1cccc(c1)C1CCNC1	56,5
213	OC(=O)CCC(=O)c1ccc2OCCc2c1	27,5
214	FC(F)(F)c1nc(N2CCCCC2)c2ncn(Cc3ccccc3)c2n1	117,1
215	COc1ccc(cc1OC)C#CC(O)=O	25,0
216	O=S(=O)(c1ccccc1)n1ccc2ccccc12	96,1
217	CC(C)(C)OC(=O)N1C(Cc2ccccc12)C(O)=O	37,1
218	CCOC(=O)CCC(=O)c1ccc2nccc(O)c2c1	41,0
219	CCOC(=O)CCC(=O)c1ccc(O)cc1	59,2

220	CCOC(=O)CN1N=C(CCC1=O)c1ccc(OC)c(OC)c1	63,8
221	OC(=O)c1cnc2ccc(nn12)-c1ccccc1	25,8
222	CC(C)(C)OC(=O)NC(CCCNC(=N)NN(=O)=O)C(O)=O	25,3
223	Ic1c(nc2ncccn12)-c1ccccc1	64,0
224	COC(=O)C=CCN1C(COC1=O)c1ccccc1	64,3
225	Cc1nccc2cc3OCOc3cc12	58,1
226	NC(=O)C1CCc2cc(Cl)ccc12	56,0
227	C1Oc2cc3CCNCc3cc2O1	25,3
228	COC(=O)C1CC(=O)c2ccccc12	59,9
229	NC(=O)Cc1ccc(Cl)cc1	47,7
230	O=C1CSCCN1	12,7
231	Clc1ccc(Oc2ccc(cc2)N(=O)=O)c(Cl)c1	112,1
232	CCCN1CCCC1=CN(=O)=O	41,6
233	N#CCN1CCC(Cc2ccccc2)CC1	87,6
234	CC(C)C1COC(=O)N1C(=O)Cc1ccc(Cl)cc1	94,8
235	COC1=CC(=O)OC(CCc2ccccc2)C1	77,1
236	NNC(=O)C1CCCCC1	30,6
237	COC(=O)CC1C(CCN1S(=O)(=O)c1ccc(C)cc1)C(=O)OC	79,3
238	Cc1ccc(cc1)S(=O)(=O)NCCC1CCC=C1	94,0
239	OC(c1ccccc1)(c1ccccc1)c1ccccc1	100,7
240	O=N(=O)c1ccc(cc1)S(=O)(=O)N1CCC2CC=CC12	88,9
241	OC1(CCCc2ccccc2)CCCCC1	103,9
242	COC1=CC(C)(C=C(C1)OC)C(N)=O	19,7
243	CC(COc1ccccc1)NC(C)C(O)c1ccc(O)cc1	50,4
244	FC(F)(F)C1CC(=O)CC(=C1)c1ccco1	77,4
245	FC(F)(F)C1CC(=O)CC(=C1)c1ccc2ccccc2c1	100,7
246	COC(=O)C1Cc2ccccc2CN1	51,2
247	COc1cc(cc(OC)c1O)C(O)C(CO)Oc1c(OC)cc(CO)cc1OC	34,7
248	COC1=C(Oc2ccccc2C1=O)c1ccccc1	88,6
249	CC(=O)Oc1cc2OC(=Cc3ccccc3)C(=O)c2c(OC(C)=O)c1	59,7
250	CC(=O)Oc1ccccc1C(=O)Oc1cccc(OC(=O)c2ccccc2OC(C)=O)c1C(C)=O	94,0
251	O=C1Oc2ccccc2C1=Cc1ccccc1	101,2
252	[Na]OC(=O)C1(Oc2ccccc2C1=O)c1ccccc1	36,8
253	COc1cc(CC(C)(C(N)=O)c2ccccc2)cc(OC)c1OC	63,0
254	OC1(Cc2ccccc2)Oc2ccccc2C1=O	73,1
255	OCc1cn(c2cc(F)ccc12)S(=O)(=O)c1ccccc1	78,2
256	COc1ccc2cc(ccc2c1)C(O)C(F)(F)F	83,3
257	CC1NC(OC1c1ccccc1)C(F)(F)F	86,1
258	COC(OC)C12CCC=CC1=CC(=O)CC2	67,1
259	FC(F)(F)C1CC2CC(=O)CC2=CN1C(=O)OCc1ccccc1	83,9
260	OC(=O)CC(c1c[nH]c2ccccc12)C(F)(F)F	39,3
261	Cc1noc(CC(C)(C)O)c1C(=O)NN	24,6
262	CC1(C)OC(=O)C(=Cc2c[nH]c3ccccc23)C(=O)O1	72,2
263	COc1ccc2[nH]cc(CC(C(O)=O)C(O)=O)c2c1	14,7
264	CC(=O)NCCc1cn(c2ccc(O)cc12)S(=O)(=O)c1ccccc1	58,3
265	OCCNC(=O)c1cccc2cnccc12	21,1
266	O=C(Nc1ccncc1)N1CCNCC1	9,9
267	OCCCNS(=O)(=O)c1cccc2cnccc12	35,3
268	CC(C)(C)OC(=O)NCCCC(=O)Nc1ccncc1	45,4
269	Fc1ccccc1N1CCN(Cc2ccc3NC(=O)Oc3c2)CC1	74,1
270	Nc1cc2C(CC(=O)c2cc1O)c1ccc(Cl)cc1	67,9
271	OCc1ccc2NC(=O)Oc2c1	22,7
272	CN(C)CCN1C(=O)[Se]c2ccccc12	51,8
273	CC1C(Br)C(=O)c2cc3OC(=O)N(C)c3cc12	71,4
274	CC(CN1CCN(CC1)c1cccc(c1)C(F)(F)F)C(O)c1ccc2N(C)C(=O)Oc2c1	98,1
275	NS(=O)(=O)c1ccc(cc1)N(CCO)C(=O)c1ccc(Cl)c(Cl)c1	82,3
276	OC(=O)CNc1ccc(cc1S)C(=O)c1ccccc1C(O)=O	13,3

277	O=C(CCN1C(=O)Oc2ccccc12)N1CCCCC1	65,5
278	CS(=O)(=O)c1ccc(cc1)N1C(CNC1=O)c1ccccc1	54,3
279	CN1C(=O)Sc2cc(CCNC(=O)C(F)(F)F)ccc12	68,7
280	CN1c2ccccc2C(NCCc2ccc(O)cc2)c2ccccc2S1(=O)=O	73,8
281	CSCCC(NCc1cc(Br)cc(Br)c1N)C(O)=O	62,3
282	Cc1ccc(cc1)S(=O)CC(=O)c1ccccc1	69,2
283	CN(C(=O)C1=CCCCC1)c1ccccc1I	93,3
284	CCn1c(C(c2ccccc2F)n2ccnc2)c(C)c2ccccc12	94,1
285	CON(C)C(=O)c1cn(c2ccc(Br)cc12)S(=O)(=O)c1ccccc1	102,6
286	NCC(O)(Cn1cncn1)c1ccc(Cl)cc1Cl	36,4
287	FC(F)(F)c1ccc(Cn2cc(Cn3ccnc3)c3ccccc23)cc1	87,1
288	CN1CCN(C)C1c1ccccc1O	62,7
289	CC1(C)OC(C(O1)C(=O)N1CCOCC1)C(=O)N1CCOCC1	39,1
290	OC(C(O)C(=O)NCc1ccccn1)C(=O)NCc1ccccn1	22,2
291	OC(C(O)C(=O)Nc1ccccc1)C(=O)Nc1ccccc1	53,1
292	OCC(NC(C#N)(c1ccccc1)C(F)(F)F)c1ccccc1	87,9
293	CCOC(=O)CC1(NC(CO1)c1ccccc1)C(F)(F)F	94,4
294	CCOC(=O)c1ccccc1C#N	73,7
295	CC1C2C(=O)C(=O)C3(C)C(O)CC4OCC4(OC(C)=O)C3C(OC(=O)c3ccccc3)C(O)(CC1=O)C2(C)C	58,9
296	OC1(CC=C)c2ccccc2-c2ccccc12	85,5
297	CN(C)CCn1c(Cl)c(Cl)c2c1C=NN(CO)C2=O	40,2
298	CC1Cc2c(C)nc(C)nc2N=N1	28,1
299	COc1ccc(cc1OC)-c1ccc2cc3OCOc3cc2c1N(C(C)=O)C(C)=O	85,0
300	COc1cc2C(N=O)C(CCc2cc1O)c1ccc(OC)c(OC)c1	61,6
301	COc1ccc(cc1)-c1ccc2cc(OC)ccc2c1C(N)=O	71,3
302	CN(C)c1ncnc2n(Cc3ccccc3)c(C)cc12	78,5
303	CC(=O)CC1=C(O)N=C(NC1=O)c1ccccc1	12,4
304	Cc1cc2c(NCc3ccccc3)ncnc2n1CCCO	57,0
305	Cc1[nH]ccc1C(=O)NN=Cc1ccc(o1)N(=O)=O	49,0
306	S=C1NC=Nc2n[nH]cc12	19,5
307	O=C1OC(C2=C1C=CNC2=O)c1ccsc1	43,4
308	O=C1OC(C2=C1C=CNC2=O)c1ccc2ccccc2c1	61,3
309	Brc1cc2OCOc2cc1C=O	79,0
310	COC(=O)C=Cc1c(COC(C)=O)[nH]c(COC(C)=O)c1C=CC(=O)OC	69,1
311	Cc1cc2[nH]c3CCCCc3c2c(Cl)n1	40,7
312	Brc1csc(C(=O)c2ccccc2)c1Br	105,4
313	O=C(NC1CC=CC1)c1ccccc1	61,5
314	Cn1ccc2c(nccc12)N1CCNCC1	26,8
315	N#Cc1cocc1C#N	45,0
316	COc1ccc(OC)c2C(C)N(CC(O)c12)S(=O)(=O)c1ccc(C)cc1	85,9
317	CC(CO)NCc1ccccc1	21,2
318	CSC1=NC(CCO)=CC(=O)N1	16,3
319	CCOC(=O)CC1=CC(=O)NC(SC)=N1	36,4
320	CN1NC(=O)C(Cc2occc2C(=O)NN)=C1C	19,3
321	CCOC(=O)C(Cc1occc1C(=O)OC)C(=O)c1ccccc1	91,8
322	Nc1nc(CCl)nc(NCC2CCOC2)n1	37,0
323	CCOC(=O)Cc1ccc2CCc3cccc1c23	104,9
324	Cn1ccc2c1C(O)CNC2=O	12,6
325	CC(C)N(C(C)C)C(=O)c1cnccc1C(C)(O)c1cc2c(Cl)nccc2n1C	65,3
326	CC1(OC(=O)c2cnccc12)c1ccco1	57,0
327	COc1cccc(O)c1Cc1occc1C(O)=O	41,2
328	CN(C)CCCNC(=O)c1ccoc1COc1ccc(Cl)cc1	53,3
329	Cc1cc2occc2c(n1)N1CCCC1	73,6
330	O=C1CCc2[nH]c3C=CNC(=O)c3c2C1	22,0
331	O=C1NC2=C(CCC2)c2occc12	42,3
332	CCOC(=O)C1=C(Cl)C=C(C)NC1=O	46,1

193

333	ClC1=C(C#N)C(=O)NC=C1	31,1
334	O=NC1CCOC1=O	7,0
335	CCOC(=O)C(Cc1c(C)c(C#N)c(OC)nc1OC)(NC(C)=O)C(=O)OCC	80,1
336	COc1cc2c(C)c(C)oc2cc1CCNC(=O)c1ccccc1	92,7
337	COc1cc2ccccc2cc1O	72,1
338	Oc1ccc2ccccc2c1Cl	80,9
339	Oc1ccccc1C#N	47,4
340	C1Cc2cc3ccccc3nc2O1	58,4
341	COC(=O)c1ccc(O)c(Cl)c1	59,4
342	COC(=O)c1ccc(Cl)c(c1)N(=O)=O	83,8
343	NC(=O)c1cc2ccccc2o1	45,2
344	N#Cc1cc2ccc3ccoc3c2o1	89,7
345	OC1=CC(=O)NC(Cl)=C1c1ccccc1	30,8
346	COc1ccc2occ(C(C)C)c2c1	106,3
347	NC(=O)c1cc(cc(c1)N(=O)=O)N(=O)=O	47,7
348	CCN(CC)CCOc1cc(C)c(Cl)cc1C(C)C	89,7
349	CC(=O)Nc1ccc(cc1)S(=O)(=O)Nc1nc2ccccc2[nH]1	57,1
350	CC1=CC(=O)C=C(O1)N1CCCCC1	48,5
351	CC(C)C(=O)Oc1ccccc1C(=O)Nc1ccc(I)cc1	100,4
352	OCc1cc(Br)cc(Br)c1O	69,9
353	Oc1cccc2CCCCc12	80,9
354	CN(C=CN(=O)=O)C1CCCCC1	64,8
355	COc1cc(cc(OC)c1OC)C(CN(=O)=O)CN(=O)=O	71,7
356	ClCC1=CC(=O)Oc2ccc3ccccc3c12	85,9
357	OC(=O)C=Cc1ccc2CCCCc2c1	47,5
358	Clc1ccc2OCC(C=O)=Cc2c1	86,8
359	Cc1c(C=O)oc2cc(C)ccc12	84,3
360	CCN(CC)CCOc1ccc2cc(oc2c1C(C)=O)C(C)=O	41,6
361	COc1cc(NS(C)(=O)=O)ccc1NC(C)=O	32,7
362	O=C1NC=CC2=C1NC(=O)c1ccccc1O2	39,2
363	CCNC(=O)c1ccccc1O	61,2
364	Cc1cccc(n1)-c1cccc(C)n1	80,6
365	OC(=O)c1cc2ccccc2c(c1O)-c1c(O)c(cc2ccccc12)C(O)=O	40,3
366	CC(C)(C)NCC(O)COc1ccc(I)cc1	37,3
367	CC(C)(C)OCC#CC(N(O)Cc1ccccc1)c1ccccc1	105,3
368	CCC(N(O)Cc1ccccc1)c1ccco1	90,2
369	CC(C)CC(N(O)C(=O)OC(C)(C)C)c1cccs1	100,7
370	CC(=C)C1CC=C(C)C(O)(C1)C(C)(C)O	80,3
371	COc1ccc(C(C)C)c2C(O)CC(C)c12	90,5
372	CC(C)C1=CC2=C(Cl)C(=O)CC2C(C)C=C1	107,2
373	CCC1(OCc2ccccc2)C=C(COC1=O)C(O)=O	38,7
374	COC(=O)C1=C2CCCN2C(=O)C(OS(C)(=O)=O)=C1	49,2
375	COc1ccc(cc1)C(NOCCCCOCC#C)(c1ccccc1)c1ccc(OC)cc1	115,5
376	ON1S(=O)(=O)c2ccccc2S1(=O)=O	20,7
377	ON1C(=O)c2ccc(Cl)cc2C1=O	33,0
378	Cc1ccc(cc1)S(=O)(=O)OCC1(C)COC(C)(C)OC1	53,6
379	CC1(C)SSC(C)(C)C=NC(CCCCC#N)CN=C1	83,3
380	CC(C)(SSC(C)(C)C=NCc1ccccc1)C=NCc1ccccc1	84,9
381	OC(COC(=O)c1ccccc1)c1ccccc1	79,3
382	OC(=O)CCC12CCCCC1=Nc1ccccc21	30,8
383	C1CCN2CCc3c([nH]c4ccccc34)C2C1	49,3
384	OC(=O)Cc1c2CCC(=O)n2c2ccccc12	27,3
385	OCCCc1[nH]c2ccccc2c1CCO	45,0
386	COC(=O)C1=CC(=O)C=CO1	24,4
387	CC1=CCCC(C)(C)C1C(O)=O	32,7
388	COC(=O)N(C)C(=O)c1ccccc1	66,7
389	CC1(C)CCCC(C)(C)N1C(=O)c1ccccc1	105,6

390	Ic1ccc(OCC2CO2)cc1	88,7
391	C=CCc1ccccc1C#N	86,8
392	O=S1CCCSC1	14,5
393	COc1ccccc1-c1c(CO)c(CO)c(-c2ccccc2OC)c2ccccc12	93,5
394	Cc1cc(cc(C)c1O)C1=CC(C)(C)NC(C)(C)C1	43,2
395	CN(C)S(=O)(=O)c1ccc(C)cc1	73,9
396	CN(C)c1cccc2c(cccc12)S(=O)(=O)NCCCN	44,2
397	CC(=O)OCC1OC(C2OC(C)(C)OC12)n1cnc2c1N=CNC2=O	39,9
398	CC(C)(C)OC(=O)NCCCNC(=S)Nc1ccc(C2=C3C=CC(=O)C=C3Oc3cc(O)ccc23)c(c1)C(O)=O	43,2
399	CCOC(=O)CN(CC(=O)OCC)Cc1cc(ccc1O)N(=O)=O	91,1
400	CC(N(C1CC(C)(C)N(O)C(C)(C)C1)C(=O)Cc1ccc(cc1)N(=O)=O)c1ccccc1	97,5
401	NC1(CCc2ccccc2C1)C(O)=O	18,9
402	COC(=O)c1sc(cc1NC(=O)NCc1ccco1)-c1ccc(OC)cc1	63,3
403	O=C1CC(CO1)c1ccccc1	64,0
404	CC1(O)CC(N(Cc2ccccc2)C1=O)c1ccccc1	64,9
405	O=S1(=O)NC(CCO1)c1ccccc1	67,4
406	CCOC(=O)NC(=S)NNC(=O)c1cc2c(cn1)[nH]c1ccccc21	68,5
407	Oc1ccc2OC(CC(=O)c2c1O)c1ccccc1	79,0
408	FC(F)(F)S(=O)(=O)Cc1ccccc1	83,5
409	CCc1oc2c(O)cccc2c1C(=O)c1ccccc1	84,1
410	C=CCN(CCC#N)Cc1ccccc1	86,0
411	COc1ccc(cc1)C(=O)CC(=O)c1c(O)cc(OC)cc1OC	89,1
412	O=C(CCCC(=O)OCc1ccc(cc1)N(=O)=O)OCc1ccc(cc1)N(=O)=O	91,4
413	N#CCC(OCc1ccccc1)C(CC#N)OCc1ccccc1	91,5
414	COc1ccc(C(=O)Nc2ccccc2C(F)(F)F)c(Cl)c1Cl	91,7
415	ClCc1ccc(CCl)cc1	91,8
416	OC(Cc1ccc(cc1)N(=O)=O)c1ccc(cc1)C(F)(F)F	94,9
417	CCc1ccc2[nH]c3C(N(CCc3c2c1)C(C)=O)c1ccccc1	96,0
418	CC(=O)OC1CC(OC1COC(=O)c1ccc(OCCCCC(=O)OC(C)(C)C)cc1)N1C=C(I)C(=O)NC1=O	96,3
419	CNC(CCCc1ccccc1)C(F)(F)F	98,8
420	COc1ccc(CN(O)C(C#CC(=O)OC(C)(C)C)C(C)C)c(OC)c1	108,0
421	COc1ccc(C=CC(=O)c2ccccc2OCc2ccccc2)cc1	109,0
422	COC(=O)c1sc(cc1NC(=O)N(CCC#N)Cc1ccccc1)-c1ccc(OC)cc1	110,1
423	CC(N=C(c1ccccc1)C(F)(F)F)c1ccccc1	115,0
424	OC(=O)CN(CC(O)=O)Cc1cc(F)ccc1O	14,2
425	CC(C)C1NC(=NN=C1NCCN1CCCCC1)c1ccccc1	35,9
426	CCN(CC)CCNC(=O)C1CCCN1c1ccnc2cc(Cl)ccc12	41,9
427	Oc1c(sc2ccccc12)C(=O)Nc1ccccc1	54,3
428	CS(=O)(=O)Nc1cc(ccc1O)C(=O)c1ccccc1	54,5
429	Oc1ccc2oc(cc2c1C=O)N(=O)=O	61,9
430	CCOC(=O)N(C(C)C)C(C)Cc1cc(I)c(O)c(Cl)n1	64,6
431	NNc1ncnc2sc3CCCCc3c12	50,7
432	CC1(C)NNc2ncccc2-n2cccc12	89,3
433	COc1cccc(C(=O)NN)c1O	28,7
434	CN(C)CCON=C1c2cccn2-c2c(csc12)-c1ccc(Cl)cc1	101,7
435	OS(=O)(=O)c1ccc2-c3ccc(cc3C(=O)c2c1)S(O)(=O)=O	9,6
436	CCC1(O)C(=O)OCC2=C1C=C1N(Cc3cc4ccccc4nc13)C2=O	50,9
437	O=C1N(C(=O)c2ccccc12)c1ccccc1	79,5
438	CCC12CCCN3CCC4(C(C(C1)N(=O)=O)N(C(=O)OC)c1ccc(Br)cc41)C23	132,2
439	COc1ccc2N(CC(=O)NC(N)=N)C=C3C(=O)NC(=N)N=C3c2c1	23,6
440	COC(=O)CC1CC(=O)c2cc(OC)c(OC)cc12	54,7
441	CC1C=CC(Cl)C2=C1SC1=C(C(=O)NC(C)=C1)C2=O	56,7
442	CC1=CC(O)=C(C(=O)OC2CCCCC2)C(=O)N1	62,1
443	CNc1nccn2c(cnc12)C(=O)C(C)(C)C	67,2
444	CC(C)(C)OC(=O)N1CCc2ncc(cc2C1)N(=O)=O	79,5

195

445	CC(C)(C)OC(=O)N1CCc2nnc(cc2C1)-c1ccccc1	80,1
446	COc1cc2C(=O)CC(c2cc1NC(C)=O)c1ccc(Cl)cc1	80,2
447	CCOC(=O)c1c(Cl)nccc1C1=NC(C)(C)CO1	81,4
448	FC(F)(F)c1cccc(c1)-c1cc(C#N)c(NCCN2CCOCC2)nn1	82,1
449	Nc1ccc(CCN2CCN(CC2)c2cccc(c2)C(F)(F)F)cc1O	89,5
450	C=CCOc1ccc(cc1OCC=C)C(=O)NC1CCCCC1	92,1
451	Brc1ccc(cc1OCc1ccccc1)C(=O)OCc1ccccc1	92,9
452	CN1C(=O)[Se]c2cc(ccc12)C(=O)c1ccc(Cl)cc1	97,6
453	CN1c2cccnc2N(C)c2nccc(I)c12	102,8
454	CCC(C(=O)OC1CC(C)(C)N(O)C(C)(C)C1)c1ccccc1	110,1
455	CCOC(=O)C(C)Oc1c(C=O)cc(Cl)c2ccccc12	110,9
456	COc1ccc2c(C)c3ccccc3c(C)c2c1	122,4
457	CC(=O)N1C(=O)Oc2cc3C(=O)C=C(c3cc12)c1ccc(Cl)cc1	103,6
458	CCc1c(Cc2[nH]c(C=O)c(CC)c2CC)[nH]c(C=O)c1CC	93,7
459	COC(=O)c1c(OC)ccc2cc(oc12)C(=O)c1ccc(Cl)cc1	96,2
460	Oc1c(C=O)ccc2c1ccc1ccccc21	108,3
461	OC1=C(C(=O)c2ccccc2)C(=O)OC2=C1CCCC2	33,0
462	Oc1ccc2oc(cc2c1)C(=O)c1cc2cc(O)ccc2o1	62,1
463	CCN(CC)C1=Nc2c(C)c3c4cc(Br)ccc4[nH]c3c(C)c2C(=O)N1	69,2
464	Oc1ccc2n3C(=O)c4ccccc4-c3cc2c1	73,8
465	CC(CO)C(=O)c1ccc2ccccc2c1O	80,5
466	CN1c2ccccc2C(=O)c2ccc(c(C)c12)N(=O)=O	83,3
467	O=C(Nc1cccn1-c1ccccc1)OCc1ccccc1	90,6
468	COc1cccc2cc(C=CC(=O)N3CCN(Cc4ccccc4)CC3)oc12	92,1
469	CCOC(=O)N1CCN(CC1)c1nc2ccsc2n2cccc12	92,8
470	O=N(=O)C=C(c1ccccc1)c1ccccc1	97,8
471	CCOC(=O)c1ccc(cc1)-n1cccc1	98,7
472	Clc1ccc(cc1)S(=O)(=O)Nc1ccc(Cl)cc1C(=O)Nc1ccc(Br)cc1	99,0
473	OC(c1cc2cc(Br)ccc2o1)c1ccc(Br)cc1	107,1
474	CC(C)C1CCC(C)CC1OC=CS(=O)c1ccccc1	111,6
475	COC1(CC=C)c2ccccc2-c2ccccc12	113,4
476	COc1cc(C=CC(=O)c2ccccc2O)ccc1OCc1ccccc1	114,6
477	COc1cc(C)c(cc1C(C)C)C(=O)c1ccc(Cl)c(Cl)c1	116,8
478	COc1ccc(cc1)-c1oc-2c(CNC(C)(C)c3cccn-23)c1-c1ccc(OC)cc1	117,1
479	C[Si](C)(C)OC(CC1(SCCCS1)[Si](C)(C)C)C(CC1(SCCCS1)[Si](C)(C)C)O[Si](C)(C)C	138,8
480	CC(C)c1cc(C(=O)C(Br)(Br)Br)c(C)cc1O	105,0
481	COc1cc(OC)cc(OCC(=O)NN=Cc2ccc(s2)N(=O)=O)c1	63,9
482	NC1=C(Oc2ccccc2C1=O)c1ccc(O)c(I)c1	76,6
483	Cc1ccc(OC(CCN2CCN(CCCC(c3ccc(F)cc3)c3ccc(F)cc3)CC2)c2ccccc2)cc1	131,7
484	COc1cccc2OCC(Cc12)(NCc1ccccc1)C#N	96,6
485	COc1ccccc1OC(CCN1CCN(CC1)c1ccccn1)c1ccccc1	94,7
486	O=C1C(Cc2ccccc12)c1ccccc1	94,3
487	COC(=O)C1=C(C(CC(=O)C1)C(=O)N(C)C)C(=O)OC	25,9
488	CC1CC(=O)N(CCO)N=C1	30,8
489	NC1C(O)c2ccccc2C1c1ccccc1	44,1
490	COC(=O)c1sc(cc1NC(=O)NS(=O)(=O)c1ccc(C)cc1)-c1ccc(OC)c(OC)c1	62,3
491	ClC(N=Nc1ccccc1)c1ccccc1	57,2
492	COc1ccc(Cn2c(N)c(C#N)c3CCCCc23)cc1	63,4
493	CCC1(NN=C2N1C=CC=C2n1cccc1)c1ccc(F)cc1	79,6
494	Oc1cccc2OC(O)(CC(=O)c12)c1ccccc1	80,6
495	COC(=O)C1OC(N=C1C(=O)OC)c1ccccc1	79,0
496	CN=Nc1c(O)[nH]c2ccccc12	38,4
497	NNC1=Nc2ccc(Cl)cc2C(=S)N2CSCC12	71,1
498	COc1ccc(NC2c3ccccc3N(C)S(=O)(=O)c3cc(Cl)ccc23)cc1	99,6
499	COc1ccccc1N=NCc1ccccc1	105,5

500	CC1=CC2=C(C)NC(=O)N=C2O1	22,2
501	CC1=Cc2occc2C(=O)O1	49,8
502	CN(C)C=Nc1nnc(cc1C)-c1ccccc1	67,7
503	OCCCCC1COC(O1)c1ccccc1	47,4
504	O=C1CC(CO1)c1ccc(cc1)N(=O)=O	63,2
505	OC(=O)C1CCCCN1C(=O)c1cc2OCOc2cc1N(=O)=O	32,1
506	CC(C)N=Nc1ccnc2cc(Cl)ccc12	68,3
507	Oc1c(Br)oc2ccc3ccccc3c12	62,0
508	CCCCCCCNC(=O)Nc1scc(c1C(O)=O)-c1ccccc1	62,2
509	COC(=O)C1C(c2ccsc2)C(C(=O)OCCCl)=C(C)N=C1C	83,9
510	CCN(C)C(=O)Nc1c(csc1C(O)=O)-c1ccc(F)cc1	43,5
511	COC(=O)C1C(c2ccc(s2)C(=O)NCCN2CCCCC2)C(C(=O)OC)=C(C)N=C1C	47,8
512	CN1C=Nc2sc3CCCCc3c2C1=O	67,1
513	O=C(NC1CC2N(C1)C(=O)c1ccccc1NC2=O)Nc1ccc(cc1)N(=O)=O	54,0
514	O=S(CC1OCCCS1)c1ccccc1	57,8
515	BrCC(=O)Nc1ccccc1-c1ccccc1NC(=O)CBr	73,6
516	COc1ccc(cc1)-c1csc(C(O)=O)c1NC(=O)Nc1ccc(C)o1	47,3
517	CC(C)(CNCc1ccccc1)SCc1ccccc1	104,5
518	Oc1c2ccccc2cc2ccccc12	91,7
519	Brc1csc(CN=O)c1	66,3
520	CCOC(=O)CNC(=O)OCC1c2ccccc2-c2ccccc12	91,6
521	CC(C)(C)[Si](C)(C)OC1C(O)C(C[N-][N+]#N)OC1N1C=CC(=O)NC1=O	76,1
522	OC1CC(=CC(O)C1O)C(O)=O	-35,5
523	OC(=O)C1=CC(=O)C=C(O1)C(O)=O	-27,2
524	O=C1NC(=O)C(=O)C(=O)N1	-25,9
525	C[N+](C)(C)Cc1ccccc1C[Si](C)(C)C	-33,7
526	CC(CCC=C(C)C)CC(N(O)C(=O)OC(C)(C)C)S(=O)(=O)c1ccccc1	12,1
527	COc1cc(N)c(Cl)cc1C(=O)OCCC1CCCCC1	6,7
528	CCOC(=O)Cc1cccc(CBr)c1	16,8
529	COc1ccc2C(=O)OC(=O)c2c1	-21,0
530	N#CC(N1CCN(CC1)c1ccccc1)c1ccccc1	30,2
531	CCCCN(CCCC)CCOC(=O)c1cc(nc2ccccc12)-c1ccccc1	38,0
532	CC(C)C1CC(C)CCC1OC(=O)c1ccccc1C(=O)OC1OCC(=O)CC1C(C(C)=O)C(C)=O	37,5
533	CC(C)(C)OC(=O)N1CC(O)CC1C(=O)OCc1ccccc1	108,2
534	O=C1OC(=O)c2cc3ccccc3cc12	12,6
535	BrCc1cc2ccccc2cc1CBr	48,9
536	CCOC(=O)C(O)c1cnc2ccc(C)cn12	10,6
537	CC(C)(C)c1cc(C=CC(=O)c2ccc3NC(=O)[Se]c3c2)cc(c1O)C(C)(C)C	50,5
538	C[N+]1=Cc2ccccc2CC1	-33,7
539	CC1=[N+](C)OC(C)(C)C1	-34,3
540	Oc1ccc(C=C2N=C(OC2=O)c2ccccc2)cc1	25,6
541	OC(N1CCN(CC1)c1ccccn1)C(F)(F)F	20,3
542	CC1=CC(=O)OC1N1CCOCC1	6,1
543	NC(=O)c1ccc(cc1)C(O)=O	-20,5
544	CC1=CC(O)=CC(=O)O1	-5,9
545	Cc1cc(O)c(cc1N=Nc1ccccc1C(=O)c1ccccc1)C(C)(C)C	66,2

The following CHI test set was used to test our consensus model. Molecules numbered from 1 to 11 are from Huguet2007, those from 12-138 from Valko2011, those from 139-152 from Plassa1998, those from 153 to 161 from Valko1997 and those from 162 to 195 from Camurri2001.

MolNb.	#SMILES	CHI
1	CNC(C)C(O)c1ccccc1	26,73
2	COc1cc(C=O)ccc1O	39,56
3	CCC1(C(=O)NC(=O)NC1=O)c1ccccc1	51,08
4	CC(C)Cc1ccc(cc1)C(C)C(O)=O	55,26
5	COc1ccc2CC3C4CCCCC4(CCN3C)c2c1	58,6
6	Oc1c(Cl)cccc1Cl	69,33
7	Oc1ccc2ccccc2c1	70,97
8	Clc1cccc(c1)N1CCN(CCCN2N=C3C=CC=CN3C2=O)CC1	75,27
9	Oc1cc(Cl)cc(Cl)c1	81,99
10	CNCCC=C1c2ccccc2CCc2ccccc12	96,28
11	CN(C)CCCN1c2ccccc2CCc2ccccc12	100,3
12	NC1=NC(=O)N(C=C1)C1OC(CO)C(O)C1O	-23
13	Oc1cc(O)cc(O)c1	9,05
14	Nc1ccc(cc1)S(N)(=O)=O	10,54
15	NC(=O)c1cnccn1	11,89
16	Oc1ccc(O)cc1	13,47
17	CCC1C(=O)NC(=O)NC1=O	15,99
18	OCc1ccc(O)cc1	20,21
19	CC(=O)Nc1ccc(O)cc1	20,48
20	On1nc2ccccc2n1	23,81
21	OCc1cccc(O)c1	24,82
22	Oc1cccc(O)c1	25,18
23	Cc1cc(NS(=O)(=O)c2ccc(N)cc2)no1	25,62
24	NC(=O)c1ccccc1	29
25	O=C1NC(=O)c2ccccc2N1	29,27
26	OC(=O)c1cccc(O)c1	34,36
27	Cc1cc(C)nc(NS(=O)(=O)c2ccc(N)cc2)n1	37,53
28	CCC1C(Cc2cncn2C)COC1=O	39,82
29	O=C1CCCCC1	43,76
30	CC12CCC(=O)C=C1CCC1C3CCC(C(=O)CO)C3(CC(O)C21)C=O	44,96
31	NS(=O)(=O)c1ccccc1Cl	45,02
32	Oc1ccc(cc1)C#N	45,64
33	Nc1ccc2cc3ccc(N)cc3nc2c1	46,38
34	Nc1ccc(F)cc1	47,7
35	c1n[nH]c2ccccc12	48,67
36	OC(=O)Cc1ccccc1	50,71
37	CC(C)CC1(CC=C)C(=O)NC(=O)NC1=O	51,08
38	Oc1cccc(c1)C#N	51,29
39	COC(=O)c1ccc(O)cc1	51,67
40	CN1c2[nH]c(nc2C(=O)N(C)C1=O)-c1ccccc1	51,7
41	CC12CC(O)C3(F)C(CCC4=CC(=O)CCC34C)C1CCC2(O)C(=O)CO	52,2
42	CCC[N+]([O-])=O	52,41
43	CC(=O)Nc1cccc(c1)[N+]([O-])=O	52,47
44	CC12CC(=O)C3C(CCC4=CC(=O)CCC34C)C1CCC2(O)C(=O)CO	52,58
45	Nc1ccc(cc1)[N+]([O-])=O	52,76
46	OC(=O)C1=CC(=O)c2ccccc2O1	53,04
47	OC(=O)c1cccc(c1)[N+]([O-])=O	54,59
48	OC(=O)c1ccc(F)cc1	54,59
49	OC(=O)c1cccc(F)c1	55,34
50	Oc1ccc(cc1)[N+]([O-])=O	55,52

51	CNCCc1ccccn1	55,58
52	OC(=O)c1ccc(cc1)[N+]([O-])=O	56,03
53	Oc1cccc(F)c1	56,05
54	OC(=O)c1ccccc1O	57,46
55	Nc1cccc(c1)[N+]([O-])=O	57,67
56	CC12CC(O)C3C(CCC4=CC(=O)CCC34C)C1CCC2C(=O)CO	59,82
57	O=C1NC(=O)C(N1)(c1ccccc1)c1ccccc1	61,09
58	Oc1ccccc1Cl	61,13
59	CC1=CC(=O)C=C(C)C1=O	61,25
60	CC12CCC3C(CCC4=CC(=O)CCC34C)C1CCC2(O)C(=O)CO	61,44
61	CCN(CC)CCOC(=O)c1ccc(N)cc1	62,14
62	COc1cc(cc(OC)c1OC)C1C2C(COC2=O)C(O)c2cc3OCOc3cc12	63,38
63	Sc1nc2ccccc2s1	63,59
64	N#Cc1ccccc1	64,06
65	CC12CCC3C(CCC4CC(O)CCC34C)C1(O)CCC2C1=CC(=O)OC1	64,55
66	CC1(C)SC2C(NC(=O)COc3ccccc3)C(=O)N2C1C(O)=O	64,71
67	COC(=O)c1ccccc1C(=O)OC	65
68	Oc1ccccc1[N+]([O-])=O	67,49
69	CCc1ccccc1N	68,78
70	CC12CC(=O)C3C(CCC4=CC(=O)CCC34C)C1CCC2(O)C(=O)C(O)CC(O)=O	69,29
71	CCCC[N+]([O-])=O	69,68
72	COc1ccc2nccc(C(O)C3CC4CCN3CC4C=C)c2c1	70,25
73	[O-][N+](=O)c1ccc(cc1)[N+]([O-])=O	70,39
74	CC12CCC3C(CCc4cc(O)ccc34)C1CCC2O	70,58
75	O=Nc1ccccc1N=O	70,58
76	CC(=O)C=Cc1ccccc1	70,86
77	CCCOC(=O)c1ccc(O)cc1	71,06
78	Oc1ccc(I)cc1	71,49
79	Oc1cccc(c1)C(F)(F)F	71,9
80	CC12CCC3C(CCC4=CC(=O)CCC34C)C1CCC2O	72,11
81	Oc1c(F)c(F)c(F)c(F)c1F	72,28
82	Oc1cccc2ccccc12	73,72
83	Oc1ccc(Cl)c(Cl)c1	75,36
84	OC(=O)CCC(=O)c1ccc(cc1)-c1ccccc1	76,22
85	c1ccccc1	77,82
86	CC(=O)C1CCC2C3CCC4CC(O)CCC4(C)C3C(=O)CC12C	78,88
87	CC12CCC3C(CC=C4CC(O)CCC34C)C1CCC2=O	79,24
88	CCCCC[N+]([O-])=O	79,86
89	C1COc2ccccc2OCCOCCOc2ccccc2OCCO1	80,04
90	CC(=O)SC1CC2=CC(=O)CCC2(C)C2CCC3(C)C(CCC33CCC(=O)O3)C12	81,69
91	CCOC(=O)c1ccccc1C(=O)OCC	81,86
92	CCN(CC)CC(=O)Nc1c(C)cccc1C	85,21
93	CC12CCC3C(CCC4CC(=O)CCC34C)C1CCC2O	85,54
94	COc1ccc2n(C(=O)c3ccc(Cl)cc3)c(C)c(CC(O)=O)c2c1	87,98
95	Oc1ccc(Cl)cc1Cc1cc(Cl)ccc1O	88,26
96	OC(=O)Cc1ccccc1Nc1c(Cl)cccc1Cl	88,63
97	Cc1ccccc1	88,65
98	CCCCCC[N+]([O-])=O	89,27
99	Clc1ccccc1	89,99
100	CN(C)CCOC(c1ccccc1)c1ccccc1	94,69
101	c1ccc2ccccc2c1	94,98
102	N1c2ccccc2Sc2ccccc12	97,33
103	CCc1ccccc1	97,43
104	Cc1cccc(Nc2ccccc2C(O)=O)c1C	98,12
105	CC(C)c1cccc(C(C)C)c1O	101,76
106	CCCc1ccccc1	106,43
107	c1ccc2c(c1)sc1ccccc21	108,26

108	c1ccc(cc1)N=Nc1ccccc1	110,64
109	c1ccc2cc3ccccc3cc2c1	113,79
110	CCCOc1ccccc1	114,38
111	c1cc2ccc3cccc4ccc(c1)c2c34	116,79
112	CN(C)CCC=C1c2ccccc2CCc2ccccc12	118,2
113	CN(C)CCCN1c2ccccc2Sc2ccc(Cl)cc12	121,84
114	CCCCCCc1ccccc1	128,41
115	OC(=O)c1ccccc1	49,7
116	Nc1ncnc2[nH]cnc12	8,65
117	CN1c2nc[nH]c2C(=O)N(C)C1=O	18,4
118	c1ccc(cc1)-c1nnn[nH]1	23,6
119	CN1c2ncn(C)c2C(=O)N(C)C1=O	25,08
120	c1nc2ccccc2[nH]1	34,3
121	Nc1ccccc1	43,22
122	COc1cc2CCC(NC(C)=O)C3=CC(=O)C(OC)=CC=C3c2c(OC)c1OC	43,9
123	CC(=O)Nc1ccccc1	41,19
124	CC12CC(O)C3C(CCC4=CC(=O)CCC34C)C1CCC2(O)C(=O)CO	49,56
125	Oc1ccccc1	47,52
126	Cc1ccc(N)cc1	56,13
127	CC1CC2C3CCC4=CC(=O)C=CC4(C)C3(F)C(O)CC2(C)C1(O)C(=O)CO	57,25
128	CCOC(=O)c1ccc(O)cc1	61,45
129	CC(=O)c1ccccc1	64,1
130	c1cc2ccccc2[nH]1	72,1
131	CCC(=O)c1ccccc1	77,4
132	CC(C)NCC(O)COc1cccc2ccccc12	77,53
133	COc1ccccc1	75,98
134	CCCCOC(=O)c1ccc(O)cc1	80,04
135	CC12CCC3C(CCC4=CC(=O)CCC34C)C1CCC2C(=O)CO	76,65
136	CC(=O)OCC(=O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	92,98
137	CC(=O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	96,34
138	CCCCC(=O)c1ccccc1	96,4
139	COC(=O)C(NC(=O)C(CC(N)=O)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	48,3
140	COC(=O)C(Cc1ccccc1)NC(C)=O	50,05
141	COC(=O)C(NC(=O)C(CO)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	51,57
142	COC(=O)C(NC(=O)C(Cc1ccc(O)cc1)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	59,57
143	COC(=O)C(NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	63,41
144	COC(=O)C(NC(=O)C(NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C)C(C)C	66,1
145	COC(=O)C(NC(=O)C(Cc1c[nH]c2ccccc12)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	68,42
146	COC(=O)C(Cc1ccccc1)NC(=O)C(NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	68,61
147	COC(=O)C(NC(=O)C(Cc1ccccc1)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	70,35
148	COC(=O)C(NC(=O)C(CC(C)C)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	70,46
149	COC(=O)C(CC(C)C)NC(=O)C(NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	70,72
150	COC(=O)C(NC(=O)C(COCc1ccccc1)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	73,65
151	COC(=O)C(NC(=O)C(CC(=O)OCc1ccccc1)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	74,14
152	COC(=O)C(NC(=O)C(Cc1ccc(OCc2ccccc2)cc1)NC(=O)C(C)NC(=O)OCc1ccccc1)C(C)C	80,51
153	CC(=O)Oc1ccccc1C(O)=O	21,62
154	CNS(=O)(=O)Cc1ccc2[nH]cc(CCN(C)C)c2c1	28,03
155	CNC(NCCSCc1nc[nH]c1C)=NC#N	30,81
156	CN1c2nc([nH]c2C(=O)N(C)C1=O)-c1ccccc1	53,94
157	[O-][N+](=O)c1ccccc1	59,79
158	COc1ccc(CN(CCN(C)C)c2ccccn2)cc1	67,87
159	COc1ccc(CCN2CCC(CC2)Nc2nc3ccccc3n2Cc2ccc(F)cc2)cc1	79,99
160	c1cc2ccccc2o1	81,66
161	c1ccncc1	34,16

162	CCCCC1C(=O)N(N(C1=O)c1ccccc1)c1ccccc1	46,17
163	CC(C)N=C1C=C2N(c3ccc(Cl)cc3)c3ccccc3N=C2C=C1Nc1ccc(Cl)cc1	124,88
164	NS(=O)(=O)c1cc2c(NCNS2(=O)=O)cc1C(F)(F)F	44,82
165	CCC1(C(=O)NCNC1=O)c1ccccc1	36,28
166	[c]1ccccc1	99,15
167	NC(=O)NN=Cc1ccc(o1)[N+]([O-])=O	37,18
168	CC12CC(=O)C3C(CCC4=CC(=O)C=CC34C)C1CCC2(O)C(=O)CO	53,82
169	Oc1ncnc2[nH]ncc12	-14,55
170	CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	84,4
171	Cc1cc(O)cc(C)c1Cl	79,9
172	NS(=O)(=O)c1cc2c(NC(CSCc3ccccc3)=NS2(=O)=O)cc1Cl	56,52
173	COc1cc(Cc2cnc(N)nc2N)cc(OC)c1OC	42,12
174	Cc1ccc(C)c(OCCCC(C)(C)C(O)=O)c1	65,51
175	CN(C)C1C2CC3C(=C(O)C2(O)C(=O)C(C(N)=O)=C1O)C(=O)c1c(O)cccc1C3(C)O	48,87
176	NS(=O)(=O)c1cc2c(NC(Cc3ccccc3)NS2(=O)=O)cc1C(F)(F)F	75,86
177	CCCCNc1ccc(cc1)C(=O)OCCN(C)C	84,4
178	CCCCOC(=O)c1ccc(N)cc1	80,8
179	NC(=O)NC(=O)Cc1ccccc1	43,92
180	CCCCC1C(=O)N(N(C1=O)c1ccc(O)cc1)c1ccccc1	34,48
181	OC1=C(Oc2cc(O)cc(O)c2C1=O)c1ccc(O)cc1O	30,43
182	CCCC(C)(COC(N)=O)COC(=O)NC(C)C	28,63
183	CC(=O)Nc1nnc(s1)S(N)(=O)=O	11,54
184	CCC1CN2CCC1CC2C(O)c1ccnc2ccc(OC)cc12	67,31
185	OC(=O)c1c(I)cc(I)c(NC(=O)CCCCC(=O)Nc2c(I)cc(I)c(C(O)=O)c2I)c1I	26,38
186	NC(=O)N1c2ccccc2C=Cc2ccccc12	61,46
187	Fc1ccc(cc1)C(=O)CCCN1CCC(=CC1)N1C(=O)Nc2ccccc12	76,75
188	CCOC(=O)c1ccc(N)cc1	63,26
189	CCCNC(=O)NS(=O)(=O)c1ccc(Cl)cc1	32,68
190	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC2OC(C)CC(C2O)N(C)C)C(C)(O)CC(C)C(=O)C(C)C(O)C1(C)O	66,53
191	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC2OC(C)CC(C2O)N(C)C)C(C)(CC(C)C(=O)C(C)C(O)C1(C)O)OC	81,08
192	CCCC(=O)OC1C(C)OC(CC1(C)OC(=O)CC)OC1C(C)OC(OC2C(CC=O)CC(C)C(O)C=CC=CCC(C)OC(=O)CC(O)C2OC)C(O)C1N(C)C	101,81
193	CCC1OC(=O)CC(O)C(C)C(OC2OC(C)C(OC3CC(C)(O)C(O)C(C)O3)C(C2O)N(C)C)C(CC=O)CC(C)C(=O)C=CC(C)=CC1COC1OC(C)C(O)C(OC)C1OC	71,38
194	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC2OC(C)CC(C2O)N(C)C)C(C)(O)CC(C)CN(C)C(C)C(O)C1(C)O	78,88
195	COC1CC(OC2C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)CC3(CO3)C(=O)C(C)C(O)C(C)C(C)OC(=O)C2C)OC(C)C1O	62,56

Appendix C

Supporting Information of "Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules" article

This appendix contains the supporting information for the "Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules" article published in Molecular Informatics, 2014, volume 33, pages 477-487. First the additional details of the modelling are given, followed by the data from the pK_{BHX} database.

Supporting information for: “Individual hydrogen-bond strength QSPR modelling with ISIDA local descriptors: a step towards polyfunctional molecules”

Fiorella Ruggiu[a], Vitaly Solov’ev[b], Gilles Marcou[a], Dragos Horvath[a], Jérôme Graton[c], Jean-Yves Le Questel[c], Alexandre Varnek*[a]

[a] Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France

[b] Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Leninskiy prospect, 31a, 119991, Moscow, Russian Federation

[c] Université de Nantes, UMR CNRS 6230, Chimie Et Interdisciplinarité: Synthèse, Analyse, Modélisation (CEISAM), UFR Sciences & Techniques, 2, rue de la Houssinière, BP 92208, 44322 NANTES Cedex 3, France

* Corresponding author, e-mail: *varnek@unistra.fr*

1. How to use the webserver to access the SVM CM

The SVM CM is freely available online on our web server: http://infochim.u-strasbg.fr/webserv/VSEngine.html. We invite all interested users to try our web server. Accounts are free and can be made readily. To make a prediction, one needs to select the “QSAR-based Property Predictions” in the menu (indicated with a ***1*** in Figure SI 1). Afterwards a menu appears to enter a project name and then, molecules may be drawn in ***2***or any format supported by ChemAxon may be uploaded in ***3***(see Figure SI 1).Information about available models can be found in the menu below (see 4 in Figure SI 1); by clicking on the link, an explanatory pdf will open. Once you have drawn or uploaded your molecules, click on “Validate”. The web server will then launch the standardisation of the compounds.

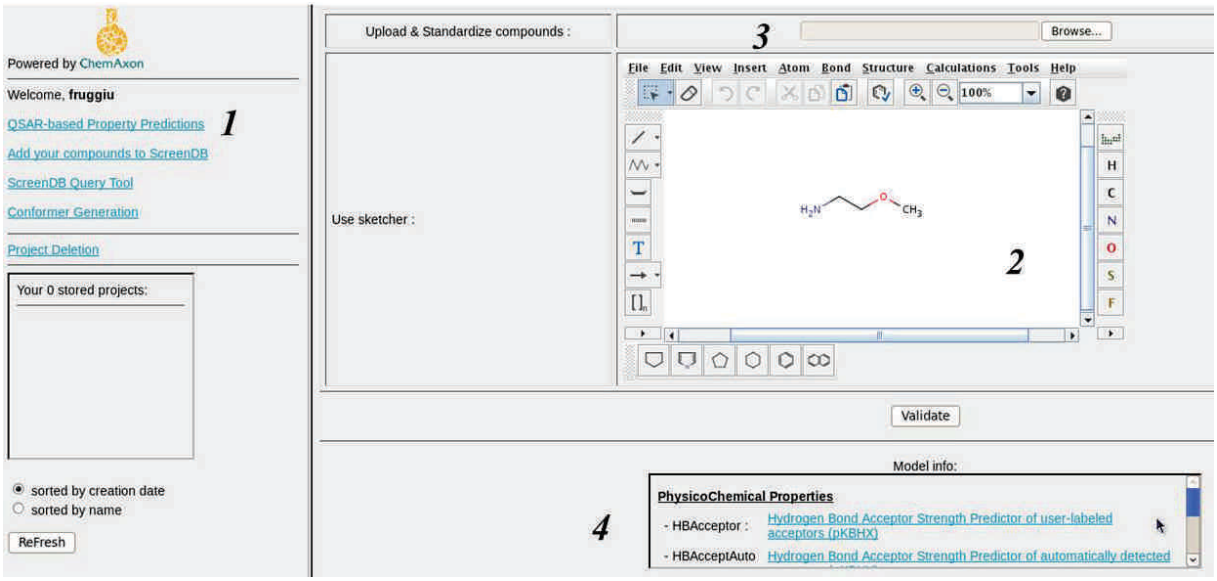


Figure SI 1. Input menu from the web server

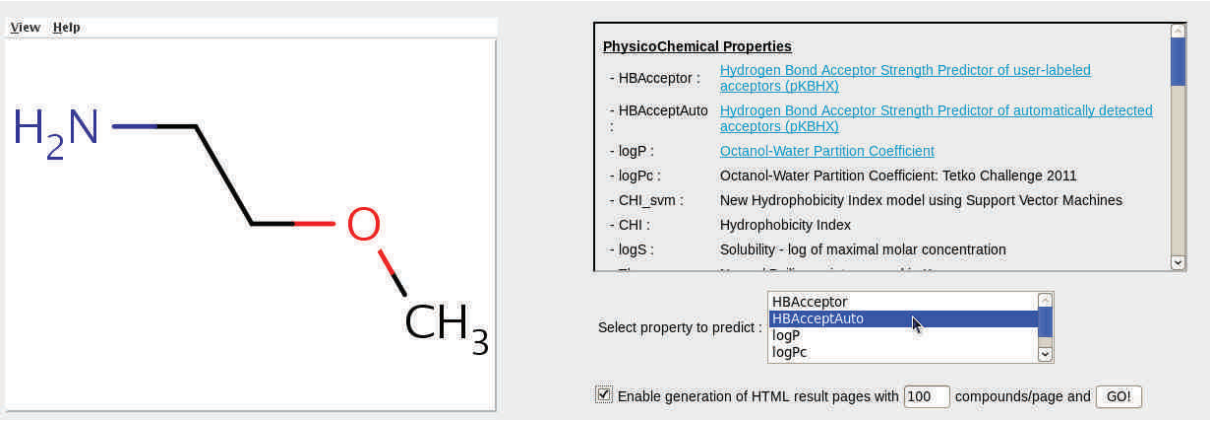


Figure SI 2. Prediction selection menu from the web server

After standardisation, the menu for the property prediction will appear (see Figure SI 2). The consensus model for pK_{BHX} described in this article is available as "HBAcceptorAuto" (see Figure SI 2). This option will automatically detect potential acceptor sites based on the ChemAxon property calculator. We recommend you use this one. After pressing "GO!", the webserver will apply the SVM models and output the most reliable prediction in the column "**Hbond_acceptor**"(see Figure SI 3). Results are available in html or csv format.

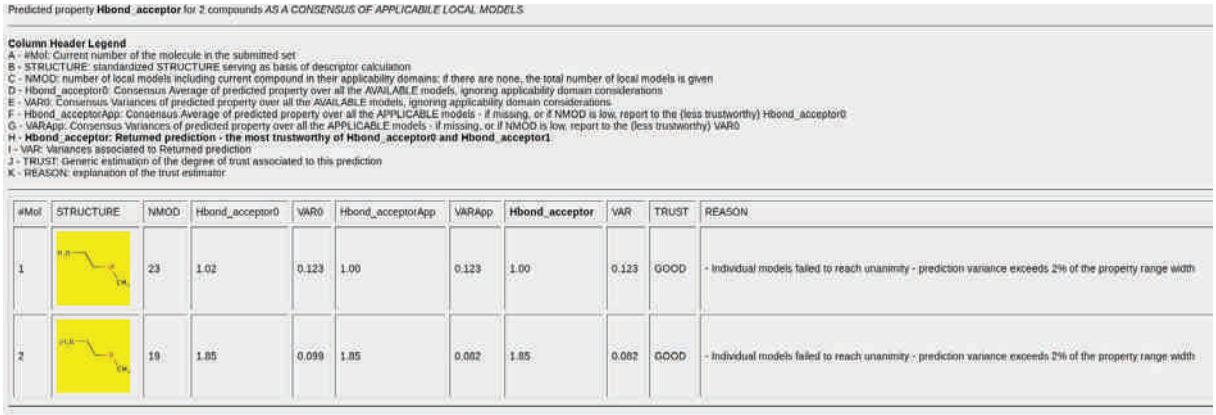


Figure SI 3. Results from the prediction web server

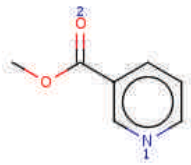
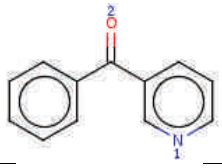
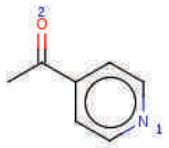
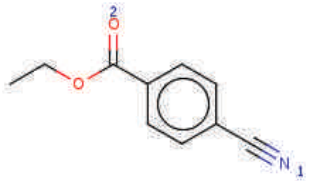
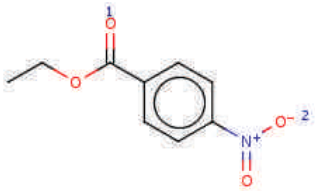
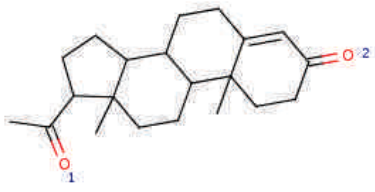
If you wish to indicate the acceptor site yourself, then you may use the mapping format and indicate your acceptor site as 1. Please generate one molecule per predicted acceptor sites; don't indicate two mapped 1 in the same entry. The property to predict then is HBAcceptor. The mapping will then be converted into the ISIDA Fragmentor2012 format for marked atom. More details are given in the pdf and ISIDA Fragmentor2012 manual.

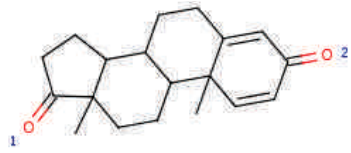
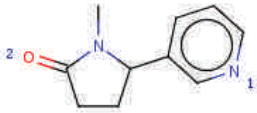
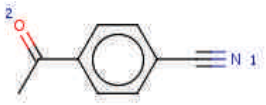
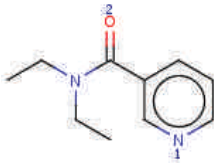
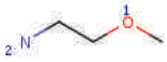
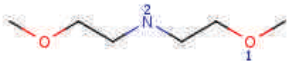
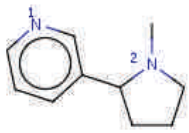
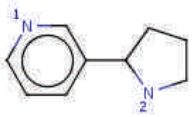
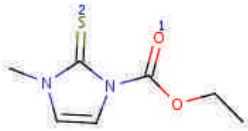
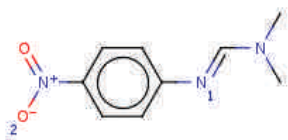
2. Bifunctional molecules in the test set

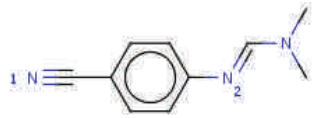
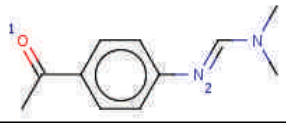
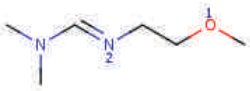
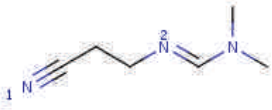
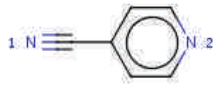
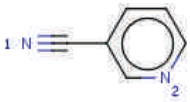
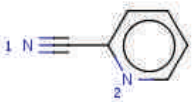
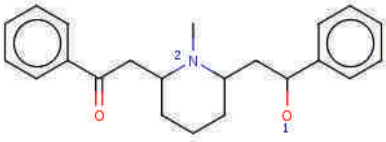
Table SI 1. Performances of the SVM CM on the 27 bifunctional acceptors in the test set

Minimum number of applicable models	Number of predictions	R^2_{det}	R^2_{corr}	RMSE	MAE
0	54	0.515	0.793	0.437	0.328
1	51	0.569	0.794	0.417	0.315
5	37	0.587	0.801	0.425	0.308
9	23	0.671	0.844	0.365	0.263
14	14	0.849	0.928	0.254	0.174

Table SI 2. Structure, experimental and predicted values of the polyfunctional molecules in the pK_{BHX} database external test set

Mol. nb.	Structure		Exp. pK_{BHX}	CM SVM			CM MLR	
				Nb. of models in AD	Predicted pK_{BHX} with AD	Predicted pK_{BHX} no AD	Nb. of models in AD	Predicted pK_{BHX} with AD
1		1	1,44	2	1.78	1.70	0	/
		2	0,51	8	0.72	0.74	0	/
2		1	1,42	13	1.42	1.46	0	/
		2	0,68	20	0.85	0.86	0	/
3		1	1,41	20	1.79	1.79	0	/
		2	0,78	8	0.97	1.04	0	/
4		1	0,66	2	0.97	0.86	0	/
		2	0,53	2	0.41	0.72	0	/
5		1	0,50	24	0.43	0.42	33	0,44
		2	0,16	2	0.48	0.30	0	/
6		1	1,2	2	1.41	1.32	0	/
		2	1,75	2	1.41	1.49	0	/

7		1	1,36	2	1.63	1.36	0	/
		2	1,84	2	1.63	2.00	0	/
8		1	1,62	2	1.73	2.45	0	/
		2	2,16	8	2.08	2.12	0	/
9		1	0,57	24	0.60	0.59	53	0,67
		2	0,15	24	0.14	0.13	30	0,00
10		1	0,65	25	0.70	0.69	26	0,53
		2	0,60	2	0.59	0.89	0	/
11		1	1,63	9	2.13	2.20	0	/
		2	1,98	8	2.01	2.07	0	/
12		1	1,09	23	1.00	1.02	0	/
		2	2,26	19	1.85	1.85	0	/
13		1	1,16	9	1.22	1.27	0	/
		2	2,22	13	2.22	2.09	0	/
14		1	0,70	13	1.42	1.35	0	/
		2	1,33	27	1.36	1.36	0	/
15		1	0,99	17	1.04	1.03	0	/
		2	2,31	19	1.63	1.63	0	/
16		1	2,03	13	2.63	2.58	0	/
		2	1,11	8	1.42	1.40	0	/
17		1	2,08	13	2.63	2.62	0	/
		2	1,52	8	1.64	1.70	0	/
18		1	0,72	0	/	1.83	0	/
		2	1,32	0	/	1.13	0	/
19		1	1,20	20	1.28	1.27	0	/
		2	0,48	7	1.45	1.25	0	/

20		1	1,23	8	1.99	1.82	0	/
		2	1,32	2	0.73	1.58	0	/
21		1	1,70	8	2.33	2.12	0	/
		2	1,52	24	1.67	1.66	30	1,74
22		1	1,00	8	1.32	1.42	0	/
		2	2,74	8	2.52	2.51	0	/
23		1	1,00	10	1.62	1.72	0	/
		2	2,12	20	1.89	1.97	0	/
24		1	0,47	8	0.70	0.74	0	/
		2	0,92	2	1.37	1.69	0	/
25		1	0,53	8	0.66	0.69	0	/
		2	0,82	2	1.37	1.59	0	/
26		1	0,61	0	/	0.66	0	/
		2	0,48	2	1.33	1.59	0	/
27		1	1,94	5	0.80	1.10	0	/
		2	2,11	10	1.53	1.51	0	/

3. External test set of 36 log*K*values of H-bond complexes with phenol

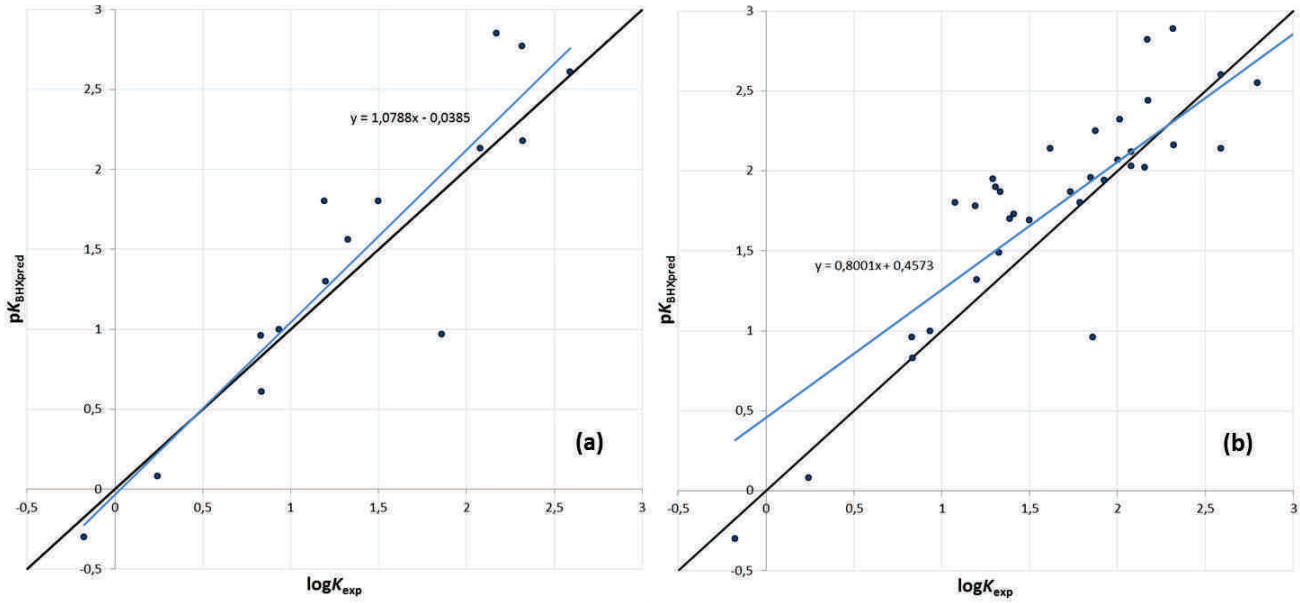
The values of the stability constant of the formation of a H-bond complex with phenol were taken from two publications[1, 2].

TableSI 3. Statistical parameters of both consensus models on the external test set of 36 log*K* values of H-bond complexes with phenol[a]

	CM SVM					log*K* vs pK_{BHX} by CM MLR				
	Nb. mol	RMSE	MAE	R^2_{det}	R^2_{corr}	Nb. mol	RMSE	MAE	R^2_{det}	R^2_{corr}
With AD	15	0.38	0.28	0.77	0.92	35	0.56	0.35	0.49	0.86
No AD	36	0.40	0.31	0.73	0.87	36	0.39	0.28	0.75	0.83

[a] For ISIDA/MLR, marked atom strategy type is MA1.

Figure SI 4. Predicted pK_{BHX} values vs experimental log*K* values with phenol for the external test set on 36 molecules (a) CM SVMwith AD and (b) CM SVM without AD. In light blue, the trend line is given.



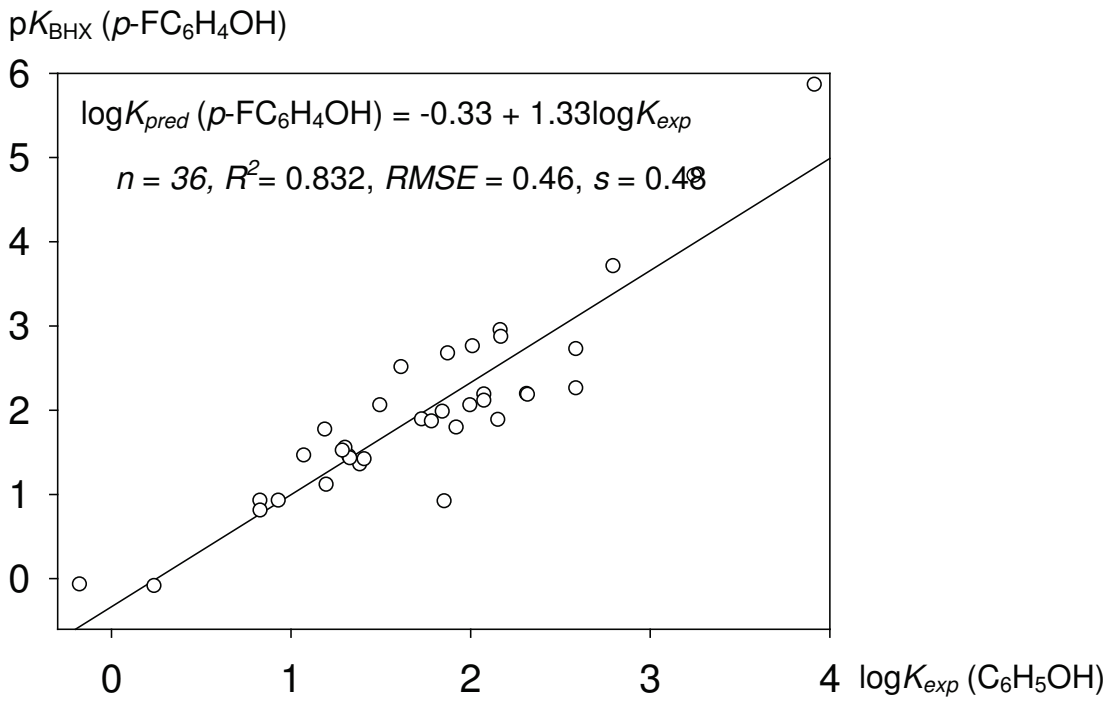
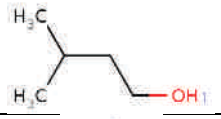
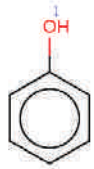
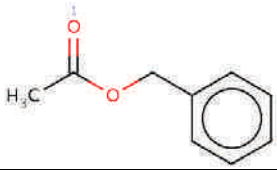
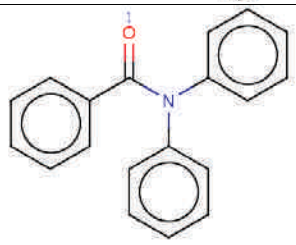
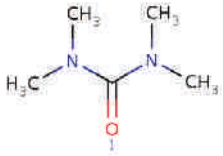
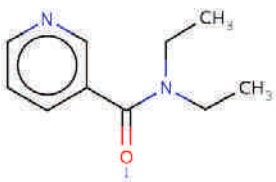
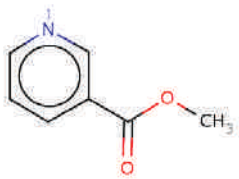
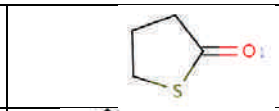
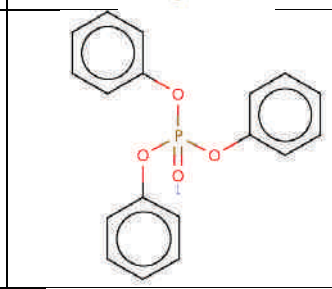
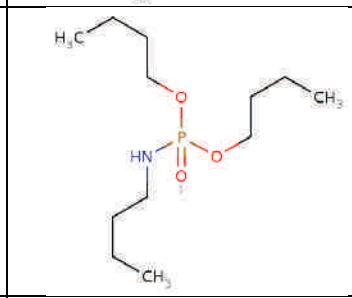
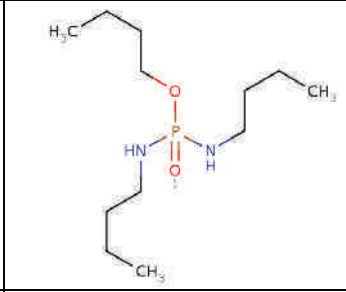
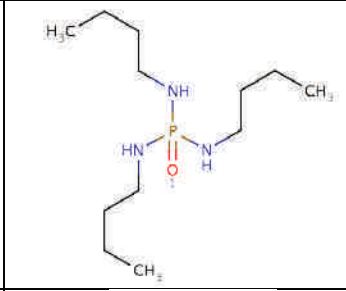
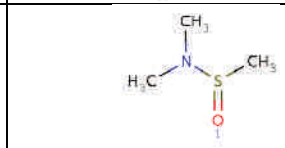
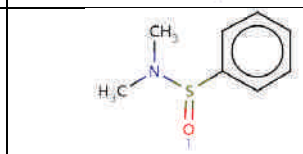
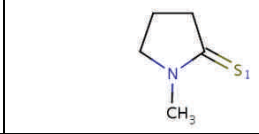
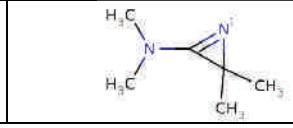
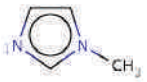
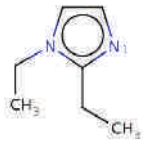
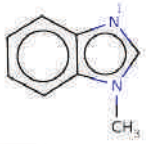
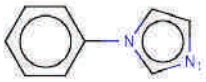
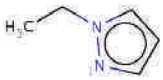
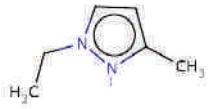
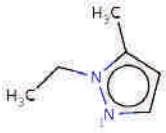
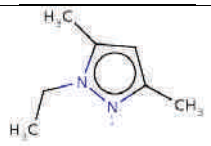
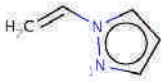
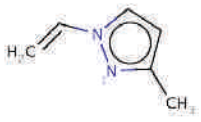
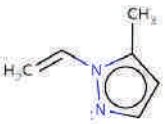
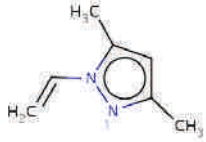


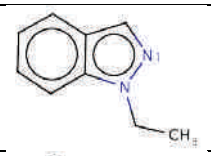
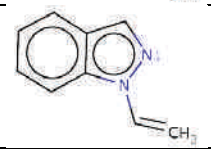
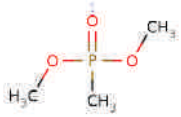
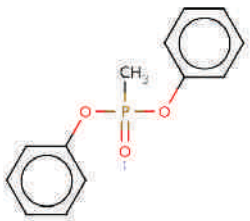
Figure SI 5. Predicted pK_{BHX} values by MLR CM vs experimental log*K* values with phenol for the external test set on 36 molecules

TableSI 4. Structure, experimental log*K* values and predictions by both consensus models of the external test set of 36 log*K* values of H-bond complexes with phenol[a]

	Structure	Exp. log*K*	pK_{BHX} by CM SVM			pK_{BHX} by CM MLR		
			Nb. of models in AD	Predicted log*K* with AD	Predicted log*K* no AD	Nb. of models in AD	Predicted log*K* with AD	Predicted log*K* no AD
1		0.93	27	1	1.00	74	0.92	0.92
2		0.83	27	0.96	0.96	87	0.92	0.92
3		0.24	27	0.08	0.08	84	-0.09	-0.09
4		-0.17	26	-0.30	-0.30	79	-0.08	-0.08
5		1.86	19	0.97	0.96	49	0.91	1.03
6		1.50	18	1.80	1.69	59	2.07	2.05
7		2.17	23	2.85	2.82	42	2.71	2.94
8		2.00	8	/	2.07	95	2.05	2.05
9		1.39	2	/	1.70	50	1.35	1.38

10		0.83	17	0.61	0.83	98	0.80	0.80
11		1.62	27	/	2.14	0	/	2.50
12		2.80	27	/	2.55	48	3.70	3.31
13		3.25	27	/	2.73	50	4.78	3.97
14		3.92	4	/	3.10	50	5.86	4.62
15		2.17	27	/	2.44	22	2.86	2.18
16		1.88	27	/	2.25	46	2.67	1.93
17		1.20	25	1.30	1.32	90	1.11	1.11
18		2.32	19	2.77	2.89	104	2.19	2.19

19		2.32	23	2.18	2.16	32	2.32	2.18
20		2.59	27	/	2.14	29	2.56	2.25
21		2.08	17	2.13	2.12	36	2.33	2.18
22		2.08	9	/	2.03	50	2.11	2.14
23		1.33	15	1.56	1.49	78	1.45	1.45
24		1.19	15	1.80	1.78	102	1.76	1.76
25		1.73	9	/	1.87	49	1.88	1.67
26		1.85	9	/	1.96	49	1.97	1.71
27		1.92	27	/	1.94	49	1.79	1.58
28		2.16	27	/	2.02	49	1.88	1.72
29		1.08	9	/	1.80	49	1.46	1.41
30		1.31	9	/	1.90	49	1.55	1.42
31		1.33	27	/	1.87	43	1.42	1.41
32		1.29	27	/	1.95	43	1.51	1.49

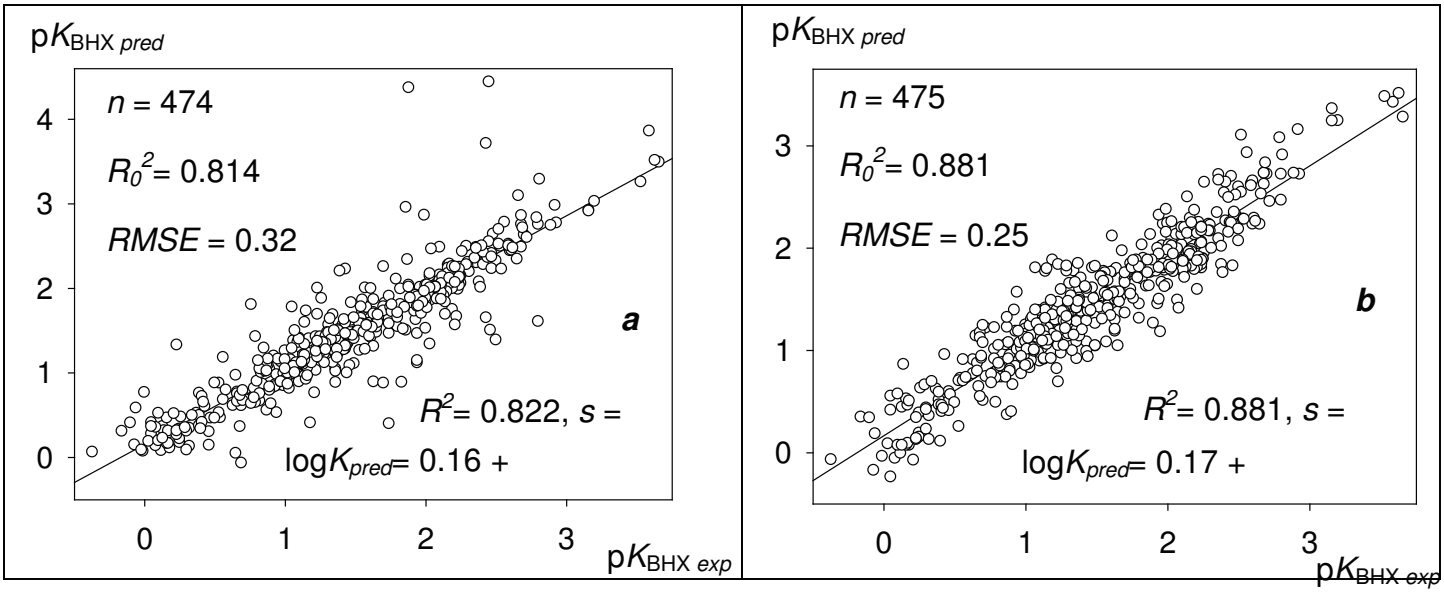
33		1.79	9	/	1.80	49	1.86	1.73
34		1.41	4	/	1.73	47	1.41	1.30
35		2.59	25	2.61	2.60	101	2.72	2.72
36		2.02	4	/	2.32	46	2.75	2.81

[a] For ISIDA/MLR, labeled fragment type is MA1.

4. Comparison of the different Marked Atom strategies

Table SI 5. Comparison of the 5CV-RMSE distribution for individual SVM models over the different iterations and folds using a Student t-test. The models significantly different at a confidence interval of 0.95 are in red. The ones in green indicate that it cannot be stated those are significantly different at the given confidence interval. The probability for the distributions to be equivalent is given.

	MA0	MA1	MA2	MA3
MA0		0,14	10^{-6}	10^{-6}
MA1	0,14		10^{-5}	10^{-5}
MA2	10^{-6}	10^{-5}		0,36
MA3	10^{-6}	10^{-5}	0,36	



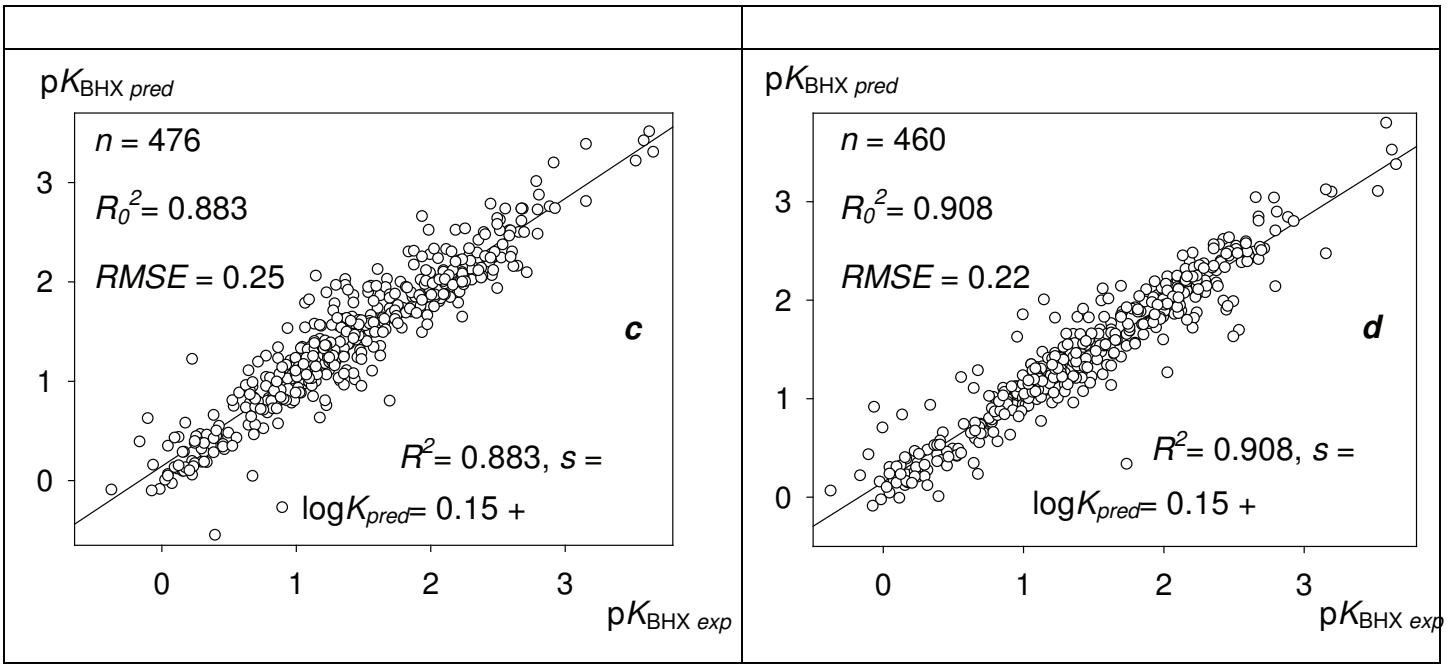


Figure SI6. MLR CM by the ISIDA/QSPR program. Predicted *vs.* experimental values of pK_{BHX} with ISIDA descriptors: (*a*) MA0 , (*b*) MA1, (*c*) MA2 and (*d*) MA3. R^2 is the determination coefficient and *RMSE* is the root-mean squared error. Predicted values correspond to all test sets of external 5-fold cross-validation.

5. SVM consensus model: Descriptors, SVM cost and statistical parameters of individual models

The ISIDA descriptors use a specified nomenclature where first the fragmentation type is given (I: sequences or II: augmented atom-centred fragments), then the property used (A: atom symbols), whether bonds are included (B), followed by the lengths of the fragments and finally the options used: P: atom pairs, AP: all path exploration, FC: use of formal charges and MA*X*: Marked Atom, the*X* stands for the strategy number. More details about the ISIDA descriptors and their nomenclature can be found on our website: http://infochim.u-strasbg.fr/spip.php?rubrique49.

The only varying option in SVM was the cost and it is given in Table 8. The following libsvm options were used: -s 3 -k 0 -p 0.09.

Table SI 6. Descriptors, SVM cost and statistical parameters of individual models in the SVM CM

Descriptors	SVM Cost	Average5CV-RMSE	Average5CV-R^2	Average 5CV-R^2_{corr}	Average 5CV-MAE
IAB2-5_-P-FC-MA3	0.7	0.27	0.87	0.94	0.18
IAB2-4_-AP-FC-MA3	0.5	0.27	0.87	0.93	0.16
IAB2-5_-FC-MA3	1	0.27	0.87	0.94	0.16
IAB2-5_-AP-FC-MA3	0.2	0.27	0.87	0.93	0.16
IIAB2-3_-R-P-FC-MA3	0.4	0.27	0.87	0.93	0.18
IAB2-4_-AP-MA3	0.3	0.27	0.87	0.93	0.17
IAB2-6_-P-FC-MA3	0.8	0.27	0.87	0.93	0.19
IAB2-4_-P-FC-MA3	0.7	0.27	0.87	0.93	0.19

IAB2-4_-FC-MA3	0.2	0.27	0.87	0.93	0.17
IAB2-4_-MA3	0.2	0.27	0.87	0.93	0.17
IAB3-5_-AP-FC-MA3	0.2	0.28	0.86	0.93	0.17
IAB2-5_-AP-MA3	0.2	0.28	0.86	0.93	0.17
IAB2-4_-P-MA3	0.5	0.28	0.86	0.93	0.19
IIAB2-3_-R-FC-MA3	0.2	0.28	0.86	0.93	0.17
IIAB2-3_-R-MA3	0.2	0.28	0.86	0.93	0.17
IAB2-5_-MA3	1	0.28	0.86	0.93	0.17
IAB2-7_-FC-MA3	0.4	0.28	0.86	0.93	0.18
IAB2-3_-P-FC-MA3	0.8	0.28	0.86	0.93	0.20
IIAB2-3_-R-P-MA3	0.4	0.28	0.86	0.93	0.18
IAB2-6_-AP-MA3	0.2	0.28	0.86	0.93	0.17
IAB2-5_-P-MA3	0.3	0.28	0.86	0.93	0.19
IAB2-7_-P-FC-MA3	0.4	0.28	0.86	0.93	0.20
IAB2-3_-FC-MA3	0.4	0.28	0.86	0.92	0.20
IAB2-8_-P-FC-MA3	0.4	0.29	0.85	0.93	0.20
IAB3-5_-AP-MA3	0.5	0.29	0.85	0.92	0.18
IAB2-6_-AP-FC-MA3	0.2	0.29	0.85	0.92	0.18
IAB2-6_-FC-MA3	0.4	0.29	0.85	0.92	0.18

References

1. Raevsky, O. A.; Solotnov, A. F.; Solov'ev, V. P., The Electron-Donor and -Acceptor Functions of the Physiologically Active and Model Compounds. IX. Thermodynamic Parameters of the Interaction of Methyldimethyl- and Methyldiphenyl-Phosphonate with Phenol. *Zhurnal Obshchei Khimii (Rus)* **1987,***57*, 1240-1243.

2. Raevsky, O. A.; Solov'ev, V. P.; Grigor'ev, V. Y., *Thermodynamic Characteristics of Hydrogen Bond of Phenols with Organic Bases*. VINITI Depos. N 1001-V88: Moscow, 1988; p 83.

#SMILES	pK_BHX	Comment
CC(=[O:1])c1cccnc1	0,90	polyfunctional
C1COCC[NH:1]1	1,78	polyfunctional
C1CSCC[NH:1]1	1,67	polyfunctional
C1CSC[NH:1]1	1,10	polyfunctional
C[N:1]1CCOCC1	1,56	polyfunctional
CC(=O)c1ccc[n:1]c1	1,39	polyfunctional
C1C[O:1]CCN1	1,10	polyfunctional
CN1CC[O:1]CC1	0,96	polyfunctional
C1C[S:1]CCN1	0,34	polyfunctional
C1C[S:1]CN1	0,32	polyfunctional
CCOC(=[O:1])C#C	0,63	
CCOC=[O:1]	0,66	
CCOC(=[O:1])CCl	0,67	
CCOC(=[O:1])CF	0,74	
COC(C)=[O:1]	1,00	
CCOC(=[O:1])C(C)(C)C	1,04	
CCOC(=[O:1])Cc1ccccc1	1,05	
CCOC(=[O:1])C12CC3CC(	1,06	
CCOC(C)=[O:1]	1,07	
CCOC(=[O:1])CC	1,08	
CCOC(=[O:1])C(C)C	1,09	
CCOC(=[O:1])C1CC1	1,12	
CCOC(=[O:1])C=Cc1ccccc	1,14	
CC(C)OC(C)=[O:1]	1,15	
CCOC(=[O:1])c1ccc(Br)cc1	0,78	
COC(=[O:1])c1ccccc1	0,89	
CCOC(=[O:1])c1ccccc1	0,94	
CCOC(=[O:1])c1ccc(C)cc1	1,05	
CCOC(=[O:1])c1ccc(OC)cc	1,13	
CCOC(=[O:1])c1ccc(cc1)N(	1,45	
COC(=[O:1])OC	0,82	
CCOC(=[O:1])OC	0,84	
CCOC(=[O:1])OCC	0,88	
CC1COC(=[O:1])O1	1,22	
[O:1]=C1CCO1	0,86	
CC1CC(=[O:1])O1	0,97	
[O:1]=C1Oc2ccccc2C=C1	1,30	
[O:1]=C1CCCO1	1,32	
CC1CCC(=[O:1])O1	1,43	
CC=[O:1]	0,65	
[O:1]=Cc1ccccc1	0,78	
Clc1ccc(C=[O:1])cc1	0,63	
COc1ccc(C=[O:1])cc1	1,10	
CN(C)c1ccc(C=[O:1])cc1	1,53	
[O:1]=CC=Cc1ccccc1	1,13	
COc1ccccc1C=[O:1]	1,11	
CC(C)=[O:1]	1,18	
CCC(C)=[O:1]	1,22	
CC(C)C(=[O:1])C(C)C	1,08	
CC(C)(C)C(=[O:1])C(C)(C)(	0,96	
CC(=[O:1])C12CC3CC(CC(	1,30	
[O:1]=C(C1CC1)C1CC1	1,36	
CC(C)(C)C(=[O:1])C12CC3	1,08	
[O:1]=C(C12CC3CC(CC(C3	1,17	
[O:1]=C1CCC1	1,00	
[O:1]=C1CCCC1	1,27	

[O:1]=C1CCCCC1	1,39
[O:1]=C1CCCCCC1	1,41
[O:1]=C1CCCCCCC1	1,45
[O:1]=C1CCCCCCCCCC1	1,20
[O:1]=C1CCCCCCCCCCC	1,23
[O:1]=C1CCCCCCCCCCC(	1,22
CC1(C)C2CCC1(C)C(=[O:1	1,31
CC(=[O:1])C(Cl)(Cl)Cl	0,00
CC(=[O:1])C(Cl)Cl	0,25
ClCC(=[O:1])CCl	0,32
CC(=[O:1])CCl	0,66
CC1=CC(=[O:1])C=C(C)O1	2,50
[O:1]=C1c2ccccc2Oc2cccc	1,36
[O:1]=C1c2ccccc2Cc2cccc	1,25
CCN(CC)c1ccc(cc1)C(C)=[	1,82
CN(C)c1ccc(cc1)C(C)=[O:1	1,76
CC(=[O:1])c1ccc(cc1)N1CC	1,71
COc1ccc(cc1)C(C)=[O:1]	1,33
CC(=[O:1])c1ccc(cc1)C12C	1,30
CC(=[O:1])c1ccc(cc1)C(C)(	1,25
CC(C)c1ccc(cc1)C(C)=[O:1	1,21
CCc1ccc(cc1)C(C)=[O:1]	1,25
CC(=[O:1])c1ccc(C)cc1	1,24
CSc1ccc(cc1)C(C)=[O:1]	1,21
CC(=[O:1])c1ccccc1	1,11
CC(=[O:1])c1ccc(F)cc1	1,00
CC(=[O:1])c1ccc(Cl)cc1	0,93
COc1cccc(c1)C(C)=[O:1]	1,16
CC(=[O:1])c1cccc(C)c1	1,10
CC(=[O:1])c1cccc(F)c1	0,83
CC(=[O:1])c1cccc(Cl)c1	0,82
CC(=[O:1])c1cccc(c1)C(F)(	0,72
CC(=[O:1])c1ccccc1Cl	0,90
COc1ccccc1C(C)=[O:1]	1,34
[O:1]=C(c1ccccc1)c1ccccc	1,07
COc1ccc(cc1)C(=[O:1])c1c	1,27
COc1ccc(cc1)C(=[O:1])c1c	1,49
CN(C)c1ccc(cc1)C(=[O:1])c	1,67
CC(=[O:1])C#C	0,68
[O:1]=C1c2ccccc2-c2ccccc	1,09
CC(=[O:1])c1ccc2ccccc2c1	1,13
CC(=[O:1])C=Cc1ccccc1	1,38
[O:1]=C1C=CC=CC=C1	1,97
CNC=[O:1]	1,96
CNC(=[O:1])c1ccccc1	2,03
CCC(=[O:1])NC	2,24
CCNC(C)=[O:1]	2,29
CNC(C)=[O:1]	2,30
CN(C)C(Cl)=[O:1]	1,00
CN(C)C(=[O:1])C(F)(F)F	1,04
CN(C)C(=[O:1])CCl	1,74
CN(C=[O:1])c1ccccc1	1,74
CC(=[O:1])N(c1ccccc1)c1c	1,94
CC(C)(C)C(=[O:1])N(C1CC	2,06
CN(C)C(=[O:1])C(C)(C)C	2,10
CN(C)C=[O:1]	2,10
CN(C(C)=[O:1])c1ccccc1	2,19

CCC(=[O:1])N(C1CCCCC1	2,22
CN(C)C(=[O:1])c1ccccc1	2,23
CC(C)C(=[O:1])N(C1CCCC	2,24
CC(C)C(=[O:1])N(C)C	2,26
CCC(=[O:1])N(C)C	2,36
CC(=[O:1])N(C1CCCCC1)C	2,41
CN(C)C(C)=[O:1]	2,44
CCN(CC)C(C)=[O:1]	2,47
CN1CCCC1=[O:1]	2,38
CN1CCCCCC1=[O:1]	2,53
CN1C=CC=CC1=[O:1]	2,50
CN1CCCCC1=[O:1]	2,60
CN(C)C(=[O:1])Oc1ccccc1	1,70
CCOC(=[O:1])N(C)C	1,83
CCN(CC)C=[O:1]	2,16
COc1ccccc1C(=[O:1])N(C)C	2,48
COC(=[O:1])c1ccccc1OC	1,49
COC(=[O:1])c1ccc(OC)cc1	1,08
CN(C)C=NC(=[O:1])c1cccc	2,10
CCN(CC)C(=[O:1])C(F)(F)F	1,06
CCN(CC)C(=[O:1])SC	1,56
CN(C)C(=[O:1])C12CC3CC	2,30
[O:1]=CN1CCCC1	2,31
[O:1]=CN1CCCCC1	2,17
CCCCN(CCCC)C=[O:1]	2,17
CC(C)N(C=[O:1])C(C)C	2,24
CC(=[O:1])N1CCCC1	2,61
CC(C)N(C(C)C)C(C)=[O:1]	2,47
CC(=[O:1])N1CCCCC1	2,45
CCCCC(=[O:1])CCCC	1,21
CN(C)C=NC(=[O:1])N(C)C	2,92
CC12CCC3C(CCC4CCCC	1,36
CCC(=[O:1])c1ccccc1	1,04
CCCC(=[O:1])c1ccccc1	1,04
CN1N(C(=[O:1])C=C1C)c1c	2,80
CN(C)C1=C(C)N(C)N(C1=[	2,45
CC(=[O:1])N1CCOCC1	2,16
[O:1]=CN1CCOCC1	1,93
CC(=[O:1])c1ccc(cc1)N1CC	1,61
[NH2:1]c1cccc(Cl)c1	0,13
[NH2:1]c1cccc(F)c1	0,20
CC(C)c1cccc(C(C)C)c1[NH	0,37
CCc1cccc(CC)c1[NH2:1]	0,39
[NH2:1]c1ccccc1	0,46
Cc1cccc(C)c1[NH2:1]	0,47
Cc1ccc([NH2:1])cc1	0,56
C[NH:1]c1ccc(Cl)cc1	0,05
C[NH:1]c1ccccc1	0,26
C[NH:1]c1ccc(C)cc1	0,43
C1C[NH:1]c2ccccc2C1	0,70
C1Cc2ccccc2[NH:1]1	0,79
C[N:1](C)c1ccc(Br)cc1	0,17
CC[N:1](CC)c1ccccc1	0,05
C1CC[N:1](C1)c1ccccc1	0,16
C[N:1](C)c1ccccc1	0,39
C[N:1](C)c1cccc(C)c1	0,41
C1C[N:1]2CCCc3cccc(C1)c	0,39

C1CC[N:1](CC1)c1ccccc1	0,68
C[N:1](C)c1ccc(C)cc1	0,69
[NH3:1]	1,74
C1CC[NH:1]CC1	2,38
CC[NH:1]C	2,25
CC[NH:1]CC	2,25
C[NH:1]C1CCCCC1	2,24
CCCC[NH:1]C	2,24
C1CCC[NH:1]CC1	2,24
C[NH:1]C(C)(C)C	2,21
C[NH:1]C(C)C	2,20
CCCC[NH:1]CCCC	2,11
CC(C)[NH:1]C(C)C	2,00
C[NH:1]CC=C	2,00
CC1(C)CCCC(C)(C)[NH:1]	1,88
C=CC[NH:1]CC=C	1,70
C[NH:1]CC#C	1,64
C([NH:1]Cc1ccccc1)c1cccc	1,34
ClCC[NH:1]CCCl	1,19
C1C[NH:1]C(C1)c1ccccc1	1,93
Fc1cccc(c1)C1CCC[NH:1]1	1,65
FC(F)(F)c1cccc(c1)C1CCC	1,38
C1CC2(CC[N:1]1CC2)c1cc	2,46
C[N:1]1C2CCC1CCC2	2,39
C[N:1]1CCCC1	2,19
CC[N:1](C)C	2,17
C[N:1](C)C1CCCCC1	2,15
C[N:1](C)C	2,13
CC(C)[N:1](C)C	2,11
C[N:1]1CCCCC1	2,11
CCCC[N:1]1CCCC1	2,04
CC[N:1](CC)CC	1,98
ClC1C[N:1]2CCC1CC2	1,97
C[N:1](C)CC=C	1,92
C[N:1]1CCCC(Cl)C1	1,74
C[N:1]1CCC(Cl)CC1	1,70
C[N:1](C)CC#C	1,60
C[N:1](C)Cc1ccccc1	1,59
CCCCCCCC[N:1](CCCCC(	1,57
CCCC[N:1](CCCC)CCCC	1,55
C[N:1](C)CCCCl	1,54
CCC[N:1](CCC)CCC	1,47
ClCC[N:1]1CCCC1	1,45
C=CC[N:1](CC=C)CC=C	1,34
C[N:1]1C(C)(C)CCCC1(C)C	1,23
CC[N:1](C1CCCCC1)C1CC	1,14
C[N:1]1CCCC1c1ccccc1	1,38
C[N:1]1CCCC1c1cccc(F)c1	1,09
C[N:1]1CCCC1c1cccc(c1)C	0,92
[NH2:1]Cc1cccc(c1)[N+]([O	1,26
[NH2:1]Cc1cc(Cl)cc(Cl)c1	1,27
[NH2:1]Cc1cc(F)cc(F)c1	1,28
[NH2:1]Cc1cccc(c1)C(F)(F)	1,43
COc1cccc(C[NH2:1])c1	1,94
Cc1cccc(C[NH2:1])c1	1,97
[NH2:1]Cc1cccc(F)c1	1,58
[NH2:1]Cc1cccc(Cl)c1	1,55

C1CC=CC[NH:1]1	2,16
C[N:1]1CCC=CC1	2,02
CC[N:1](C)Cc1ccccc1	1,47
C[N:1](C)Cc1cccc(F)c1	1,27
C[N:1](C)Cc1cccc(c1)C(F)(	1,16
CC[N:1](C)Cc1cccc(c1)C(F	1,03
CC[N:1](C)Cc1cccc(F)c1	1,14
C[N:1](C)CCCn1c2ccccc2c	1,88
C[N:1](C)CCC(=O)c1ccccc	1,94
C[N:1](C)n1cccc1	0,23
Brc1ccc(NC(=[N:1]c2ccc(Br	0,87
Cc1ccc(NC(=[N:1]c2ccccc2	1,47
Cc1cccc(NC(=[N:1]c2ccccc	1,40
Fc1ccc(cc1)[N:1]=C(Nc1cc(	1,25
Clc1ccc(cc1)[N:1]=C(Nc1cc	1,13
Clc1cccc(c1)[N:1]=C(Nc1cc	1,05
CN(C)C=[N:1]c1ccc(Br)cc1	1,65
CN(C)C=[N:1]c1ccccc1	1,90
CN(C)C=[N:1]c1ccc(C)cc1	2,07
CN(C)C=[N:1]c1ccccc1Br	1,37
CN(C)C=[N:1]c1ccccc1C	1,63
CN(C)C=[N:1]Cc1cc(Cl)cc(	2,00
CN(C)C=[N:1]Cc1cccc(Cl)c	2,10
CN(C)C=[N:1]Cc1ccc(Cl)cc	2,12
CN(C)C=[N:1]Cc1ccccc1	2,35
CN(C)C=[N:1]Cc1ccc(C)cc1	2,36
CCC[N:1]=CN(C)C	2,59
CC(C)[N:1]=CN(C)C	2,60
CC(C)C[N:1]=CN(C)C	2,52
CN(C)C=[N:1]C(C)(C)C	2,41
C[N:1]=CN(C)C	2,70
CN(C)C=[N:1]C1CCCCC1	2,59
CCC(C)(C)[N:1]=CN(C)C	2,26
CN(C)C=[N:1]CC=C	2,48
CN(C)C=[N:1]CC(F)(F)F	1,60
COc1cccc(C[N:1]=CN(C)C)	2,40
COc1ccc(cc1)[N:1]=CN(C)C	2,08
CN(C)C=[N:1]c1ccc(F)cc1	1,80
COc1ccc(cc1)[N:1]=C(N(C)	1,53
CN(C(=[N:1]c1ccc(C)cc1)c1	1,52
CN(C(=[N:1]c1ccc(Br)cc1)c	1,11
COc1ccc(cc1)N(C)C(=[N:1]	1,63
COc1ccc(cc1)N(C)C(=[N:1]	1,23
CN(C(=[N:1]c1ccc(C)cc1)c1	1,42
CN(C(=[N:1]c1ccc(Br)cc1)c	0,99
C[N:1]=C(N(C)C)c1ccc(Br)c	2,19
CN(C(C)=[N:1]c1ccccc1)c1	1,65
CN(C(C)=[N:1]c1ccc(C)cc1	1,75
CN(C(C)=[N:1]c1ccc(Br)cc1	1,36
CN(C(C)=[N:1]c1cccc(Cl)c1	1,24
CN(C(C)=[N:1]c1ccccc1)c1	1,76
CN(C(C)=[N:1]c1ccc(C)cc1	1,82
CN(C(C)=[N:1]c1ccc(Br)cc1	1,43
CN(C(C)=[N:1]c1ccccc1)c1	1,38
CN(C(C)=[N:1]c1ccc(Br)cc1	1,19
CN(C(C)=[N:1]c1ccccc1)c1	1,44
CN(C(C)=[N:1]c1cccc(Cl)c1	1,05

CN(C)C(=[NH:1])N(C)C	3,20
Cc1ccc(NC=[N:1]c2ccc(C)c	2,22
Clc1ccc(NC=[N:1]c2ccc(Cl)	1,63
CC(Nc1ccccc1)=[N:1]c1ccc	1,66
CC(Nc1ccc(C)cc1)=[N:1]c1	2,01
CC(Nc1ccc(C)cc1)=[N:1]c1	1,85
CC(Nc1ccc(C)cc1)=[N:1]c1	1,58
CC(Nc1ccccc1)=[N:1]c1ccc	1,36
CC(Nc1ccccc1)=[N:1]c1ccc	1,24
N(C(=[N:1]c1ccccc1)c1cccc	1,29
C1CCN(C1)c1cc[n:1]cc1	2,93
CCN(CC)c1cc[n:1]cc1	2,89
CN(C)c1cc[n:1]cc1	2,80
CC1CCN(CC1)c1cc[n:1]cc1	2,68
C1CCN(CC1)c1cc[n:1]cc1	2,68
CN(N)c1cc[n:1]cc1	2,58
Nc1cc[n:1]cc1	2,56
Cc1cc(C)[n:1]c(C)c1	2,29
Cc1cc[n:1]cc1C	2,24
Nc1ccc[n:1]c1	2,20
Cc1cccc(C)[n:1]1	2,14
COc1cc[n:1]cc1	2,13
Nc1cccc[n:1]1	2,12
CC(C)(C)c1cc[n:1]cc1	2,11
CNc1cccc[n:1]1	2,11
Cc1cc[n:1]cc1	2,07
CCc1cc[n:1]cc1	2,07
Cc1cccc[n:1]1	2,03
CCc1ccc[n:1]c1	2,01
Cc1ccc[n:1]c1	2,00
c1ccc(cc1)-c1cc[n:1]cc1	1,96
C=Cc1cc[n:1]cc1	1,95
CCc1cccc[n:1]1	1,94
c1ccc2[n:1]cccc2c1	1,89
CCCCc1cccc[n:1]1	1,88
c1ccc2c(c1)c[n:1]c1ccccc2	1,87
c1cc[n:1]cc1	1,86
C=Cc1cccc[n:1]1	1,65
CN(C)c1cccc[n:1]1	1,61
Clc1cc[n:1]cc1	1,54
c1ccc(cc1)-c1cccc[n:1]1	1,43
Ic1ccc[n:1]c1	1,37
Fc1ccc[n:1]c1	1,35
Clc1ccc[n:1]c1	1,31
Brc1ccc[n:1]c1	1,31
Clc1cccc[n:1]1	1,05
Fc1cccc[n:1]1	0,95
Brc1cccc[n:1]1	1,03
Clc1c[n:1]cc(Cl)c1	0,85
Fc1cccc(F)[n:1]1	0,14
Cc1cc[n:1]c(C)c1	2,21
CNc1cc[n:1]cc1	2,69
BrC(Br)C#[N:1]	0,19
BrC#[N:1]	0,19
ClCC#[N:1]	0,39
FC(F)(F)c1ccc(cc1)C#[N:1]	0,54
FC(F)(F)c1cccc(c1)C#[N:1]	0,53

Fc1ccccc1C#[N:1]	0,64
BrCc1ccccc1C#[N:1]	0,69
Brc1ccccc1C#[N:1]	0,71
Clc1ccccc1C#[N:1]	0,67
CSC#[N:1]	0,73
Fc1ccc(cc1)C#[N:1]	0,72
[N:1]#COc1ccccc1	0,77
[N:1]#Cc1ccccc1	0,80
Cc1ccccc1C#[N:1]	0,83
CC#[N:1]	0,91
CCC#[N:1]	0,93
CCCC#[N:1]	0,89
CC(C)C#[N:1]	1,00
COc1ccccc1C#[N:1]	1,06
CCCCCC#[N:1]	0,89
CC(C)(C)C#[N:1]	0,99
COc1ccc(cc1)C#[N:1]	0,97
[N:1]#CC1CC1	1,03
[N:1]#CC12CC3CC(CC(C3	1,00
CN(C)C#[N:1]	1,56
[N:1]#CN1CCCCC1	1,58
CCN(CC)C#[N:1]	1,63
Clc1ccc(cc1)C#[N:1]	0,66
Cc1cccc(C)c1C#[N:1]	0,86
C=CCC#[N:1]	0,87
[N:1]#CN1CCOCC1	1,34
[N:1]#CN1CCCC1	1,66
CC(C)N(C#[N:1])C(C)C	1,74
CCCN=C(NC)NC#[N:1]	2,09
CN(C)C(C)=NC#[N:1]	2,24
CNCC#[N:1]	0,79
CN(C)CC#[N:1]	0,76
C[O:1]C(C)(C)C	1,19
CC(C)[O:1]C(C)C	1,11
CC[O:1]C(C)(C)C	1,08
CC[O:1]CC	1,01
CCCC[O:1]CCCC	0,88
C=CC[O:1]CC=C	0,70
CC1(C)CCC(C)(C)[O:1]1	1,43
CC12CCC(CC1)C(C)(C)[O:	1,38
C1C[O:1]C1	1,36
CC1CCC[O:1]1	1,34
C1CC[O:1]C1	1,28
C1CC[O:1]CC1	1,23
C1CCC2[O:1]C2C1	1,13
CC1C[O:1]1	0,97
C1CC=C[O:1]1	0,53
C1C[O:1]C=CC1	0,41
CC(C)(C)[OH:1]	1,14
CC(C)[OH:1]	1,06
CC[OH:1]	1,02
C[OH:1]	0,82
[OH:1]CCCl	0,50
[OH:1]c1ccc(F)cc1	-0,12
[OH2:1]	0,65
C[O:1]c1ccccc1	-0,07
C[O:1]c1ccc(Cl)cc1	-0,25

CC[O:1]c1ccccc1	-0,01
C[O:1]c1ccccc1C	-0,37
C[O:1]c1ccc(C)cc1	0,08
C1Cc2ccccc2[O:1]1	0,21
CC[SH:1]	-0,16
CC(C)[SH:1]	-0,10
C[S:1]C	0,12
CC[S:1]C	0,18
CCC[S:1]C	0,16
CCCC[S:1]C	0,17
CCCCCCCC[S:1]C	0,17
C[S:1]C1CCCCC1	0,24
C[S:1]C(C)(C)C	0,25
CC[S:1]CC	0,25
CCCC[S:1]CCCC	0,23
CC(C)[S:1]C(C)C	0,30
CC(C)(C)[S:1]C(C)(C)C	0,40
C[S:1]CC=C	0,08
C[S:1]Cc1ccccc1	-0,02
C[S:1]CCCl	-0,12
CC[S:1]C=C	-0,13
C1C[S:1]1	0,03
CC1C[S:1]1	0,10
C1C[S:1]C1	0,13
C1CC[S:1]C1	0,32
C1CC[S:1]CC1	0,23
CN(C)C(Cl)=[S:1]	0,50
CN(C)C(=[S:1])c1ccccc1	1,02
CN(C)C=[S:1]	1,05
CNC(C)=[S:1]	1,14
CN(C)C(=[S:1])c1ccc(C)cc1	1,15
COc1ccc(cc1)C(=[S:1])N(C	1,16
CN(C)C(C)=[S:1]	1,22
CN(C)C(=[S:1])c1ccc(N)cc1	1,33
[S:1]=C1CCCCCN1	1,60
CN(C)C=NC(=[S:1])N(C)C	1,79
CN1CCN(C)C1=[S:1]	1,32
CN(C)C(=[S:1])N(C)C	1,35
CN1CCCNC1=[S:1]	2,00
CN(C)C=NC(=[S:1])c1cccc	1,23
CSC(=[S:1])N=CN(C)C	1,12
CN(C)C=NC(=[S:1])SCc1cc	1,10
CN(C)C(C)=NC(=[S:1])SCc	1,49
CN=C=[S:1]	-0,05
CCP(=[O:1])(CC)CC	3,66
CCCCP(=[O:1])(CCCC)CC	3,63
CN(C)P(=[O:1])(N(C)C)N(C	3,60
CCCCCCCCP(=[O:1])(CCC	3,59
CP(C)(C)=[O:1]	3,53
[O:1]=P(c1ccccc1)(c1ccccc	3,16
CCOP(C)(=[O:1])OCC	2,81
CCOP(=[O:1])(OCC)OCC	2,68
CCCCOP(=[O:1])(OCCCC)	2,66
COP(=[O:1])(OC)OC	2,50
ClP(=[O:1])(c1ccccc1)c1cc	2,17
ClP(Cl)(=[O:1])c1ccccc1	1,26
ClP(Cl)(Cl)=[O:1]	0,56

223

[O-:1][N+](=O)C=CN1CCC(	1,62
CCN(CC)C=C[N+]([O-:1])=(	1,58
[O-:1][N+](=O)C=CN1CCC	1,53
CN(C)[N+]([O-:1])=O	0,82
COc1ccc(cc1)[N+]([O-:1])=(	0,50
Cc1ccc(cc1C)[N+]([O-:1])=(	0,46
[O-:1][N+](=O)c1ccccc1	0,30
[O-:1][N+](=O)c1ccc(cc1)C	0,42
Cc1cccc(C)c1[N+]([O-:1])=(	0,29
CC(=O)c1cccc(c1)[N+]([O-:	0,18
c1csc[n:1]1	1,37
C=Cn1cc[n:1]c1	2,35
c1ccc(cc1)-c1[n:1]c2ccccc2	1,18
c1[n:1]c2ccccc2s1	1,29
[O:1]=S1OCCO1	0,87
COS(=[O:1])OC	0,94
CCOS(=[O:1])OCC	1,07
[O:1]=S(c1ccccc1)c1ccccc1	2,04
CS(=[O:1])c1ccccc1	2,24
Cc1ccc(cc1)S(=[O:1])c1ccc	2,21
[O:1]=S(Cc1ccccc1)Cc1ccc	2,43
CS(C)=[O:1]	2,54
[O:1]=S1CCCC1	2,47
CCCCS(=[O:1])CCCC	2,65
FC(F)(F)c1cccc(c1)C1=[N:1	1,52
Fc1cccc(c1)C1=[N:1]CCC1	1,66
C1C[N:1]=C(C1)c1ccccc1	1,98
C(=[N:1]c1ccccc1)c1ccccc1	0,87
C([N:1]=Cc1ccccc1)c1cccc	1,18
[NH:1]=C(c1ccccc1)c1cccc	1,80
CC(C)(C)[N:1]=Cc1ccccc1	1,29
CCP(=[S:1])(CC)CC	1,46
[O-:1][n+]1cc(Cl)cc(Cl)c1	1,56
C[S:1]CCl	-0,37
[O-][N+](=O)c1cccc(c1)[N:1	0,86
c1c[n:1]oc1	0,81
c1[n:1]oc2ccccc12	0,68
Cn1ccc[n:1]1	1,84
CCCCCCCCP(=[S:1])(CCC	1,54
COP(=[S:1])(OC)OC	0,58
CCOP(=[S:1])(OCC)OCC	0,76
[S:1]=P(c1ccccc1)(c1ccccc	1,00
C[NH:1]CCC#N	1,37
C[N:1](C)CCC#N	1,15
COc1cccc(c1)[N:1]=C(N(C)	1,41
CC(C)(C)c1cccc[n:1]1	1,42
C=CC#[N:1]	0,70
CN(C)C=[N:1]Cc1cccc(c1)C	1,99
C[N:1]1CCc2ccccc2C1	1,80
Cc1[n:1]c2ccccc2o1	1,48
[O:1]=C1OC=CO1	0,69
C[N:1]=C(N(C)C)N(C)C	3,16
CN(C)C=CC#[N:1]	1,70
CN(C)C=[N:1]N(C)C	2,43
CC1(C)C2CCC1(C)C(=[S:1	0,34
CC(C)c1cccc[n:1]1	1,76
C[O:1]c1cccc(C)c1	0,05

CC(=O)n1cc[n:1]c1	1,86
Clc1ccc(cc1)S(=[O:1])c1cc	1,68
[N:1]#CCc1ccccc1	0,81
c1ccc2c[n:1]ccc2c1	1,94
[S:1]=C1SCCS1	0,30
[OH:1]C12CC3CC(CC(C3)(	1,27
CC(=[O:1])C(F)(F)F	-0,06
C[N:1]=C(N(C)C)c1ccccc1	2,62
CN1CCCN(C)C1=[O:1]	2,79
CN(C)C=[N:1]C1CC1	2,36
CC(C)N(C(C)C)C(=[O:1])C(	2,03
C[NH:1]CCc1ccccc1	2,14
C[N:1]=C(N(C)C)c1ccc(C)c	2,72
CCN(CC)C(=[S:1])N(C)C	1,29
[O:1]=C1C=COC=C1	2,03
CC(Nc1ccc(Cl)cc1)=[N:1]c1	1,07
CN(C)C=[N:1]c1ccc(cc1)C(	1,43
CN(C)C=[N:1]C12CC3CC((	2,52
c1coc[n:1]1	1,30
C[N:1]=Cc1ccccc1	1,49
C[N:1]=C(N(C)C)c1ccc(cc1	1,99
c1ccc2[n:1]c3ccccc3cc2c1	1,95
Cc1c[n:1]cc(C)c1	2,21
C1C2CC3CC1CC(C2)C3C	1,44
CC[O:1]C=C	0,10
[O:1]=C1CCCCCO1	1,63
[O:1]=C1CCCCO1	1,57
CC(=[O:1])c1ccc(cc1)C(F)(	0,78
CN(C)c1ccc(cc1)C#[N:1]	1,23
CCN(CC)c1ccc(cc1)[N+]([C	0,90

225

#SMILES	pKBHX	Comment
COC(=O)c1ccc[n:1]c1	1,44	bifunctional
COC(=[O:1])c1cccnc1	0,51	bifunctional
O=C(c1ccccc1)c1ccc[n:1]c1	1,42	bifunctional
[O:1]=C(c1ccccc1)c1cccnc1	0,68	bifunctional
CC(=O)c1cc[n:1]cc1	1,41	bifunctional
CC(=[O:1])c1ccncc1	0,78	bifunctional
CCOC(=O)c1ccc(cc1)C#[N:1]	0,66	bifunctional
CCOC(=[O:1])c1ccc(cc1)C#N	0,53	bifunctional
CCOC(=O)c1ccc(cc1)[N+]([O-:1])=O	0,16	bifunctional
CCOC(=[O:1])c1ccc(cc1)[N+]([O-])=O	0,5	bifunctional
CC(=[O:1])C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	1,2	bifunctional
CC(=O)C1CCC2C3CCC4=CC(=[O:1])CCC4(C)C3CCC12C	1,75	bifunctional
CC12CCC3C(CCC4=CC(=O)C=CC34C)C1CCC2=[O:1]	1,36	bifunctional
CC12CCC3C(CCC4=CC(=[O:1])C=CC34C)C1CCC2=O	1,84	bifunctional
CN1C(CCC1=O)c1ccc[n:1]c1	1,62	bifunctional
CN1C(CCC1=[O:1])c1cccnc1	2,16	bifunctional
CC(=O)c1ccc(cc1)[N+]([O-:1])=O	0,15	bifunctional
CC(=[O:1])c1ccc(cc1)[N+]([O-])=O	0,57	bifunctional
CC(=O)c1ccc(cc1)C#[N:1]	0,65	bifunctional
CC(=[O:1])c1ccc(cc1)C#N	0,6	bifunctional
CCN(CC)C(=O)c1ccc[n:1]c1	1,63	bifunctional
CCN(CC)C(=[O:1])c1cccnc1	1,98	bifunctional
C(OCc1ccccc1)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,41	bifunctional
C([O:1]Cc1ccccc1)c1ccccc1	0,65	bifunctional
CCOCC[Cl:1]	-0,1	bifunctional
CC[O:1]CCCl	0,44	bifunctional
ClCCOCC[Cl:1]	-0,33	bifunctional
ClCC[O:1]CCCl	-0,03	bifunctional
[CH:1]#[C:1]c1ccccc1	-0,44	bifunctional
C#C[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,6	bifunctional
C[C:1]#[C:1]c1ccccc1	-0,19	bifunctional
CC#C[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,57	bifunctional
c1ccc(cc1)N=N[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,53	bifunctional
c1ccc(cc1)[N:1]=Nc1ccccc1	-0,22	bifunctional
CC(Cl)C[Cl:1]	-0,61	bifunctional
CC([Cl:1])CCl	-0,61	bifunctional
ClCC(Cl)C[Cl:1]	-0,76	bifunctional
ClCC([Cl:1])CCl	-0,76	bifunctional
CO[c:1]1[c:1](C)[cH:1][cH:1][cH:1][c:1]1C	-0,21	bifunctional
C[O:1]c1c(C)cccc1C	0,32	bifunctional
CO[c:1]1[cH:1][c:1](C)[cH:1][cH:1][c:1]1C(C)(C)C	-0,16	bifunctional
C[O:1]c1cc(C)ccc1C(C)(C)C	-1,44	bifunctional
O(c1ccccc1)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,43	bifunctional
[O:1](c1ccccc1)c1ccccc1	-0,8	bifunctional
C1[c:1]2[cH:1][cH:1][cH:1][cH:1][c:1]2Oc2ccccc12	-0,43	bifunctional
C1c2ccccc2[O:1]c2ccccc12	-1,71	bifunctional
CS[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,48	bifunctional
C[S:1]c1ccccc1	-0,32	bifunctional
S(c1ccccc1)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,51	bifunctional
[S:1](c1ccccc1)c1ccccc1	-0,93	bifunctional
C[O:1]CCN	1,09	bifunctional
COCC[NH2:1]	2,26	bifunctional

C[O:1]CCCN	1,16	bifunctional
COCCC[NH2:1]	2,22	bifunctional
NCCC#[N:1]	0,7	bifunctional
[NH2:1]CCC#N	1,33	bifunctional
COCCNCC[O:1]C	0,99	bifunctional
COCC[NH:1]CCOC	2,31	bifunctional
CN1CCCC1c1ccc[n:1]c1	2,03	bifunctional
C[N:1]1CCCC1c1cccnc1	1,11	bifunctional
C1CNC(C1)c1ccc[n:1]c1	2,08	bifunctional
C1C[NH:1]C(C1)c1cccnc1	1,52	bifunctional
[CH:1]#CCN(CC#C)CC#C	-0,64	bifunctional
C#CC[N:1](CC#C)CC#C	0,83	bifunctional
C(N(C[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1)Cc1ccccc1)c1ccccc1	-0,39	bifunctional
C([N:1](Cc1ccccc1)Cc1ccccc1)c1ccccc1	-1,5	bifunctional
CCOC(=[O:1])N1C=CN(C)C1=S	0,72	bifunctional
CCOC(=O)N1C=CN(C)C1=[S:1]	1,32	bifunctional
S=C=N[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,63	bifunctional
[S:1]=C=Nc1ccccc1	-0,55	bifunctional
CN(C)C=Nc1ccc(cc1)[N+]([O-:1])=O	0,48	bifunctional
CN(C)C=[N:1]c1ccc(cc1)[N+]([O-])=O	1,2	bifunctional
CN(C)C=Nc1ccc(cc1)C#[N:1]	1,23	bifunctional
CN(C)C=[N:1]c1ccc(cc1)C#N	1,32	bifunctional
CN(C)C=Nc1ccc(cc1)C(C)=[O:1]	1,7	bifunctional
CN(C)C=[N:1]c1ccc(cc1)C(C)=O	1,52	bifunctional
C[O:1]CCN=CN(C)C	1	bifunctional
COCC[N:1]=CN(C)C	2,74	bifunctional
CN(C)C=NCCC#[N:1]	1	bifunctional
CN(C)C=[N:1]CCC#N	2,12	bifunctional
[N:1]#Cc1ccncc1	0,47	bifunctional
N#Cc1cc[n:1]cc1	0,92	bifunctional
[N:1]#Cc1cccnc1	0,53	bifunctional
N#Cc1ccc[n:1]c1	0,82	bifunctional
[N:1]#Cc1ccccn1	0,61	bifunctional
N#Cc1cccc[n:1]1	0,48	bifunctional
[Cl:1]c1ccccc1	-1,02	bifunctional
Cl[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,65	bifunctional
[Br:1]c1ccccc1	-0,92	bifunctional
Br[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,66	bifunctional
Ic1ccccc1	-0,92	bifunctional
I[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,64	bifunctional
CN1C(CC([OH:1])c2ccccc2)CCCC1CC(=O)c1ccccc1	1,94	bifunctional
C[N:1]1C(CC(O)c2ccccc2)CCCC1CC(=O)c1ccccc1	2,11	bifunctional
COC(=[O:1])C(Cl)(Cl)Cl	0,11	
CCOC(=[O:1])C(F)(F)F	0,08	
CCOC(=[O:1])C(Cl)(Cl)Cl	0,15	
CCOC(=[O:1])C(=O)OCC	0,65	
CC(=[O:1])OC(C)(C)C	1,1	
CCOC(=[O:1])C=CN(C)C	2,09	
CCOC(=[O:1])c1ccc(cc1)C(=O)OCC	0,76	
[O:1]=COc1ccccc1	0,35	
COC=[O:1]	0,65	
CCSC(=[O:1])OC	0,73	
CC(C)(C)OC(=[O:1])c1ccccc1	0,97	

CCOC(=[O:1])c1ccc(F)cc1	0,88
COC(=[O:1])C1CCCCC1	1,06
CCOC(=[O:1])c1cccc(C)c1	1,01
COC(=[O:1])C1CC1	1,12
CC(=O)C(C)=[O:1]	0,23
[O:1]=C(C(=O)c1ccccc1)c1ccccc1	0,44
[O:1]=C1C=CC(=O)C=C1	0,51
[O:1]=C1C=C(Oc2ccccc12)c1ccccc1	1,99
CC(=O)c1ccc(cc1)C(C)=[O:1]	0,92
CC(=[O:1])c1cccc(c1)C(C)=O	0,86
CCN(CC)c1ccc(cc1)C(=[O:1])c1ccc(cc1)N(CC)CC	2,33
CC1(C)CC(=[O:1])C=C(Cl)C1	1,21
CC1=CC(=[O:1])CC(C)(C)C1	1,74
[O:1]=C1C(=C1c1ccccc1)c1ccccc1	2,3
CN(C)C1=CC(=[O:1])CC(C)(C)C1	2,92
CCN(CC)C(=[O:1])N(CC)CC	2,43
CN(C)C(=[O:1])N(C)C	2,44
CN1CCN(C)C1=[O:1]	2,46
ClC(=[O:1])N(c1ccccc1)c1ccccc1	0,75
CCN(CC)C(Cl)=[O:1]	1,08
CN(C)C(=[O:1])C(Cl)(Cl)Cl	1,17
[O:1]=CN(c1ccccc1)c1ccccc1	1,41
[O:1]=C(N(c1ccccc1)c1ccccc1)c1ccccc1	1,61
CC(C)(C)C(=[O:1])N(c1ccccc1)c1ccccc1	1,64
COc1ccc(cc1)C(=[O:1])N(c1ccccc1)c1ccccc1	1,67
CN(C)C(=[O:1])c1ccc(cc1)[N+]([O-])=O	1,9
CN(C)C(=[O:1])c1ccc(cc1)C(F)(F)F	1,97
CN(C)C(=[O:1])c1ccc(Br)cc1	2,07
CN(C)C(=[O:1])c1ccc(F)cc1	2,14
CCN(CC)C(=[O:1])c1ccccc1	2,26
CN(C)C(=[O:1])c1ccc(C)cc1	2,27
COc1ccc(cc1)C(=[O:1])N(C)C	2,31
CCN(CC)C(=[O:1])c1ccc(OC)cc1	2,35
CN(C)C(=[O:1])c1ccc(cc1)N(C)C	2,49
[O:1]=C(Oc1ccccc1)N(c1ccccc1)c1ccccc1	1,18
COC(=[O:1])N(c1ccccc1)c1ccccc1	1,41
CCOC(=[O:1])N(c1ccccc1)c1ccccc1	1,45
COC(=[O:1])N(C)C	1,8
CCOC(=[O:1])N(CC)CC	1,95
[O:1]=C(N(c1ccccc1)c1ccccc1)N(c1ccccc1)c1ccccc1	1,74
CCN(CC)C(=[O:1])N(c1ccccc1)c1ccccc1	2,07
CN(C)C(=[O:1])N(c1ccccc1)c1ccccc1	2,08
CCN(C(=[O:1])N(CC)c1ccccc1)c1ccccc1	2,16
CN1C(=[O:1])CCC1=O	1,06
CCC1(C)CC(=[O:1])NC(=O)C1	1,05
CCN1C(=[O:1])C=CC1=O	0,68
CC(=[O:1])Nc1ccccc1	1,69
CC(=O)OC(C)=[O:1]	0,55
ClS(Cl)=[O:1]	-0,38
CS(=[O:1])c1ccc(cc1)[N+]([O-])=O	1,58
[O:1]=[Se](Cc1ccccc1)Cc1ccccc1	3,3
C[Se](C)=[O:1]	3,43
CCOS(=[O:1])(=O)OCC	0,5

CCOS(C)(=[O:1])=O	0,72
O=S(=[O:1])(c1ccccc1)c1ccccc1	0,91
[O:1]=S1(=O)CCCC1	1,17
CCCCS(=O)(=[O:1])CCCC	1,22
CN(C)S(=O)(=[O:1])c1ccccc1	0,89
CN(C)S(C)(=O)=[O:1]	1
CCN(CC)S(=O)(=[O:1])N(CC)CC	1,17
CN(C)C=NS(=[O:1])(=O)c1ccccc1	1,51
CC[O:1]CCOCC	1,09
C[O:1]CCOC	1,02
CC(C)(C)[O:1]C(C)(C)C	0,75
C[O:1]COC	0,58
C[O:1]C(OC)OC	0,55
COC([O:1]C)(OC)OC	0,26
C[Si](C)(C)[O:1][Si](C)(C)C	-0,53
C[O:1]C(C(F)(F)F)C(F)(F)F	-0,41
C1C2CC3CC1CC(C2)C31[O:1]C11C2CC3CC(C2)CC1C3	1,44
C1C[O:1]CCO1	0,73
C1COC[O:1]C1	0,63
C1C[O:1]CO1	0,45
C1OC[O:1]CO1	0,02
c1cc[o:1]c1	-0,4
CC(C)C12CCC(C)(O[O:1]1)C=C2	0,92
C1C2CC3CC1CC(C2)C31[O:1]OC11C2CC3CC(C2)CC1C3	0,63
CC(C)(C)O[O:1]C(C)(C)C	0,13
C1C[O:1]CCOCCOCCOCCOCCO1	1,2
C1COCCOCC[O:1]CCO1	1,13
C1COCCOCCOCCOCC[O:1]1	1,12
c1c[c:1]2[cH:1][cH:1][c:1]3cccc4ccc(c1)[c:1]2[c:1]34	-0,56
c1cc[c:1]2[c:1](c1)[cH:1][cH:1][c:1]1cccc[c:1]21	-0,53
c1cc[c:1]2[cH:1][cH:1][cH:1][cH:1][c:1]2c1	-0,48
[cH:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,49
c1ccc(cc1)-[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,47
C([c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1)c1ccccc1	-0,41
[cH:1]1[cH:1][cH:1][c:1]([cH:1][cH:1]1)C(c1ccccc1)c1ccccc1	-0,5
C[Si](C)(C)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,36
C[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,36
CC[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,36
CC(C)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,34
C1CCC(CC1)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,32
CC(C)(C)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,32
C1C2CC3CC1CC(C2)(C3)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,28
C[c:1]1[cH:1][cH:1][cH:1][cH:1][c:1]1C	-0,27
C[c:1]1[cH:1][cH:1][cH:1][c:1](C)[cH:1]1	-0,28
C[c:1]1[cH:1][cH:1][c:1](C)[cH:1][cH:1]1	-0,29
C[c:1]1[cH:1][c:1](C)[cH:1][c:1](C)[cH:1]1	-0,18
CC(C)[c:1]1[cH:1][c:1]([cH:1][c:1]([cH:1]1)C(C)C)C(C)C	-0,18
CC(C)(C)[c:1]1[cH:1][c:1]([cH:1][c:1]([cH:1]1)C(C)(C)C)C(C)(C)C	-0,23
C[c:1]1[cH:1][c:1](C)[c:1](C)[cH:1][c:1]1C	-0,15
C[c:1]1[cH:1][cH:1][c:1](C)[c:1](C)[c:1]1C	-0,14
C[c:1]1[cH:1][c:1](C)[c:1](C)[c:1](C)[c:1]1C	-0,07
C[c:1]1[c:1](C)[c:1](C)[c:1](C)[c:1](C)[c:1]1C	0,02
C[C:1](C)=[C:1](C)C	-0,85

C[C:1]=[CH:1]CCCC1	-0,74
C1CC[CH:1]=[CH:1]C1	-0,82
C1C=CC[CH:1]=[CH:1]1	-0,88
CCCCC[CH:1]=[CH2:1]	-0,67
CC[C:1]#[C:1]CC	-0,1
CCCC[C:1]#[CH:1]	-0,22
[CH:1]#[C:1]CCC#C	-0,43
[OH:1]C1CCCCC1	1,14
CCCCCCCC[OH:1]	1,04
CCCC[OH:1]	1,02
CCC[OH:1]	1
[OH:1]CCc1ccccc1	0,97
[OH:1]Cc1ccccc1	0,86
[OH:1]CC=C	0,79
[OH:1]CCF	0,55
[OH:1]CCBr	0,54
[OH:1]CC#C	0,38
Cc1ccc([OH:1])cc1	0,03
Cc1cccc([OH:1])c1	0,01
[OH:1]c1ccccc1	-0,07
[OH:1]CC(Cl)(Cl)Cl	-0,13
[OH:1]CC(F)(F)F	-0,28
[OH:1]c1cccc(c1)C(F)(F)F	-0,36
[OH:1]C(C(F)(F)F)C(F)(F)F	-0,96
N(c1ccccc1)[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,3
CN(C)c1cccc2[cH:1][cH:1][cH:1][c:1](N(C)C)c12	-0,24
C[N:1](C)c1ccc(cc1)N(C)C	1,13
Cc1ccc2N3CN(Cc2c1)[c:1]1[cH:1][cH:1][c:1](C)[cH:1][c:1]1C3	-0,26
[O:1]=P(N1CCCCC1)(N1CCCCC1)N1CCCCC1	3,74
[O:1]=P(Oc1ccccc1)(Oc1ccccc1)Oc1ccccc1	1,89
CC1=[N:1]CCC1	2,59
[F:1]C1C2CC3CC(C2)C([F:1])C1C3	0,48
[F:1]C12CC3CC(CC(C3)C1)C2	0,31
[F:1]C1C2CC3CC(C2)CC1C3	-0,02
[F:1]C1CCCCC1	0,09
[F:1]C12CC3CC(C1)CC(F)(C3)C2	0
CCCCCCCC[F:1]	0,02
CCCCC[F:1]	-0,06
FCCC[F:1]	-0,27
CC([Cl:1])(Cl)Cl	-1,15
ClC[Cl:1]	-0,8
CC([Cl:1])Cl	-0,72
[Cl:1]CCCl	-0,61
ClCCC[Cl:1]	-0,51
ClCCCC[Cl:1]	-0,47
CCCC[Cl:1]	-0,41
[Cl:1]CCCCCCl	-0,39
CCCCC[Cl:1]	-0,38
CC(C)[Cl:1]	-0,3
[Cl:1]C1CCCCC1	-0,27
CC(C)(C)[Cl:1]	-0,28
[Cl:1]C12CC3CC(CC(C3)C1)C2	-0,18
[Cl:1]C1CCCCC1Cl	-0,5

BrC[Br:1]	-0,7
[Br:1]CCBr	-0,63
BrCCC[Br:1]	-0,53
[Br:1]C1CC1	-0,47
BrCCCC[Br:1]	-0,47
CC[Br:1]	-0,4
CCC[Br:1]	-0,38
CCCCC[Br:1]	-0,35
CCCC[Br:1]	-0,34
CC(C)[Br:1]	-0,3
[Br:1]C1CCCCC1	-0,25
CC(C)(C)[Br:1]	-0,22
BrC12CC3CC(CC(C3)C1)C2	-0,17
IC[I:1]	-0,68
[I:1]CCI	-0,65
ICCC[I:1]	-0,51
C[I:1]	-0,47
CC[I:1]	-0,47
ICCCC[I:1]	-0,46
CCCCC[I:1]	-0,37
CC(C)[I:1]	-0,37
CC(C)(C)[I:1]	-0,33
[I:1]C1CCCCC1	-0,32
[I:1]C12CC3CC(CC(C3)C1)C2	-0,19
CN(C)C=C[N+]([O-:1])=O	1,47
CC(C)[N+]([O-:1])=O	0,41
C[N+]([O-:1])=O	0,27
CC[N+]([O-:1])=O	0,32
CC(C)(C)[N+]([O-:1])=O	0,42
CC1(C)C2CCC1(C)C(C2)=N[N+]([O-:1])=O	0,34
[O:1]=Nc1ccccc1	0,15
CCN(CC)c1ccc(cc1)N=[O:1]	1,33
[O:1]=NN1CCCC1	1,49
[Se:1]=P(c1ccccc1)(c1ccccc1)c1ccccc1	0,94
[O-][N+](=O)c1cc[n+]([O-:1])cc1	1,05
[O-:1][n+]1ccc(Cl)cc1	2,44
[O-:1][n+]1ccccc1	2,72
[O-:1][n+]1ccc(cc1)-c1ccccc1	2,85
Cc1ccc[n+]([O-:1])c1	2,92
Cc1cc[n+]([O-:1])cc1	3,12
COc1cc[n+]([O-:1])cc1	3,7
[O:1]=[As](c1ccccc1)(c1ccccc1)c1ccccc1	4,15
C[S:1]SC	-0,49
CC[S:1]SCC	-0,4
C[S:1]CSC	-0,22
C1C[S:1]CS1	-0,38
C1C[S:1]CCS1	-0,14
C1CSC[S:1]C1	-0,06
c1c[n:1]c2c(c1)ccc1ccc[n:1]c21	3,1
c1ccc(cc1)-c1cc[n:1]c2c1ccc1c(cc[n:1]c21)-c1ccccc1	3,26
Cc1cc[n:1]c2c1ccc1c(C)cc[n:1]c21	3,34
Cc1c[n:1]c2c(ccc3c(C)c(C)c[n:1]c23)c1C	3,46
C[O:1]c1ccccc1OC	1,16

COc2ccc([O:1])cc2	-0,08
C[O:1]c1ccc(OC)cc1	0,12
C1Oc2ccccc2[O:1]1	-0,45
C1C[O:1]c2ccccc2O1	-0,23
[NH2:1]C12CC3CC(CC(C3)C1)C2	2,3
[NH2:1]C1CCCCC1	2,29
CCCCCCCC[NH2:1]	2,27
CCCCCCCCCCCCCCCC[NH2:1]	2,26
CC(C)(C)[NH2:1]	2,23
CC(C)[NH2:1]	2,2
CCCC[NH2:1]	2,19
CCC[NH2:1]	2,2
CC[NH2:1]	2,17
[NH2:1]C1CC1	1,72
[NH2:1]CCN	2,25
NCCC[NH2:1]	2,31
NCCCC[NH2:1]	2,21
NCCCCCC[NH2:1]	2,21
[NH2:1]CCc1ccccc1	2,16
[NH2:1]CC=C	1,93
[NH2:1]Cc1ccccc1	1,84
[NH2:1]CC#C	1,56
[NH2:1]CC(F)(F)F	0,71
C[NH2:1]	2,2
C1CC[NH:1]C1	2,59
C1C[NH:1]C1	2,59
C[NH:1]CCNC	2,29
CC1C[NH:1]1	2,28
C[NH:1]C	2,26
C1CNCC[NH:1]1	2,11
C1Cc2ccccc2C[NH:1]1	2,04
C[NH:1]Cc1ccccc1	1,82
C[Si](C)(C)[NH:1][Si](C)(C)C	-0,45
C1C[N:1]2CCC1CC2	2,71
C1C[N:1]2CCN1CC2	2,33
C[N:1](C)CCCCCCN(C)C	2,05
C[N:1](C)CCN(C)C	2,02
CN1CC[N:1](C)CC1	1,88
C[N:1]1CN(C)CN(C)C1	1,58
C1N2CN3CN1C[N:1](C2)C3	1,33
CC[N:1](C(C)C)C(C)C	1,05
CC(C)C[N:1](C(C)C)C(C)C	0,3
CCC(CC)[N:1](C(C)C)C(C)C	-0,34
CN(C)C(=[S:1])N=C(C)N(C)C	2,06
C[Se:1]C	-0,01
[cH:1]1[cH:1][cH:1]s[cH:1]1	-0,5
ClC(Cl)(Cl)C#[N:1]	-0,26
C[Si](C)(C)C#[N:1]	0,93
C#CC#[N:1]	0,3
[O-][N+](=O)c1ccc(cc1)C#[N:1]	0,35
ClC(=C)C#[N:1]	0,36
[N:1]#Cc1ccccc1C#N	0,38
[O-][N+](=O)c1cccc(c1)C#[N:1]	0,43

[N:1]#Cc1cccc(c1)C#N	0,48
Clc1cc(Cl)cc(c1)C#[N:1]	0,52
FC(F)(F)c1ccccc1C#[N:1]	0,57
BrCC#[N:1]	0,57
Clc1cccc(c1)C#[N:1]	0,65
Brc1cccc(c1)C#[N:1]	0,65
ICC#[N:1]	0,67
BrCc1cccc(c1)C#[N:1]	0,7
FC(F)(F)c1cccc(CC#[N:1])c1	0,7
Brc1ccc(CC#[N:1])cc1	0,72
ClCCC#[N:1]	0,73
Brc1ccc(cc1)C#[N:1]	0,73
Clc1cccc(CC#[N:1])c1	0,74
BrCc1ccc(cc1)C#[N:1]	0,75
CSCC#[N:1]	0,77
Fc1ccc(CC#[N:1])cc1	0,77
COc1cc(OC)cc(c1)C#[N:1]	0,78
COc1cccc(c1)C#[N:1]	0,8
Clc1ccc(CC#[N:1])cc1	0,82
ClCCCC#[N:1]	0,83
[N:1]#Cc1ccc(cc1)-c1ccccc1	0,83
Cc1ccc(CC#[N:1])cc1	0,84
Cc1cccc(c1)C#[N:1]	0,85
ClCCCCC#[N:1]	0,87
Cc1cccc(CC#[N:1])c1	0,87
COc1ccc(CC#[N:1])cc1	0,87
Cc1ccc(cc1)C#[N:1]	0,88
CCCCC#[N:1]	0,92
[N:1]#CC1CCCCC1	0,97
CN(C)C=[N:1]CC#C	2,3
CN(C)C=[N:1]c1ccc(cc1)N(C)C	2,32
CN(C)C=[N:1]c1cc(cc(c1)[N+]([O-])=O)[N+]([O-])=O	0,6
CC(C)[N:1]=C(N(C(C)C)C(C)C)N(C(C)C)C(C)C	1,06
CN1CCCN2CCC[N:1]=C12	3,48
C1C[N:1]=C2CCCCCN2C1	3,85
C1CN2CCC[N:1]=C2C1	3,89
N(C=[N:1]c1ccccc1)c1ccccc1	2,13
CNC(=[N:1]c1ccccc1)c1ccccc1	1,83
Cn1c[n:1]cc1Br	2,22
[cH:1]1[cH:1][cH:1][nH][cH:1]1	0,15
Cn1cc[n:1]c1	2,72
C1CCC(CC1)c1c[n:1]cn1C1CCCCC1	3,12
Cn1[cH:1][cH:1][cH:1][cH:1]1	0,23
c1ccc2c[n:1]ncc2c1	1,97
c1ccn[n:1]c1	1,65
Nc1[n:1]cccn1	1,55
c1ccc2[n:1]c3ccccc3nc2c1	1,22
c1ccc([n:1]c1)-c1ccccn1	1,15
c1c[n:1]cnc1	1,07
c1c[n:1]ccn1	0,92
c1ccc2c(c1)ccc1ccc[n:1]c21	1,16
Brc1c[n:1]cnc1	0,59
c1nc[n:1]cn1	0,32

Fc1[n:1]c(F)c(F)c(F)c1F	-0,49
Fc1c(F)c(F)c(C#[N:1])c(F)c1F	0,01
CN(C)C=NC#[N:1]	2,09
CCCC[N+](CCCC)(CCCC)[N-]C#[N:1]	3,24
Cc1c[n:1]c(C)cn1	1,29
Cc1cc(C)nc[n:1]1	1,47
C1CC1[c:1]1[cH:1][cH:1][cH:1][cH:1][cH:1]1	-0,37
C1[c:1]2[cH:1][cH:1][cH:1][cH:1][c:1]2-c2ccccc12	-0,38
C[c:1]1[cH:1][c:1](n[c:1]([cH:1]1)C(C)(C)C)C(C)(C)C	-0,45
CC(C)(C)[c:1]1[cH:1][cH:1][cH:1][c:1](n1)C(C)(C)C	-0,54
CC[N:1](CC)CC(F)(F)F	0,23
[N:1]#CC1CCCC1	0,88
CCN(CC)C(=[O:1])C(C)(C)C	2,13
CCC[N:1](C)C	1,98
CCOP(C)=[O:1]	2,66
CCCOP(C)=[O:1]	2,73
CCP(=[O:1])CC	3,06
CCCCOP(=[O:1])OCCCC	2,49
[O:1]=P(C1CCCCC1)C1CCCCC1	3,05
CN(C)P(C)(=[O:1])N(C)C	3,41
CCCCOP(=[O:1])(CC)OCCCC	2,84
CCOP(=[O:1])(N(C)C)N(C)C	3,18
CC1CCOP(O)(=[O:1])O1	2,4
OP1(=[O:1])OCCC(O1)c1ccccc1	2,38
CC(C)C1OP(O)(=[O:1])OCC1(C)C	2,4
CN1CN(C)C1=[O:1]	2,46
CCN(CC(=[O:1])N(CC)c1ccccc1)c1ccccc1	2,16
CN(C)C=[N:1]CC(C)(C)C	2,26

Molnb.	#SMILES	pK_BHX
1	CCCC[OH:1]	0,934
2	CC(C)CC[OH	0,832
3	[OH:1]c1cccc	0,242
4	C[O:1]c1ccc(	-0,173
5	CC(=[O:1])O	1,858
6	[O:1]=C(N(c1	1,5
7	CN(C)C(=[O:	2,17
8	CCN(CC)C(=	2,001
9	COC(=O)c1c	1,388
10	[O:1]=C1CC(	0,833
11	[O:1]=P(Oc1	1,618
12	CCCCNP(=[O	2,798
13	CCCCNP(=[O	3,25
14	CCCCNP(=[O	3,919
15	CN(C)S(C)=[	2,173
16	CN(C)S(=[O:	1,877
17	CN1CCCC1=	1,201
18	CN(C)C1=[N	2,317
19	Cn1cc[n:1]c1	2,322
20	CCc1[n:1]ccn	2,591
21	Cn1c[n:1]c2c	2,079
22	c1cn(c[n:1]1)	2,079
23	c1ccn[n:1]c1	1,327
24	c1cnc[n:1]c1	1,193
25	CCn1ccc[n:1	1,732
26	CCn1ccc(C)[	1,847
27	CCn1[n:1]ccc	1,924
28	CCn1[n:1]c(C	2,157
29	C=Cn1ccc[n:	1,076
30	Cc1ccn(C=C	1,306
31	Cc1cc[n:1]n1	1,333
32	Cc1cc(C)n(C	1,291
33	CCn1[n:1]cc2	1,787
34	C=Cn1[n:1]cc	1,412
35	COP(C)(=[O:	2,591
36	CP(=[O:1])(C	2,015