



Développement de méthodes et d'outils chémoïnformatiques pour l'analyse et la comparaison de chimiothèques

Vincent Le Guilloux

► To cite this version:

Vincent Le Guilloux. Développement de méthodes et d'outils chémoïnformatiques pour l'analyse et la comparaison de chimiothèques. Chimie thérapeutique. Université d'Orléans, 2013. Français. NNT : 2013ORLE2079 . tel-01265387

HAL Id: tel-01265387

<https://theses.hal.science/tel-01265387>

Submitted on 1 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE

Pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE D'ORLEANS

Université d'Orléans

Diplôme National - Arrêté du 7 août 2006

ECOLE DOCTORALE SCIENCES ET TECHNOLOGIES

Discipline : Chimie théorique et informatique

Par

Vincent LE GUILLOUX

**Développement de méthodes et d'outils
chémoinformatiques pour l'analyse et la comparaison de
chimiothèques.**

Directeur de thèse : **M. Luc Morin-Allory**

Co-encadrant : **M. Philippe Vayer**

Soutenue le 13 Décembre 2013

JURY

M. Bruno Villoutreix	MTi, Université Paris 7	Président
M. Alexandre Varnek	UPL, Université de Strasbourg	Rapporteur
M. Ronan Bureau	CERMN, Université de Caen	Rapporteur
M. Xavier Morelli	iSCB, CRCM, Université de Marseille	Examineur
M. Pascal Bonnet	ICOA, Université d'Orléans	Examineur
M. Philippe Vayer	Technologie Servier	Examineur, Co-encadrant
M. Luc Morin-Allory	ICOA, Université d'Orléans	Directeur

Je dédie cette thèse à Noémie et Eliot, à Ophélie, et à mon frère, Jonathan.

Remerciements

Après un sprint final plutôt épique, quel n'est pas mon soulagement d'avoir pu finir cette thèse. Il va de soit que cela n'aurait pas été possible sans l'aide et le soutien (et les quelques coups de pieds au derrière qui vont bien) de tant de personnes qu'il convient de remercier très sincèrement.

Je remercie tout d'abord les membres de mon jury pour avoir pris le temps de lire cette thèse et de venir jusqu'à Orléans afin de m'écouter la présenter.

Je tiens ensuite à remercier tout particulièrement le Pr. Luc Morin-Allory ainsi que le Dr. Philippe Vayer pour m'avoir encadré de la meilleure des manières durant ces trois ans. Merci d'abord du fond du coeur à tous les deux de m'avoir permis de mettre mon début de thèse entre parenthèses lorsque cela s'est avéré nécessaire. Merci Luc pour tous vos conseils, votre générosité, votre humanité et votre honnêteté, et pour avoir supporté mes retards dans la rédaction. Et surtout profitez bien d'une retraite bien méritée et peut être un peu studieuse quand même?! Merci Philippe pour m'avoir fait profiter de ton expertise dans le domaine de l'ADME, et pour m'avoir appris à mieux gérer mon temps et mes projets; on peut dire que je partais de loin! Merci enfin à tous les deux de la patience dont vous avez fait preuve pour cette fin de thèse pour le moins laborieuse!

Merci ensuite à tous les membres de l'ICOA et de Servier avec qui j'ai pu passer ces trois années. Merci à Stéphane pour ta bonne humeur quotidienne, ta gentillesse permanente et pour tes conseils précieux notamment pour la rédaction. Merci à Lionel pour ta franchise, nos débats enflammés vont me manquer! Merci Alban pour ta bonne humeur, ton enthousiasme et

Remerciements

ton ouverture d'esprit, toutes nos discussions auront toujours été un grand plaisir (et vivent les SOM!). Merci à Véro également pour toutes ces discussions philosophiques et parfois déprimantes. Je remercie également Guillaume pour le travail qu'il a pu faire durant son stage sur les DRCS et les cartes de Kohonen. Et tous ceux de l'ICOA que j'oublie : Laurent, Julie, Mathieu, Jérôme, Marie-Aude...

Je remercie également toute l'équipe du Master de Bioinformatique de l'université Paris 7, pour la qualité de la formation proposée, sans laquelle je n'aurais certainement pas entrepris de faire une thèse, et pour la bonne humeur de toute l'équipe. Je remercie également Nicolas Baurin qui m'a transmis son goût pour la chimoinformatique (et qui m'a fait découvrir l'ICOA) durant un stage fort intéressant à Sanofi.

Un merci tout particulier à Peter pour tout le bon travail qu'il a pu accomplir autour de notre cher fpocket, et pour m'avoir permis durant ma thèse de m'échapper de temps en temps de la chimoinfo et d'explorer un peu le côté obscur des protéines. Pour la suite, voyons voyons... ;). Un grand merci également à Jean-Tristan et Guilhem pour toutes ces soirées parisiennes mémorables. Un petit coucou à Alex également pour m'avoir ponctuellement inondé de sa bonne humeur et de ses blagues pourries.

Je remercie bien sûr tous mes amis d'enfance : Alice, Cec, Karen, Mag, Tiphannie, Valéria, David, Greg, Nicolas, Renaud, Seb, (...) pour toutes ces années (et les années à venir !) de bonheur. Un merci tout particulier à Morgane, Tiphannie, Nicole, Laurent et David pour m'avoir hébergé un nombre incalculable de fois à Paris durant ces trois dernières années !

Un remerciement affectueux à mes parents, Ghislaine et Bernard, pour m'avoir élevé de la meilleure des manières !

Table des matières

Liste des figures	vii
Liste des tableaux	ix
Introduction générale	1
Contributions	5
I Chémoinformatique et recherche de médicaments	9
1 Contexte : la recherche de médicaments	9
1.1 Recherche et Développement (R&D)	10
1.2 Essais cliniques et commercialisation	12
2 Les chimiothèques au coeur du processus	13
2.1 Cribler plus pour gagner plus	14
2.2 Optimiser la sélection : les faux positifs	15
2.3 Rationaliser la sélection	16
2.3.1 Composés "drug-like"	17
2.3.2 Composés "lead-like"	17
2.3.3 Chimiothèques focalisées	19
2.3.4 Chimiothèques diverses	19
2.3.5 L'oeil du chimiste	19
2.4 Chimiothèques et HTS dans le milieu académique	20
2.5 Conclusion	21
3 L'apport de la chémoinformatique	21

3.1	Généralités	21
3.2	Représentation de l'information chimique	23
3.2.1	Représentations visuelles	23
3.2.2	Représentations textuelles	24
3.2.3	Représentation en mémoire vive	26
3.3	Descripteurs moléculaires	27
3.3.1	Descripteurs 1D	27
3.3.2	Descripteurs 2D	27
3.3.3	Descripteurs 3D	28
3.3.4	Descripteurs de descripteurs	28
3.3.5	Empreintes moléculaires	28
3.3.6	Descripteurs basés sur les fragments	30
3.3.7	Schémas de simplification du graphe moléculaire	31
3.4	Espaces chimiques	33
3.4.1	Définition générale	33
3.4.2	Coordonnées et métriques	34
3.4.3	Visualisation d'espaces chimiques	36
3.4.4	Conclusion	45
3.5	La Diversité	45
3.5.1	Méthodes de quantification de la diversité	48
3.5.2	Méthodes de sélection par diversité	52
3.6	Conclusion	55
Références bibliographiques		57
II Délimitation d'espaces chimiques réduits		77
1	Introduction	77
2	Visual Characterization and Diversity Quantification of Chemical Libraries : 1. Creation of Delimited Reference Chemical Subspaces	80
III Quantification de la diversité à l'aide d'espaces chimiques délimités		133
1	Introduction	133
2	Visual Characterization and Diversity Quantification of Chemical Libraries : 2. Analysis and Selection of Size-Independent, Subspace-Specific Diversity Indices .	135

IV	Conception et développement de Screening Assistant 2	195
1	Bref historique du projet	195
2	L'open-source en chémoinformatique	197
2.1	Avantages et inconvénients	197
2.2	Librairies chémoinformatiques	200
2.3	Cartouches moléculaires : la chimie en base de données	201
2.4	Outils graphiques	202
3	Objectifs	203
4	Screening Assistant 2	205
4.1	Architecture	205
4.1.1	Langage	205
4.1.2	Plateforme NetBeans	205
4.1.3	Base de données	207
4.2	Interface graphique	209
4.3	Gestion des molécules	209
4.3.1	Identifiants	209
4.3.2	Format d'entrée	209
4.3.3	Standardisation et unicité	210
4.4	Insertion des molécules	212
4.4.1	Perception des molécules	213
4.4.2	Descripteurs	213
4.4.3	Marquage des molécules	213
4.4.4	Squelettes moléculaires	214
4.4.5	Travailleurs	215
4.5	Gestion de sous-ensembles de molécules	216
4.6	Gestion des descripteurs	217
4.7	Fonctionnalités de Recherche	219
4.7.1	Optimisation de la recherche par sous-structure	220
4.8	Espaces chimiques	221
4.8.1	Création d'espaces chimiques	221
4.8.2	Visualisation d'espaces chimiques	222
4.8.3	DRCS	222
4.9	Comparaison de chimiothèques et mesures de diversité	222
4.9.1	Mesures basées sur les scaffolds	223
4.9.2	Mesures basées sur les fingerprints	224

4.9.3	Histogrammes	224
4.10	Génération de sous-ensembles divers	225
5	Discussions et perspectives	226
Références bibliographiques		231
6	Mining collections of compounds with Screening Assistant 2	234
Conclusion générale		257
Annexes		261
1	Descripteurs calculés automatiquement par SA2	261
2	Schéma de base de données pour la gestion des descripteurs et des espaces chimiques	262
3	Guide de démarrage rapide SA2	262

Liste des figures

I.1	Le processus (simplifié) de recherche de médicaments découpé en 4 grandes étapes.	10
I.2	Évolution de la taille moyenne des chimiothèques propriétaires de quatre grandes entreprises pharmaceutiques entre 2001 et 2009. Figure extraite de [11]	14
I.3	Différentes représentations de l'aspirine disponibles dans la base de données ChEMBL [118]	26
I.4	Les différentes règles de réduction de molécules pour en générer le squelette présentées pour la molécule d'Azelaistine (antihistaminique) : (a) le scaffold, tel que défini par Bemis et Murcko, contenant l'ensemble des systèmes cycliques sans les chaînes latérales ; (b) le framework, équivalent du scaffold avec les atomes transformés en carbone et les liaisons en liaisons simples ; (c) les assemblages de cycles, comme étant l'ensemble des cycles fusionnés ou connectés par une seule liaison ; (d) les systèmes cycliques, comme étant le plus grand sous-ensemble de cycles fusionnés. Notons que les hydrogènes sont systématiquement supprimés avant l'application du schéma de réduction et qu'à l'exception du framework les doubles liaisons exo-cycliques sont systématiquement conservées.	31
I.5	Une représentation de la base de données Pubchem <i>via</i> un modèle ACP calculé sur 42 descripteurs (les MQNs). Figure extraite de Van Deursen et al. [183] . .	39
I.6	Une représentation de l'espace chimique peuplé par la base GDB. Les couleurs représentent les différents types de molécules, classées selon leur composition chimique. Figure extraite de Fink et al. [182]	42

I.7	Un exemple de molécules très similaires mais ayant une activité biologique différente. En vert, des molécules ciblant une tyrosine kinase et en noir, des molécules ciblant une monoamine oxidase. Figure extraite de [229]	47
IV.1	Historique d'apparition des librairies en chémoinformatique. Figure réalisée par Andrew Dalke et présentée lors de l'EuroQSAR 2008	198
IV.2	Périmètre d'utilisation de SA2.	204
IV.3	Vue globale du schéma de la base de données pour la dernière version de SA2. Les tables en bleu correspondent aux tables principales pour la gestion des molécules, des scaffolds, des frameworks, et des sous-ensembles. Les tables en vert correspondent aux tables utilisées pour la gestion des espaces chimiques. Les tables en rouge correspondent aux tables gérant les propriétés et les fingerprints. Les tables en orange et magenta correspondent aux tables de stockage de descripteurs (resp. fingerprints) calculables par SA2, à l'exception des tables MOE2D et MOE3D. Les tables en gris correspondent à des tables de configuration.	208
IV.4	Tables principales pour la gestion des molécules.	210
IV.5	Processus d'insertion de molécules dans SA2.	212
IV.6	Exemple de scaffold et de framework généré pour la molécule de Cinolazepam (Drugbank ID : DB01594).	215
IV.7	Conversion d'un fingerprint binaire vers une représentation sous forme de liste d'entiers permettant un stockage plus compact dans la base de données. On fait l'hypothèse ici qu'un fingerprint de 50 bits doit être inséré dans la base. Il est nécessaire d'ajouter des bits à 0 (en rouge) pour ce fingerprint afin de le représenter sous la forme de deux entier 32 bits. On obtient ainsi deux entiers qui peuvent être aisément stockés en base de données.	219
A.1	Tables permettant la gestion des propriétés et des espaces chimiques.	262

Liste des tableaux

- I.1 Métriques utilisées en chémoinformatique pour comparer des vecteurs numériques. N représente le nombre de variables composant chaque vecteur. xA_i et xB_i représentent la valeur de la variable i pour les molécules A et B, respectivement. Pour les fingerprints binaires (seconde colonne), a représente le nombre de bits à 1 dans le fingerprint A, b le nombre de bits à 1 dans le fingerprint B et c le nombre de bits à 1 communs à A et B. 35
- IV.1 Liste des librairies principales de chémoinformatique open-source. La colonne 'Actif' indique si le projet est toujours maintenu. Trois licences peuvent être trouvées ici. La GPL est la plus restrictive, puisque que tout outil basé sur une librairie GPL doit lui même être redistribué sous cette même licence. La LGPL est assez similaire, mais permet la redistribution de la librairie sous sa forme binaire sans imposer la licence du produit final. La BSD est la plus permissive et permet globalement au développeur de faire ce qu'il souhaite avec les outils distribués sous cette licence. 200
- IV.2 Présence de fonctionnalités essentielles à la manipulation de molécules dans les librairies principales de chémoinformatique. Note : l'outil de tracé de molécule du CDK fait partie du projet JChemPaint, mais celui-ci est intimement lié au CDK et les auteurs de ces deux outils proposent une librairie combinant les fonctionnalités du CDK et de JChemPaint. Par ailleurs, GGA software (l'entreprise derrière indigo) propose un outil de tracé web (utilisable uniquement dans un navigateur internet) qui ne peut donc pas être utilisé en client lourd. 201
- IV.3 Liste des méthodes permettant de créer de nouvelles librairies dans SA2. . . . 216

A.1	Liste des propriétés calculées par SA2 lors de l'insertion des molécules. Les propriétés marquées par une étoile ont des valeurs spécifiques du Molecular Handler utilisé.	261
-----	--	-----

Abréviations

ACE	Angiotensin Converting Enzyme
ACP	Analyse en Composantes Principales
ADME	Absorption, Distribution, Métabolisme et Excrétion
API	Application Programming Interface
CATS	Chemically Advanced Template Search
CDK	Chemistry Developement Kit
ChemGPS	Chemical Global Positioning System
ChemGPS-NP	Chemical Global Positioning System for Natural Products
COBRA	COllection of Bioactive Reference Analogues
DRCS	Delimited Reference Chemical Space
DUD	Directory of Usefull Decoys
ECFP	Extended Connectivity FingerPrints
FCFP	Functional Connectivity FingerPrints
GDB	Generated DataBase
GTM	Generative Topographic Map
HTS	High Throughput Screening
InChI	IUPAC International Chemical Identifier

ISIDA	In Silico Design and Data Analysis
IUPAC	International Union of Pure and Applied Chemistry
MDS	MultiDimensional Scaling
MOE	Molecular Operating Environment
NBP	NetBeans Platform
NIH	National Institute of Health
PAI	Promiscuous Aggregating Inhibitors
QSAR	Quantitative Structure-Activity Relationships
RCP	Rich Client Platform
SA	Screening Assistant
SA2	Screening Assistant 2
SDF	Structure Data Format
SGBD	Système de Gestion de Bases de Données
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry Specification
SOM	Self-Organizing Map
SPE	Stochastic Proximity Embedding
SQL	Structured Query Language
SSSR	Smallest Set of Smallest Rings
SWT	Standard Widget Toolkit

Introduction générale

Le processus de recherche de médicaments ne s'apparente pas à un long fleuve tranquille. Malgré de nouvelles technologies toujours plus innovantes, ces dernières décennies ont en effet vu le nombre de mises sur le marché de nouveaux médicaments stagner voire diminuer. Cet état de fait est d'autant plus surprenant qu'il n'est pas si récent (15-20 ans) et pourtant toujours autant d'actualité, tandis que la quantité de données explose et que l'accès à l'information n'a jamais été aussi facile qu'aujourd'hui. On peut aisément mettre en avant deux grandes révolutions qui ont transformé notre manière de faire de la recherche au cours des deux dernières décennies : les techniques expérimentales haut débit et l'informatique. L'avènement de ces nouvelles technologies a eu pour conséquence principale, à défaut de produire plus de médicaments, la génération d'une quantité immense de données, et l'apparition de nouvelles problématiques liées à leur exploitation. D'un côté, les technologies haut débit ont permis, par exemple, de séquencer le génome humain et de tester l'activité biologique de près d'un million de molécules en quelques jours seulement. De l'autre côté, l'informatique s'est introduite dans le quotidien de tout chercheur, et permet en quelques clics d'accéder à d'innombrables bases de données à travers l'utilisation d'Internet.

De nouveaux domaines ont alors vu le jour, à l'interface entre biologie, chimie et informatique, afin de répondre aux multiples problématiques liées à la recherche de médicaments. Cette thèse se situe à l'interface de plusieurs de ces domaines, regroupés sous la bannière de la **chémo-informatique**. Récent à l'échelle humaine, ce domaine fait néanmoins déjà partie intégrante de la recherche pharmaceutique. De manière analogue à la bioinformatique, son pilier fondateur reste le stockage, la représentation, la gestion et l'exploitation par ordinateur de données provenant de la chimie. La chémo-informatique est aujourd'hui utilisée principalement dans les

phases amont de la recherche de médicaments. En combinant des méthodes issues de différents domaines (chimie, informatique, mathématique, apprentissage, statistiques, etc.), elle permet la mise en oeuvre d'outils informatiques adaptés aux problématiques et données spécifiques de la chimie, tels que le stockage de l'information chimique en base de données, la recherche par sous-structure, la visualisation de données, ou encore la prédiction de propriétés physico-chimiques et biologiques.

Dans ce cadre pluri-disciplinaire, le travail présenté dans cette thèse porte sur deux aspects importants liés à la chémoinformatique : (1) le développement de nouvelles méthodes permettant de faciliter la visualisation, l'analyse et l'interprétation des données liées aux *ensembles de molécules*, plus communément appelés **chimiothèques**, et (2) le développement d'outils informatiques permettant de mettre en oeuvre ces méthodes.

Dans une première partie, nous présenterons le contexte de ce travail, ainsi que les concepts et méthodes clés sur lesquels il repose. Nous décrirons tout d'abord le processus de recherche de médicaments et la manière dont les chimiothèques sont utilisées durant les campagnes de criblage. Le domaine de la chémoinformatique sera ensuite introduit et un état de l'art sera réalisé sur les méthodes chémoinformatiques permettant de gérer, d'analyser, et de comparer les chimiothèques. On définira ainsi les notions de descripteurs, d'espace chimique et de diversité, ainsi que les différentes méthodes et applications en chémoinformatique qui apportent des solutions concrètes quant à leur analyse.

Dans une deuxième partie, on présentera une nouvelle méthode permettant de délimiter les zones les plus denses de l'espace chimique : les Espaces Chimiques Délimités de Référence (DRCS - Delimited Reference Chemical Space). Ces espaces se présentent sous la forme d'espaces réduits obtenus à l'aide de l'Analyse en Composantes Principales (ACP). La méthode repose en particulier sur la création d'une enveloppe convexe (convex hull) moyenne calculée à partir de la projection sur les premières composantes principales de plusieurs sous-ensembles de molécules dont on aura filtré au préalable les plus "exotiques" afin de délimiter la zone de l'espace la plus dense. Plusieurs chimiothèques seront analysées et projetées à l'aide de cette méthode afin de mettre en avant ses avantages, tout en discutant ses limitations.

Dans une troisième partie, on présentera une application de cette nouvelle méthode à travers la création d'indices de diversité permettant de comparer des chimiothèques de tailles différentes. Complémentaires à l'analyse visuelle, ces indices permettent de comparer des chimiothèques de

manière plus fine, afin de sélectionner la plus diverse en vue d'un criblage haut débit, ou la plus complémentaire à un ensemble de molécules existant en vue d'une étude d'enrichissement. Des indices existants seront adaptés à notre méthode et leurs comportements seront comparés dans des situations fictives afin de mieux comprendre les leurs différences. Cette analyse nous permettra de mettre en évidence l'absence d'indice(s) parfait(s) et d'extraire un nombre réduit d'indices les plus pertinents à utiliser pour comparer des chimiothèques.

Dans une quatrième partie, on présentera le logiciel Screening Assistant 2 (SA2). SA2 est un logiciel issu du projet Screening Assistant (SA) initié au laboratoire dans le cadre de la thèse d'Aurélien Monge. Une première version du logiciel a été développée par celui-ci au cours de sa thèse et mise à disposition de tous en 2006. L'outil permet de stocker des chimiothèques en bases de données et faciliter leur analyse à travers l'utilisation de méthodes chémoinformatiques. Il a été entièrement redéveloppé afin de permettre, d'une part l'utilisation d'une plateforme logicielle avancée facilitant les bonnes pratiques de la programmation modulaire et d'autre part, l'introduction d'un nombre important de nouvelles méthodes et fonctionnalités qui nécessitaient de profonds changements dans le modèle de base de données utilisé. L'objectif de ce travail était ainsi de permettre à la fois une gestion plus simple et flexible des chimiothèques, mais également l'utilisation de méthodes chémoinformatiques avancées, dont certaines développées au cours de cette thèse.

On conclura par un résumé des résultats et des différentes contributions apportées par ces travaux, et on discutera les limitations et les perspectives qui en découlent.

Contributions

Toute thèse scientifique moderne est un travail d'équipe et s'appropriier l'ensemble des travaux en découlant serait une imposture. Les résultats finaux ne sont par ailleurs pas totalement en adéquation avec les objectifs initiaux et les résultats présentés ne reflètent pas l'ensemble des travaux effectués. Je présente ici brièvement mes différentes contributions aux travaux réalisés au cours de cette thèse, ainsi que les projets n'ayant pas été présentés dans ce manuscrit.

Ma contribution personnelle à ce travail a été principalement la mise en place de la méthodologie des DRCS présentée dans les chapitres 2 et 3, ainsi que la définition, la conception et le développement de Screening Assistant 2 (SA2) (chapitres 6 et 7) dans lequel fut intégrée la méthodologie DRCS. J'ai participé à l'étude sur les indices de diversité (chapitres 4 et 5), essentiellement pour la définition de sous-espaces ciblés, et la rédaction de l'article, et j'en profite pour remercier en particulier le Dr. Lionel Colliandre qui en est le contributeur principal. Bien que ces indices reposent entièrement sur les DRCS, ceux-ci n'ont pas encore pu être intégrés à la version actuelle de SA2. Enfin, il convient de mentionner d'autres projets n'ayant pas pu être intégrés dans ce document :

- Le développement d'une méthode permettant de coupler une visualisation intuitive à l'attribution d'un score unique combinant différentes propriétés ADME préalablement transformées à l'aide de fonctions de désirabilité. Cette méthode a permis le développement d'un logiciel ayant été déployé à Technologie Servier.
- Une évaluation de méthodes QSAR d'habitude utilisées pour la prédiction de valeurs manquantes dans les jeux de données de puces à ADN. Ces méthodes ont été appliquées à des jeux de données issues de projets de chimie médicinale internes à Servier afin d'évaluer leur capacité à prédire des propriétés ADME. Nous avons constaté qu'aucune

d'entre elles n'était plus performante que la méthode classique des K plus proches voisins. Un poster a néanmoins permis de présenter une partie des résultats obtenus lors du congrès ICCS 2011 de Noordwijkerhout.

- Des études sur les cartes de Kohonen et le développement d'un logiciel interactif permettant leur utilisation sur des jeux de molécules. Ce logiciel n'a pas pu être intégré à SA2 et ne sera donc pas décrit ici.

Le travail présenté dans ce manuscrit a fait l'objet des publications, posters et logiciels open-source suivants :

Publications

Vincent Le Guilloux, Alban Arrault, Lionel Colliandre, Stéphane Bourg, Philippe Vayer and Luc Morin-Allory, Mining Chemical Libraries with "Screening Assistant 2", Journal of Cheminformatics 2012, 4 :20

Colliandre, L. ; Le Guilloux, V. ; Bourg, S. ; Morin-Allory, L. Visual characterization and diversity quantification of chemical libraries : 2. analysis and selection of size-independent, subspace-specific diversity indices J. Chem. Inf. Model. 2012, 52, 327-342.

Le Guilloux, V. ; Colliandre, L. ; Bourg, S. ; Guenegou, G. ; Dubois-Chevalier, J. ; Morin-Allory, L. Visual characterization and diversity quantification of chemical libraries : 1. creation of delimited reference chemical subspaces J. Chem. Inf. Model. 2011, 51, 1762-1774.

Posters

Le Guilloux, V. ; Bourg, S. ; Dubois-Chevalier, J. ; Colliandre, L. ; Arrault, Al. ; Vayer, P. ; Morin-Allory, L. Managing chemical libraries using Screening Assistant 2, Journées nationales de Chemoinformatique 10-2011 - Cabourg.

Le Guilloux, V. ; Bourg, S. ; Dubois-Chevalier, J. ; Colliandre, L. ; Arrault, Al. ; Vayer, P. ; Morin-Allory, L. Managing chemical libraries using Screening Assistant 2, Groupe de Graphisme et Modélisation Moléculaire (GGMM 2011) 05-2011 - La Rochelle.

Colliandre, L. ; Bourg, S. ; Dubois-Chevalier, J. ; Morin-Allory, L. ; Le Guilloux, V. Visual characterization and diversity quantification of chemical libraries, Groupe de Graphisme et Modélisation Moléculaire (GGMM 2011) 05-2011 - La Rochelle.

Le Guilloux, V. ; Bourg, S. ; Dubois-Chevalier, J. ; Colliandre, L. ; Arrault, Al. ; Vayer, P. ; Morin-Allory, L. Managing chemical libraries using Screening Assistant 2, 9th International Conference on Chemical Structures 06-2011 - Noordwijkerhout (Hollande).

Le Guilloux, V. ; Arrault, Al. ; Morin-Allory, L. ; Vayer, P. Hole filling in medicinal chemistry libraries, 9th International Conference on Chemical Structures 06-2011 - Noordwijkerhout (Hollande).

Le Guilloux, V. ; Guenegou, G. ; Bourg, S. ; Dubois, J. ; Morin-Allory, L. ; Colliandre, L. Implementation of a reference chemical space for HTS commercial compounds : application to the comparison of chemicals libraries, Ecole Thématique de Criblage 09-2010 - Marseille.

Le Guilloux, V. ; Guenegou, G. ; Bourg, S. ; Dubois, J. ; Morin-Allory, L. ; Colliandre, L. Implementation of a reference chemical space for HTS commercial compounds : application to the comparison of chemicals libraries, Seconde école d'été de Strasbourg sur la chemoinformatique 06-2010 - Obernai.

Logiciels

Screening Assistant 2, un logiciel open-source de gestion et d'analyse de chimiothèques.
[http ://sa2.sourceforge.net](http://sa2.sourceforge.net)

DRCS Tools, des outils en ligne de commande open-sources facilitant l'utilisation des DRCS.
[http ://www.univ-orleans.fr/icoa/DRCS/index.html](http://www.univ-orleans.fr/icoa/DRCS/index.html)

Chapitre I

Chémoinformatique et recherche de médicaments

1 Contexte : la recherche de médicaments

La recherche de médicaments est un processus multi-disciplinaire extrêmement long (10-20 ans) et coûteux. On trouve dans la littérature des estimations assez divergentes concernant son coût global, évalué entre 300 millions et plus de 1,7 milliard de dollars [1]. L'étude de DiMasi [2, 3], l'une des plus citées, estime ce coût à environ 802 millions de dollars, tandis qu'une étude plus récente le situe autour de 1,2 milliard [4]. La plupart des études s'accordent néanmoins sur la quantité importante de ressources nécessaires pour mener un projet à terme, ainsi que sur l'augmentation continue du coûts global du processus, celui-ci ayant été estimé à 138 millions de dollars dans les années 70 et 300 millions dans les années 80 [5]. Pour ne rien arranger, le taux de succès des projets de l'industrie pharmaceutique reste faible [6] et le nombre de nouveaux médicaments mis sur le marché stagne, voire diminue depuis les années 2000[5].

Les acteurs de la recherche pharmaceutique tentent ainsi constamment d'optimiser chacune des étapes menant à la commercialisation d'un médicament. Les entreprises pharmaceutiques font notamment partie de celles qui investissent le plus au monde, avec en moyenne 10 à 20 % de leur chiffre d'affaires réinvesti dans la recherche et le développement (R&D) [7]. L'ensemble du processus de recherche de médicaments peut se diviser en quatre grandes étapes présentées dans la figure I.1. Nous allons les décrire brièvement dans les sections qui suivent, en insistant sur les étapes pour lesquelles ont été développées les méthodes présentées dans cette thèse.



Figure I.1 – Le processus (simplifié) de recherche de médicaments découpé en 4 grandes étapes.

1.1 Recherche et Développement (R&D)

L'un des principaux objectifs de la phase R&D est de délivrer un ou plusieurs candidats médicaments actifs et innovants, présentant le moins de toxicité possible et ayant ainsi le plus de chance de passer avec succès les étapes qui suivront. Elle se divise en trois grandes étapes.

Identification et validation de cibles

La première étape consiste à identifier une ou plusieurs entités biologiques (généralement des protéines) dont la modulation de l'activité permettrait d'obtenir un effet bénéfique par rapport à une maladie ciblée. Cette étape est essentielle, en ce sens que toutes les recherches qui suivront se baseront sur l'hypothèse que ces cibles sont effectivement liées à la maladie et que l'action visée par les médicaments aura un effet positif sur l'homme. Elle est d'autant plus délicate que la pertinence de la cible par rapport à la maladie ciblée doit être mise en balance avec les effets secondaires qui pourraient apparaître en conséquence de la modification de son activité. Depuis le séquençage du génome humain et le développement de méthodes bioinformatiques permettant la comparaison de séquences et structures protéiques, la recherche de nouvelles cibles a pris un nouvel essor et le domaine est aujourd'hui très actif. De nombreux auteurs ont d'ailleurs remis en question le paradigme classique "Une maladie, Une cible, Un médicament", dans la mesure où l'effet d'une molécule sur sa cible peut être compensé par la régulation de réseaux biologiques [8, 9], sans parler du fait qu'elle a de bonnes chances d'interagir avec d'autres protéines. De nouveaux domaines ont ainsi vu le jour ces dernières décennies : la polypharmacologie [10] et la chémogénomique.

Génération de Hits et de Leads

Une fois la ou les cible(s) identifiée(s), une très grande majorité des projets de recherche de médicaments se poursuivent par une étape de criblage. Celle-ci a pour but d'identifier un premier ensemble de molécules actives, plus communément appelées *hits*. Pour être identifiée en tant que telle, une molécule doit démontrer un certain niveau d'activité (le plus souvent de l'ordre du micromolaire) lors du test de criblage. La valeur précise du niveau d'activité requis n'étant pas absolue, on retiendra globalement qu'un *hit* est une molécule ayant montré une ac-

tivité modérée ou forte lors du test expérimental. Il existe différentes méthodes expérimentales permettant l'identification de *hits*. Aujourd'hui, le criblage expérimental à haut débit - High Throughput Screening (HTS) [11–13] - est probablement la méthode la plus utilisée dans l'industrie pharmaceutique. On distingue deux grands types de criblages expérimentaux [5, 14–16] :

Criblages sur cibles : les molécules sont testées sur des systèmes biochimiques assez simples, généralement pour leur affinité ou leur capacité d'inhibition sur une protéine. C'est le type de criblage le plus commun car il permet de tester un grand nombre de molécules en peu de temps et parce qu'il correspond bien au paradigme classique utilisé dans la recherche pharmaceutique : une cible / un médicament.

Criblages phénotypiques : les molécules sont testées sur des cellules entières ou sur des modèles animaux de la maladie ciblée. Plus lents et plus coûteux, ils permettent néanmoins d'observer l'activité de la molécule dans un contexte cellulaire et se veulent ainsi plus objectifs. Ce type de criblage était beaucoup utilisé il y a quelques décennies. Il a été progressivement remplacé par le criblage sur cibles afin d'améliorer les coûts et de réduire le temps nécessaire aux tests tout en augmentant le nombre de molécules testées. Ses avantages ont par ailleurs été récemment remis en avant par différents auteurs [14, 17, 18].

Une fois identifiés, les *hits* doivent être confirmés à l'aide de tests plus poussés, principalement afin de s'assurer que l'activité observée n'est pas imputable à des artefacts liés à la méthode expérimentale ou à la présence d'impuretés. La prochaine étape sera alors de transformer les *hits* les plus prometteurs en tête de série, plus communément appelée *lead*. Un *lead* est un composé ayant démontré une activité modérée ou importante et que l'on considère comme étant un point de départ acceptable pour la recherche d'un candidat médicament. Une fois l'activité effectivement confirmée, d'autres tests plus avancés seront conduits. Ces tests auront pour objectif d'évaluer plus précisément l'activité des molécules (sélectivité, capacité d'inhibition à faible concentration) mais également d'établir les caractéristiques physico-chimiques de celles-ci (solubilité, lipophilie, stabilité métabolique, ...) afin de sélectionner les molécules les plus prometteuses. La sélection de *leads* est une étape importante étant donné que la suite du projet se focalisera généralement sur les quelques molécules ainsi obtenues.

Une fois l'identification de *leads* à terme et en cas de succès, on obtient une ou plusieurs molécules qui seront amenées vers l'étape d'optimisation et qui représentent autant de pistes possibles pour l'obtention d'un candidat médicament.

Optimisation de *lead*

Durant cette étape, les chimistes médicaux démarreront un processus itératif durant lequel des modifications chimiques seront opérées autour des quelques molécules obtenues lors des criblages afin d'optimiser l'activité et les propriétés des futurs candidats médicaments. L'objectif de cette étape est d'obtenir un nombre limité de molécules ayant, d'une part, une activité importante (à nouveau le seuil d'activité dépend du projet et de l'objectif, l'idée étant de maximiser le ratio activité / concentration) et d'autre part, des propriétés physico-chimiques, biologiques et toxicologiques optimales.

Cette étape est, de ce fait, certainement la plus difficile, puisqu'elle nécessite d'optimiser en parallèle à la fois l'activité, mais également les autres propriétés qui feront de la molécule un médicament à la fois efficace et peu (ou pas) toxique. On parle généralement d'optimisation multi-objectifs et de nombreuses recherches sont menées dans cette direction, notamment en chémoinformatique [19–21].

1.2 Essais cliniques et commercialisation

Lorsqu'un médicament est considéré comme suffisamment actif, il peut entrer en phase de test sur l'homme. Les essais cliniques sont alors portés sur un nombre très limité de molécules issues de l'étape d'optimisation. C'est la phase la plus longue et la plus coûteuse du processus de recherche et c'est également celle où le taux d'échec est le plus élevé [5]. Elle dure entre 6 et 8 ans et a pour objectif de tester l'efficacité sur l'homme du ou des candidat(s) médicament. On distingue trois grandes étapes durant les essais cliniques. La première phase permet d'évaluer sur un petit nombre de volontaires (sains ou malades), la tolérance au médicament et l'absence d'effets secondaires notables. La deuxième phase s'effectue sur un nombre croissant de patients malades, afin de mieux estimer les effets secondaires, l'efficacité thérapeutique et les doses nécessaires pour obtenir l'effet désiré. La troisième phase est une étude comparative à très grande échelle sur plusieurs groupes de patients (plusieurs milliers), durant laquelle l'effet du médicament est comparé à un traitement de référence et / ou à un placebo.

Une fois les essais cliniques passés avec succès, l'ensemble des résultats obtenus sont soumis aux autorités de régulation afin d'officialiser le statut de médicament de la molécule et permettre ainsi sa commercialisation. Le processus de pharmacovigilance débute alors, durant lequel chaque médicament sera surveillé pour ses éventuels effets indésirables à court ou long termes.

2 Les chimiothèques au coeur du processus

La démocratisation des tests haut débit et le développement conjoint de méthodes de synthèse haut débit telle que la chimie combinatoire, ont fait naître la nécessité de gérer de grands ensembles de molécules : **les chimiothèques**. Ces dernières années ont en effet vu nos capacités de criblage augmenter de manière significative. Durant les années 80, période ayant vu l'apparition de l'automatisation des tests, des plaques contenant 96 puits étaient communément utilisées [5, 14]. Au fil du temps et des avancées en termes de miniaturisation, des plaques de 384 ou même 1536 puits sont progressivement apparues. De nos jours, les groupes disposant de moyens financiers importants peuvent tester de l'ordre de plusieurs centaines de milliers de molécules en une journée [13, 22], pour un coût se situant entre 0,5 et 1 dollar par molécule testée [5].

Une conséquence naturelle de l'apparition de tests haut débit est l'augmentation significative de la taille des chimiothèques propriétaires, comme l'illustre la figure I.2. Quand on ne trouvait au plus qu'une petite dizaine de milliers de molécules dans les années 90 pour une entreprise de taille moyenne, les plus grosses entreprises pharmaceutiques possèdent aujourd'hui un catalogue atteignant plusieurs millions de produits [11, 23]. Parallèlement, un véritable marché s'est développé autour de la commercialisation de molécules et l'on distingue aujourd'hui deux principaux types de catalogues proposés à la vente : les réactifs (ou "building blocks") et les molécules destinées au criblage. Les réactifs sont généralement de petits fragments utilisés pour la synthèse de molécules mais également pour la création de chimiothèques combinatoires ou virtuelles issues de l'assemblage de ces différents fragments. Le second type de molécule représente simplement des produits déjà synthétisés et prêts à être testés et sont souvent proposés sous différentes formes et quantités. Il existe ainsi aujourd'hui plusieurs centaines de fournisseurs commerciaux proposant des molécules à la vente et de nombreuses études ont été réalisées afin d'analyser et comparer le contenu des chimiothèques proposées [24–31]. Outre des catalogues "généraux" contenant jusqu'à plusieurs dizaines de millions de références, on trouve également des chimiothèques plus spécifiques, telles que des chimiothèques de produits naturels, de fragments, d'inhibiteurs de Kinases ou de canaux ioniques, etc.

Face à un choix toujours plus important, ces dernières années ont vu une grande attention se porter sur la composition des chimiothèques et sur le choix des molécules à tester. Le HTS reste en effet une méthode relativement récente. Le premier article utilisant l'acronyme HTS a en effet été publié en 1991, tandis qu'il a fallu attendre 1997 pour voir plus d'une dizaine d'articles publiés autour de cette méthode. On notera également que les premiers standards

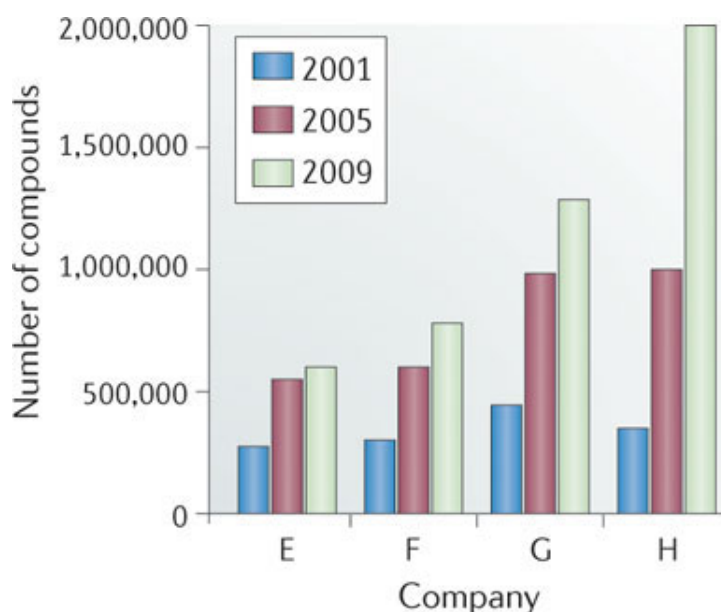


Figure I.2 – Évolution de la taille moyenne des chimiothèques propriétaires de quatre grandes entreprises pharmaceutiques entre 2001 et 2009. Figure extraite de [11]

industriels n'ont été publiés qu'en 1999 [11]. Comme n'importe quelle méthode expérimentale, le criblage haut débit comporte certaines limitations qui sont apparues au fur et à mesure de son utilisation. Quand il y a quelques années, l'enthousiasme de pouvoir tester plusieurs millions de molécules était à son paroxysme, on met aujourd'hui en avant la nécessité de privilégier la qualité des produits à tester plutôt que leur quantité.

2.1 Cribler plus pour gagner plus

L'idée derrière l'utilisation de tests haut débit était en effet assez simple et logique : tester plus de composés plus rapidement permettrait de générer plus de médicaments. Les avantages du HTS ont largement été mis en évidence depuis les années 2000, avec par exemple entre 50 et 84 % de projets de recherche de médicaments ayant été initiés suite à la découverte de nouvelles molécules actives lors de campagnes de criblage haut débit [11]. Malgré cela, le taux de succès constaté (en termes de nombre de hits par rapport au nombre de molécules testées) reste parfois relativement faible, généralement situé entre 0,1 et 2,5 % comparé aux ressources allouées à de tels tests [5, 32]. L'utilisation du HTS est de ce fait encore aujourd'hui questionnée et parfois critiquée pour trop souvent sacrifier la qualité au profit de la quantité des produits à tester.

Une étude rétrospective de Payne et al. [33] (GSK) illustre ainsi plutôt bien le fait que cribler plus n'implique pas forcément plus de succès. Durant une campagne de grande ampleur destinée à trouver de nouveaux agents anti-bactériens, 67 criblages ont été lancés sur une période de 7 ans, durant laquelle environ 400 000 molécules ont été testées. Sur ces 67 campagnes, seules 16 ont permis d'obtenir des *hits* réellement confirmés, tandis que plusieurs milliers de molécules apparemment actives n'étaient en fait que de faux positifs. Finalement, seule une molécule a pu atteindre l'étape d'optimisation de *lead*, un ratio plutôt faible si on le rapporte au coût total engendré par une telle étude.

Le problème des faux positifs est maintenant bien connu et certaines solutions ont été proposées pour en minimiser l'impact, comme nous le verrons plus loin. Mais il n'est pas le seul à rendre le HTS difficile à conduire et interpréter. Une étude de Novartis a par exemple montré que les molécules obtenues durant une campagne de criblage ne sont pas toujours les mêmes suivant la méthode expérimentale utilisée [34]. On trouve également d'autres effets de bord liés à la miniaturisation des puits [35], favorisant le transfert des composés volatiles d'un puits à un autre ou encore la fluorescence de certains composés rendant l'interprétation des résultats difficile à conduire [36].

Ces dernières années ont ainsi vu l'apparition de nombreuses méthodes ayant pour but d'optimiser le contenu des chimiothèques, à la fois de manière globale (minimiser le taux de faux positifs) mais également pour sélectionner les molécules de manière plus rationnelle afin de mieux cibler le problème étudié.

2.2 Optimiser la sélection : les faux positifs

Un faux positif, comme son nom le suggère, est un composé dont l'activité détectée lors du criblage est due à un phénomène différent de celui évalué par le test expérimental. Il existe trois grandes classes de molécules dont on sait qu'elles se transforment généralement en faux positifs durant les tests haut débit [37].

Les composés réactifs

Les composés réactifs sont des molécules contenant des groupements essentiellement électrophiles [37–39]. De part leur forte réactivité, ils forment généralement des liaisons covalentes avec la cible. A noter qu'il existe des médicaments réactifs, tels que l'aspirine.

Les composés *warheads*

Les composés *warheads* induisent des faux positifs en formant cette fois des liaisons le plus souvent réversibles non covalentes avec la cible [37]. On trouve parmi ce type de molécules, certains inhibiteurs suicides, des agents chélatants ou des composés polyioniques.

Les composés formant des agrégats

Ce type de composés, communément appelés "promiscuous aggregating inhibitors" (PAI), sont des inhibiteurs peu sélectifs et agissant de manière non compétitive. Ils forment des agrégats de taille variable (de 50 à 400 nm) qui se fixent en différents endroits de la protéine et faussent complètement la mesure [40, 41]. Ce mécanisme est un des plus étudiés car très fréquemment rencontré dans les campagnes de criblage. Deux campagnes de HTS menées par le NIH ont ainsi permis de montrer que près de 95 % des *Hits* observés lors d'un criblage avaient en fait formé des agrégats [42, 43]. Une observation similaire a été faite en 2008 lors d'une campagne de criblage sur 70 000 molécules extraites de la chimiothèque du NIH [44]. Ces études ont par ailleurs permis de confirmer que l'ajout de détergents peut largement diminuer le risque d'apparition de tels phénomènes, avec près de 90% de faux positifs supprimés.

Bien que ces phénomènes soient reconnus depuis plusieurs années maintenant, il reste toujours la difficulté de balancer le gain potentiel obtenu en termes de ressources allouées, avec la possibilité de perdre des composés potentiellement intéressants. On trouve en effet parmi les médicaments existants, des composés réactifs ou ayant été identifiés comme formant des agrégats [41]. Les études de Baell et Holloway [45] et de Jadhav et al. [42, 43] montrent également que l'apparition d'agrégats est parfois dépendante de la méthode et des conditions expérimentales utilisées. Certaines méthodes spécifiquement dédiées à leur détection peuvent même parfois donner des résultats divergents. Les spécialistes du domaine recommandent malgré tout la suppression de tels composés ou au moins de les marquer comme hautement suspects [37, 45].

2.3 Rationaliser la sélection

La suppression des faux positifs permet de diminuer le nombre de tests inutiles et ainsi de faire de substantielles économies. Il existe également d'autres méthodes permettant de sélectionner plus finement les composés à tester afin de maximiser les chances de succès par rapport aux objectifs du projet.

2.3.1 Composés "drug-like"

Idéalement, on souhaiterait ne tester que les molécules ayant un fort potentiel à devenir un médicament : absence de toxicité, grande efficacité thérapeutique, bonne absorption pour les médicaments destinés à la prescription par voie orale, etc. En l'absence de définition universelle ou de méthode prédictive parfaite, de nombreuses études ont tenté de faire la différence entre une molécule biologiquement active et une molécule "drug-like" se rapprochant du médicament idéal [46–50].

La définition la plus célèbre d'un composé "drug-like" est celle de Lipinski [51]. A partir de composés administrés par voie orale ayant passé avec succès la phase 2 des tests cliniques, il observe que les molécules ayant le plus de chances d'être absorbées par voie orale satisfont au moins trois des caractéristiques suivantes : (1) un poids moléculaire inférieur à 500 Dalton, (2) un LogP calculé inférieur à 5, (3) un nombre d'accepteurs de liaison hydrogène inférieur ou égal à 10, (4) un nombre de donneurs de liaison hydrogène inférieur ou égal à 5. Avec un objectif similaire, Veber et al. [52] étendent ces règles en ajoutant deux contraintes sur la surface polaire (inférieure à 140 Å²) et le nombre de liaisons pouvant tourner (inférieur ou égal à 10).

2.3.2 Composés "lead-like"

Le terme "drug-like", bien que souvent associé aux règles de Lipinski, est malgré tout assez flou et a donc été maintes fois décliné afin de désigner des classes de molécules plus spécifiques. De nombreuses études ont par exemple été menées dans le but de distinguer les *leads* des médicaments et autres molécules. Une fois encore, la définition d'un *lead* reste globalement suggestive et de nombreuses études ont essayé d'établir des règles analogues à celles énoncées par Lipinski. L'idée part du constat que les molécules entrant en phase clinique sont souvent plus complexes et plus grandes que les *leads* à partir desquels elles ont été dérivées [53–56]. Ainsi, des filtres plus restrictifs permettant d'obtenir des composés plus simples ont été créés, le plus cité étant celui de Hann et Oprea [55] :

- Masse moléculaire $< 460Da$;
- $-4 < \text{Log P} < 4,2$;
- $\text{Log } S_w \geq -5$;
- Nombre de liaisons pouvant tourner < 10 ;
- Nombre de cycles < 4 ;
- Nombre de donneurs de liaisons H < 5 ;

- Nombre d'accepteurs de liaisons H < 9 ;

De manière assez similaire, Congreve et al. définissent une règle analogue à celle de Lipinski mais adaptée à la sélection de molécules pour du criblage sur fragments [57]. Ici, l'idée est également de sélectionner des molécules plus petites et plus simples :

- Masse moléculaire $< 300Da$;
- Log P < 3 ;
- Nombre de donneurs de liaisons H < 3 ;
- Nombre d'accepteurs de liaisons H < 3 ;

Sans l'inclure explicitement à la règle originale énoncée dans l'article, les auteurs suggèrent également une extension de cette règle en ajoutant deux critères supplémentaires : nombre de liaisons pouvant tourner < 3 et aire de surface polaire $< 60 \text{ \AA}^2$.

Récemment, de nouveaux critères ont également été introduits afin de caractériser les molécules pouvant servir de modulateurs des interactions protéine-protéine. Ainsi, à partir d'un ensemble de 39 inhibiteurs d'interactions protéine-protéine, Morelli et al. [58] définissent une nouvelle règle de 4 permettant de définir le profil générique de tels inhibiteurs :

- Masse moléculaire $> 400Da$;
- ALog P > 4 ;
- Nombre de cycles > 4 ;
- Nombre d'accepteurs de liaisons H > 4 ;

D'autres critères ont également été introduits plus récemment. Le concept de "Ligand Efficiency" décrit par Hopkins et al. [59] suggère ainsi que le ratio entre l'énergie de liaison et le nombre d'atomes lourds est une mesure efficace pour sélectionner les *leads*. Elle permet en effet de distinguer une molécule active et très complexe, d'une molécule ayant une activité équivalente mais pour une complexité moindre, permettant ainsi de sélectionner des *leads* ayant un plus grand potentiel à être modifiés par la suite. D'autres indices de ce type ont été suggérés par d'autres groupes, en prenant cette fois en compte l'aire de la surface topologique polaire ou la lipophilicité [60–64].

2.3.3 Chimiothèques focalisées

Lorsque l'on dispose d'informations concernant l'existence de molécules actives sur une cible d'intérêt, des chimiothèques spécifiquement dédiées au criblage sur cette cible (ou sur une certaine famille de cibles, comme les Kinases) sont en général utilisées. Cette idée repose sur le principe de similarité, qui énonce que deux molécules similaires auront de fortes chances d'avoir des propriétés biologiques similaires [65–70]. En incluant dans une chimiothèque un certain nombre de molécules similaires aux actifs connus, on augmente théoriquement les chances de trouver de nouvelles molécules actives, avec pour effet de bord l'exploration d'une zone restreinte de l'espace chimique. La chémoinformatique et la modélisation moléculaire jouent alors un rôle important dans ce processus de sélection à travers l'utilisation de méthodes de recherche par similarité [69, 71, 72], d'amarrage moléculaire ou de QSAR.

2.3.4 Chimiothèques diverses

L'une des méthodes ayant été beaucoup mise en oeuvre ces dernières décennies est la sélection de molécules sur des critères de diversité. L'idée est à nouveau d'introduire plus de rationalité en sélectionnant les molécules, non pas au hasard, mais de telle sorte qu'elles maximisent la couverture d'un espace chimique connu, tout en minimisant la redondance. De nombreuses études ont pu ainsi montrer qu'utiliser une chimiothèque diverse permettait d'augmenter le taux de hits lors de campagnes de criblage [73–80], permettant ainsi de maximiser les chances de trouver des molécules actives tout en optimisant les ressources allouées. La diversité reste par ailleurs une notion subjective, qui dépend notamment de la manière dont sont décrites les molécules, ainsi que de la métrique utilisée pour les comparer. Nous l'introduirons, ainsi que les moyens de la quantifier et de créer des chimiothèques diverses, plus en détails dans les sections consacrées à la chémoinformatique.

2.3.5 L'oeil du chimiste

L'ensemble des méthodes décrites sont issues de l'expérience accumulée au fil des ans au sein des équipes de criblage et on a vu qu'il existe de nombreuses études disponibles dans la littérature sur les bonnes pratiques à appliquer lors de la sélection des composés à cribler. Un facteur parfois omis est l'expérience même des chimistes médicaux. Après tout, ce sont ces derniers qui, suite au criblage, vont travailler autour des molécules sélectionnées. Chaque chimiste ayant sa propre expérience, voire même sa propre "chimie", il sera lui même capable de reconnaître du premier coup d'oeil un composé difficile à synthétiser ou modifier, ou même un composé qui ne possède pas les propriétés physicochimiques ou ADME pour en faire un bon

médicament. Ces critères de sélection, parfois subjectifs, sont néanmoins difficiles à mettre en oeuvre durant la phase amont de recherche. Ils sont en général appliqués lors de la sélection de *Leads*.

2.4 Chimiothèques et HTS dans le milieu académique

Lorsque l'on parle de tester plusieurs millions de molécules, il va de soit que les coûts associés à de telles campagnes sont généralement prohibitifs pour les laboratoires académiques. Pourtant, au cours des dernières années, de nombreux efforts ont vu le jour afin de démocratiser le HTS et le rendre accessible aux laboratoires académiques [81–85]. Harvard fut l'une des premières institutions à tenter de mettre en place le HTS au sein d'un laboratoire académique et à analyser les difficultés inhérentes à un tel projet [81]. On recense aujourd'hui plus de 90 centres académiques de criblage [86] qui possèdent généralement leurs propres chimiothèques. Par exemple, l'université Rockefeller (<http://www.rockefeller.edu/high-throughput/highthroughput.php>) et du Michigan (<http://mhtsc.kvcc.edu/>) possèdent des chimiothèques contenant une petite centaine de milliers de molécules achetées à des fournisseurs tels que ChemDiv, Maybridge ou ChemBridge [87].

En France, la chimiothèque nationale fut au début des années 2000 l'une des premières initiatives portées à l'échelle d'un pays. Elle a pour but de fédérer et valoriser des produits issus de la recherche académique française. Aujourd'hui, la chimiothèque nationale contient plus de 50 000 produits. Depuis, d'autres initiatives de ce type ont vu le jour, par exemple en Australie [88] ou au Japon. Avec un budget annuel tournant autour de 100 millions de dollars (à titre de comparaison celui de la chimiothèque nationale était de 40 000 euros en 2009 [84]), le NIH mène depuis 2004 l'initiative la plus ambitieuse permettant de démocratiser le HTS [89, 90]. Les tests biologiques effectués dans le cadre de cette initiative sont mis à disposition de tous dans la base de données PubChem [91]. Doté de moyens très importants et étant donnés les coûts liés aux HTS et à la maintenance des infrastructures nécessaires, de nombreuses voix se sont élevées pour critiquer cette initiative, notamment parce que les molécules actives issues de telles campagnes de criblage ne peuvent pas être optimisées facilement par des laboratoires académiques qui ne disposent généralement pas de moyens suffisants pour mener à terme des projets d'optimisation de médicaments. Certains auteurs clament néanmoins qu'il faudra encore attendre quelques années afin de constater les effets du projet, étant donnée la durée nécessaire à la mise sur le marché d'un médicament. Quelques exemples sont par ailleurs fréquemment cités afin d'illustrer les opportunités qu'offre une telle plateforme de criblage pour les labo-

ratoires académiques. David Williams, un parasitologiste ayant passé une grande partie de sa carrière à étudier la *Schistosoma* (un ver qui tue environ 300 000 personnes par an dans les pays tropicaux), trouva en 2005 une enzyme nécessaire à la survie du ver et se posa alors la question de savoir comment trouver une molécule pouvant bloquer son activité. Il se tourna alors vers le NIH qui propose aux chercheurs de soumettre leurs propres enzymes afin de tester les capacités d'inhibition sur des molécules provenant de leur chimiothèque. 71 000 composés ont alors été testés et une molécule a ainsi pu être identifiée comme candidate, tuant environ 90 % des vers lors de tests sur des souris infectées [92, 93].

2.5 Conclusion

On a vu que les chimiothèques étaient au coeur du processus de génération de candidats médicaments. Avec le temps et des capacités de criblage toujours plus grandes, la taille des chimiothèques et le nombre de molécules disponibles ont augmenté de manière significative. La gestion et l'analyse d'une telle quantité de données nécessite l'aide de systèmes informatiques capables de mettre en oeuvre les différentes stratégies d'analyse et de sélection de composés. Dans la section suivante, nous introduirons le domaine de la chimoinformatique et nous discuterons des méthodes permettant de répondre aux problématiques soulevées dans la section précédente, mettant ainsi en perspective le travail de cette thèse.

3 L'apport de la chimoinformatique

3.1 Généralités

La chimoinformatique fait partie des domaines interdisciplinaires ayant vu le jour en tant que tel durant les deux dernières décennies. Bien que le problème du traitement informatique de l'information chimique ait déjà été abordé dès les années 70 dans un ouvrage consacré au sujet [94], le terme "chimoinformatique" n'est lui même apparu que dans les années 90. Les premiers ouvrages spécifiquement consacrés au domaine (contenant le mot 'chimoinformatique' dans leur titre) sont apparus en 2003 par Gillet et Leach [95] et Gasteiger (éditeur) [96]. La discipline a depuis beaucoup évolué et de nombreux ouvrages y ont été consacrés [97–102], ainsi que d'excellents articles de revues [103–109]. Un journal a même récemment été créé (2007), le *Journal of Cheminformatics* (notez l'orthographe qui fait toujours débat aujourd'hui).

Deux moteurs ayant poussé le domaine à évoluer de manière significative peuvent certaine-

ment être mis en évidence :

- Le HTS et l'augmentation significative de la quantité de données liées à la chimie. L'apparition de méthodes expérimentales haut débit a nécessité de repenser les systèmes informatiques de gestion de chimiothèques afin de permettre un traitement et une exploitation plus efficace des données.
- La prédiction de propriétés à partir de la structure des molécules. Bien que des travaux aient été réalisés dans ce sens dès la fin du *XIX^{ème}* siècle [110], il est généralement reconnu que l'étude de Hansh & Fujita [111], qui établirent une relation quantitative entre la structure et certaines propriétés des molécules, à travers l'utilisation de méthodes mathématiques, donna naissance à l'un des sous-domaines les plus actifs de la chémoinformatique : le QSAR [112] (Quantitative Structure Activity Relationship).

La chémoinformatique étant un domaine récent à l'interface de plusieurs autres domaines, ses frontières sont étendues et difficiles à tracer. Bien qu'il n'existe pas de définition officielle faisant l'unanimité au sein de la communauté, plusieurs auteurs ont tenté au fil des ans d'en tracer les limites. La première définition du terme fut publiée en 1998, où F. Brown [113] propose de regrouper sous le terme chémoinformatique, "l'ensemble des techniques permettant de transformer les données chimiques en information afin d'en extraire des connaissances pour rendre les décisions meilleures et plus rapides dans le domaine de la recherche et de l'optimisation de nouveaux médicaments". Quelques années plus tard, Johann Gasteiger définit plus simplement la chémoinformatique comme "l'application des méthodes informatiques pour la résolution de problèmes chimiques" [96]. Greg Paris proposa une définition tout aussi vaste : "La chémoinformatique est un terme générique qui couvre le design, la création, l'organisation, la gestion, la récupération, la dissémination, la visualisation et l'utilisation de l'information chimique" [114]. Le problème du traitement informatique de l'information chimique ne datant pas d'hier, Hann & Green [115] ont eux titré un de leurs articles de revue par "Chemoinformatics : new name for an old problem?". Cette définition, qui n'en est pas vraiment une, reflète cependant plutôt bien l'évolution du domaine. Enfin, une définition plus précise du domaine a été énoncée lors du premier colloque de chémoinformatique organisé à Obernai en 2006 [116] :

Chemoinformatics is a scientific discipline that has evolved in the last 40 years at the interface between chemistry and computer science. It has been realized that in many areas of chemistry, the huge amount of data and information produced by chemical research can only be processed and analyzed by computer methods. Furthermore, many of the problems faced in chemistry are so complex that novel approaches utilising solutions that are based on informatics methods

are needed. Thus, methods were developed for building databases on chemical compounds and reactions, for the prediction of physical, chemical and biological properties of compounds and materials, for drug design, for structure elucidation, for the prediction of chemical reactions and for the design of organic syntheses. Research and development in chemoinformatics is essential, for increasing our understanding of chemical phenomena, for industry to remain competitive in a global economy. Chemoinformatics methods can be applied in any field of chemistry, from analytical chemistry to organic chemistry. It is of particular importance in drug design and development.

On présentera dans cette section les concepts les plus importants de la chimoinformatique, en notant que cette présentation sera forcément biaisée par le domaine d'application de cette thèse. Nous n'aborderons ici que les méthodes apportant des solutions concrètes quant à la gestion et à l'analyse de chimiothèques.

3.2 Représentation de l'information chimique

Le traitement informatique de données chimiques nécessite des moyens pour représenter les molécules sous une forme pouvant être comprise et exploitée par un ordinateur. Il existe aujourd'hui plusieurs moyens de représenter une molécule et ceux-ci diffèrent essentiellement par la quantité d'information qu'ils encodent, ainsi que par le contexte dans lequel ils sont destinés à être utilisés. Nous présenterons ici les principaux moyens de représentation existants, afin notamment de mettre en perspective leur utilisation dans le logiciel Screening Assistant 2 (chapitre IV).

3.2.1 Représentations visuelles

Que ce soit pour la définir ou la visualiser, la manière la plus simple de représenter une molécule est à travers une représentation graphique de celle-ci, le plus souvent sous forme d'image (interactive pour le dessin de structures). Ce mode de représentation est bien entendu le plus intuitif pour tout un chacun et permet d'afficher la plupart des informations identifiant la molécule (composition atomique, liaisons, charges, stéréochimie...). L'image reste néanmoins trop peu spécifique pour être réellement interprétée par un ordinateur, une image numérique n'étant qu'un ensemble de pixels auxquels est associée une couleur. Il est intéressant de noter que des recherches sont consacrées à l'interprétation des images de structure moléculaire afin de les transformer en d'autres formats plus directement exploitables par l'ordinateur, avec pour application la fouille de données dans les bases d'articles scientifiques [117].

3.2.2 Représentations textuelles

Identifiants

Les molécules sont très souvent associées à un ou plusieurs noms, généralement courts, afin de les identifier plus rapidement. Ces noms peuvent être liés à l'usage courant de la molécule (une molécule dans un projet de recherche), à ses propriétés biologiques (le nom des acides aminés ou d'un médicament), ou encore correspondre à un identifiant dans une base de données publique (PubChem [91], ChEMBL [118], etc.). Bien qu'aucune information sur la molécule elle-même ne soit encodée, les identifiants présentent l'avantage de pouvoir être utilisés afin de retrouver très rapidement une ou plusieurs molécules, les recherches textuelles étant très performantes dans les systèmes d'informations modernes (et même anciens!).

Nomenclatures

Les systèmes de nomenclature permettent d'identifier une molécule de manière plus précise et souvent unique. Le système de nomenclature officiel en chimie est la nomenclature IUPAC (International Union of Pure and Applied Chemistry). La molécule est ici décrite à l'aide d'un ensemble de symboles et de règles qui permettent d'encoder la topologie de celle-ci. Ces nomenclatures sont plutôt destinées aux chimistes, car malgré leur complexité, elles permettent à un lecteur averti de facilement se représenter la molécule à travers l'utilisation d'un dictionnaire de termes permettant d'identifier les groupes fonctionnels (au lieu des atomes comme nous le verrons plus loin). Cette nomenclature est néanmoins peu utilisée en chémoinformatique car difficile à encoder et décoder et relativement peu compacte.

Notations compactes

La notation SMILES (Simplified Molecular Input Line Entry Specification), introduite en 1988 par David Weininger [119], est la notation la plus connue et la plus utilisée en chémoinformatique. La plupart des logiciels de modélisation supportent cette notation et l'étendent même parfois avec de nouvelles règles. Elle permet de générer une représentation compacte des molécules par un enchaînement de symboles représentant les atomes (et éventuellement leurs propriétés) et leur topologie. Elle était au départ principalement utilisée pour obtenir une représentation compacte et simple des molécules et ainsi faciliter leur stockage et leur échange. Des versions canoniques ont par la suite été créées afin de rendre unique l'identifiant et ainsi permettre de détecter les doublons.

Le programme InChI [120–122] est un logiciel libre permettant de générer un identifiant du

même nom, (Inchi - IUPAC International Chemical Identifier) propre à chaque molécule. Contrairement au SMILES, il est un peu moins compact et surtout plus difficile à interpréter sous sa forme textuelle. Il est néanmoins devenu un standard pour la génération de code unique, car il supporte un grand nombre de spécificités telles que la tautomérie, la stéréochimie, etc. Notons enfin qu'une version plus compacte de l'InChI est également disponible : l'InChI Key, qui se trouve être la version hachée (c'est à dire transformée par un algorithme numérique dit de hachage permettant d'obtenir une nouvelle représentation unique et de taille fixe) du code original.

On note enfin l'existence de la notation SMARTS [123], une notation dérivée du SMILES qui permet cette fois d'encoder les molécules sous la forme de motifs. C'est un langage utilisé non pas pour la définition de structures, mais pour la définition de requêtes utilisées pour la recherche par sous-structure. De manière analogue au SMILES, chaque élément de la chaîne de caractères représente soit un atome, soit une liaison. Cependant, la notation SMARTS introduit de nouvelles règles permettant d'utiliser des opérateurs logiques afin qu'à une position donnée dans le graphe de la molécule, on puisse définir plusieurs types d'atomes ou de liaisons autorisés. On peut ainsi utiliser la requête suivante pour récupérer toute molécule contenant un atome d'azote lié à un atome aromatique contenu dans un cycle à 6 : $N[a;r6]$. Ici, l'atome lié à l'atome d'azote, encadré par des crochets, doit être à la fois aromatique (symbole a) ET (opérateur encodé par le symbole $;$) contenu dans un cycle à 6 (symbole $r6$).

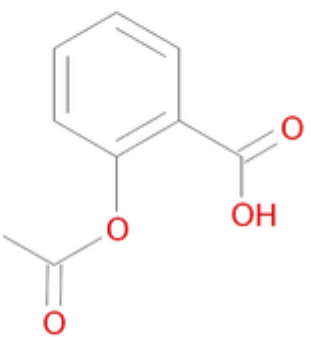
Tables de connectivité

Ces formats sont en général beaucoup moins compacts mais plus exhaustifs sur l'information portée. Parmi les plus utilisés, le format MOL [124, 125] est un format propriétaire créé par la société MDL. Ce format est le plus commun en chimoinformatique. Il permet de stocker un grand nombre d'informations sur la molécule décrite, (atomes et leur hybridation, table de connectivité avec les propriétés des liaisons, coordonnées atomiques, etc.) et son extension (le format SDF [124, 125]) permet elle de stocker plusieurs molécules dans un même fichier, auxquelles peuvent être associées des propriétés.

On notera enfin qu'une grande majorité des représentations présentées ici sont utilisées dans les bases de données publiques, comme l'illustre la fiche produit de l'aspirine extraite de la base de données ChEMBL I.3.

Compound Report Card

Compound Name and Classification

Compound ID	CHEMBL25	 <p>CHEMBL25</p>
Compound Name	Aspirin	
Synonyms	Acetylsalicylic Acid, Aspirin	
Approved Drug	Yes	
Trade Names	Ecotrin, Measurin, Equi-Prin, 8-hour bayer, Acetosalic Acid, Acetylsalicylic Acid, Salicylic Acid Acetate, Bayer extra strength aspirin for migraine pain	

Compound Representations

Molfile	Download MolFile
Canonical SMILES	<chem>CC(=O)Oc1ccccc1C(=O)O</chem>
Standard InChI	InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(... Download InChI
Standard InChI Key	BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Figure I.3 – Différentes représentations de l'aspirine disponibles dans la base de données ChEMBL [118]

3.2.3 Représentation en mémoire vive

Les représentations textuelles décrites permettent le stockage de l'information chimique dans des fichiers standards, facilitant ainsi leur échange. Un programme informatique permettant de traiter les molécules nécessite néanmoins une représentation dynamique des molécules afin de pouvoir réaliser toute sorte d'opérations : construction de molécules à partir d'une représentation textuelle, calcul de descripteurs, recherche de sous-structure etc. La plupart (si ce n'est l'ensemble) des bibliothèques de programmation pouvant traiter les molécules représentent celles-ci à l'aide d'un graphe non-dirigé. Les noeuds du graphe représentent les atomes et les arêtes représentent les liaisons covalentes entre les atomes. Les propriétés physico-chimiques des éléments composant la molécule (type atomique et hybridation, ordre des liaisons, stéréochimie, etc.) sont alors encodées en tant que propriétés des objets modélisant les différents constituants. L'avantage d'une telle représentation, outre le fait qu'elle correspond bien au modèle (simplifié) que chacun se fait d'une molécule, est qu'elle permet l'exploitation de l'ensemble

des méthodes utilisées par la théorie des graphes : recherche de plus court chemin (permettant d'obtenir la distance topologique entre deux atomes), isomorphisme de graphes (recherche de sous-structures), recherche de plus petite sous-structure commune, etc. A partir d'une telle représentation, d'autres moyens de description, très utilisés en chimoinformatique, ont été développés : les descripteurs moléculaires.

3.3 Descripteurs moléculaires

Un descripteur est une valeur numérique ou textuelle résultant d'une opération réalisée à partir d'une certaine représentation de la molécule à décrire. Les descripteurs peuvent être regroupés suivant la manière dont ils sont encodés (représentation textuelle, numérique ou vectorielle), suivant le type d'information qu'ils portent (descripteur physico-chimique, topologique, pharmacophorique, etc.), ou suivant la dimensionalité de la représentation de la molécule à partir de laquelle ils ont été calculés (1D, 2D, 3D). On en dénombre plusieurs milliers et Todeschini tente d'en faire un inventaire exhaustif dans un livre faisant référence dans le domaine [126].

3.3.1 Descripteurs 1D

Ces descripteurs encodent une information sous la forme d'un entier ou d'un nombre réel en se basant sur une propriété physico-chimique globale de la molécule, par exemple le poids moléculaire ou la composition atomique. Malgré leur simplicité, ils sont souvent utilisés afin d'obtenir une première description "grossière" des chimiothèques car faciles à interpréter et rapides à calculer.

3.3.2 Descripteurs 2D

Ils sont eux généralement basés sur la topologie de la molécule. Parmi eux, on trouve tout un ensemble de descripteurs capturant diverses informations calculées à partir du graphe de connectivité de la molécule. Les plus connus sont les indices de Kier & Hall [127, 128], Randic [129] ou Weiner [130]. De par leur nature, ils permettent de distinguer les molécules plus finement et notamment des molécules cycliques, linéaires, ou parfois même chirales [131]. Également rapides à calculer, ils sont aussi beaucoup utilisés pour décrire et analyser des chimiothèques, ainsi que pour des études de diversité ou de QSAR.

3.3.3 Descripteurs 3D

Les descripteurs 3D se basent eux sur les coordonnées tri-dimensionnelles de la molécule. Ce sont généralement des valeurs numériques décrivant les propriétés (ou leur répartition) de la surface moléculaire, le positionnement relatif de points pharmacophoriques, et parfois les configurations électroniques (descripteurs issus de la chimie quantique). L'avantage des descripteurs 3D est qu'ils encodent une information plus importante à travers la prise en compte de la configuration tri-dimensionnelle de la molécule. L'inconvénient est le temps de calcul généralement important et surtout la nécessité de sélectionner une conformation particulière de la molécule pour réaliser le calcul. Pour les molécules flexibles en particulier, la pertinence d'une conformation par rapport à une autre est toujours difficile à évaluer a priori, d'autant plus du point de vue de l'activité biologique, et l'on se retrouve souvent à devoir explorer l'espace conformationnel de la molécule afin de ne pas se limiter à une seule structure, ce qui multiplie d'autant le temps de calcul. Pour ces raisons, ces descripteurs sont rarement utilisés dans les études de diversité où la quantité de molécules à traiter est importante et où l'information sur la cible biologique n'est pas toujours disponible ni même pertinente. Ils sont plutôt utilisés lors de la phase d'optimisation, où l'on cherche à obtenir une description moléculaire plus fine et notamment pour les interactions protéine-ligand, afin de mieux comprendre et optimiser les interactions moléculaires responsables d'une activité biologique étudiée.

3.3.4 Descripteurs de descripteurs

On notera que de nombreux descripteurs sont eux même dérivés d'autres descripteurs, bien qu'ils puissent également être classés dans l'une des catégories précédemment citées. La plupart des descripteurs quantifiant la lipophilicité (LogP) sont ainsi générés à l'aide de modèles QSAR se basant sur d'autres descripteurs. Les BCUT [132] sont issus du calcul de valeurs propres à partir de la matrice de connectivité dans laquelle sont encodées des propriétés physico-chimiques au niveau de la diagonale de la matrice. Suite à cette opération, on obtient un ensemble de descripteurs décorrélés, qui ont notamment été utilisés dans de nombreuses études de diversité [132, 133].

3.3.5 Empreintes moléculaires

Les empreintes moléculaires (ou fingerprints) sont une classe de descripteurs très utilisée en chémoinformatique et notamment pour la recherche par similarité et pour les études de diversité. Un fingerprint est un ensemble fini et non ordonné de descripteurs, généralement représenté sous la forme d'un vecteur de bits (les éléments du vecteur prenant leurs valeurs

dans 0,1). Chaque élément du vecteur représente une propriété, typiquement la présence ou non d'un certain motif (fragment, atome, liaison, etc.). Il existe deux grandes classes de fingerprints, bien que ceux-ci pourraient également être classés suivant le type de représentation moléculaire sur laquelle ils se basent (2D, 3D) : (1) les fingerprints basés sur un ensemble de motifs fixes préalablement établis et (2) les fingerprints hachés, dont le nombre d'éléments du vecteur n'est pas prédéterminé.

Parmi les fingerprints basés sur un ensemble de motifs, on retrouve le fingerprint MACCS [134], dont la définition repose sur un ensemble de 166 SMARTS décrivant des groupes fonctionnels et systèmes cycliques classiquement utilisés en chimie organique. L'inconvénient de ce type de fingerprint est qu'ils encodent généralement un nombre limité de motifs, qui ne sont pas nécessairement pertinents par rapport aux molécules traitées ou à la cible biologique étudiée. Afin de pallier ce problème, des dictionnaires sont généralement générés à partir de molécules d'intérêt afin de s'assurer de la pertinence des motifs utilisés. Récemment, un nouveau dictionnaire de SMARTS a par exemple été établi en se basant sur la fréquence d'apparition de certains fragments dans des molécules actives [135], donnant naissance à un fingerprint de plus de 4800 éléments.

Les fingerprints hachés sont eux, au contraire, spécifiques des molécules à décrire. En se basant à nouveau sur le graphe moléculaire, ces fingerprints génèrent, de manière généralement exhaustive, une liste de fragments en utilisant des règles spécifiques de l'algorithme utilisé. Puisque le nombre total de fragments pouvant potentiellement être générés par un tel procédé est très important et surtout ne peut pas être connu à l'avance, ces fingerprints génèrent pour chacun des fragments, un index unique calculé à l'aide d'un algorithme de hachage utilisant une représentation canonique de chaque fragment. Ce type de fingerprint se présente donc à la base sous la forme d'un ensemble d'indices qui représentent une position dans un fingerprint de taille virtuellement infinie. Afin de les transformer en chaîne de bits de taille fixe, on utilise généralement l'opérateur modulo, qui permet ainsi de réduire la quantité d'information, avec pour inconvénient l'augmentation du risque de collisions et donc la perte d'information. Parmi les plus utilisés, on pourra citer les fingerprints ECFP et FCFP [136], MOLPRINT2D [137–139] ou encore tout un ensemble de fingerprints encodant les paires, triplets ou quadruplets pharmacophoriques présents dans les molécules [131, 140–142].

On notera enfin l'existence de fingerprints que l'on pourrait qualifier d'"hybrides", comme par exemple les Triplets Pharmacophoriques Flous [143, 144]. Ce fingerprint se base sur un dic-

tionnaire issu de l'énumération exhaustive de triplets pharmacophoriques de référence pour un ensemble de distances topologiques donné (par exemple, tous les triplets dont les distances inter-atomiques vont de 2 à 10 par incrément de 2). L'originalité de ce fingerprint est qu'au lieu de se contenter de détecter la présence ou l'absence d'un triplet de référence au sein de chaque molécule, il permet une détection dite "floue" : la correspondance entre un triplet dans une molécule et un triplet de référence sera quantifiée à l'aide d'un alignement 2D des triangles représentant les triplets pharmacophoriques. La méthode se basant sur une description 2D des molécules, cet alignement flou permet d'une certaine manière de mimer les variations conformationnelles des structures. Il encode également non pas la seule présence, mais le nombre de triplets de référence retrouvé pour chaque molécule, pondéré par le degré de recouvrement à ces mêmes triplets de référence. Il considère enfin non pas une seule représentation pour chaque molécule, mais toutes les micro-espèces présentes à l'équilibre protéolytique pour un pH donné (et leurs concentrations relatives). Le fingerprint sera finalement généré en calculant une moyenne pondérée sur l'ensemble des fingerprints obtenus pour chaque micro-espèce.

3.3.6 Descripteurs basés sur les fragments

Les descripteurs fragmentaux [145–147] (disponibles dans le programme ISIDA [148]) sont un autre type de descripteurs permettant de décrire de manière précise les différents motifs présents dans chaque molécule. Il en existe deux types :

- **Les séquences** correspondent à la présence d'un enchainement d'atomes, de liaisons, ou d'atomes et de liaisons. Les séquences d'atomes correspondent à un ensemble d'atomes de taille N reliés par un chemin au sein de la molécule. Les séquences de liaisons correspondent à la même chose mais pour les liaisons (le type de liaison étant utilisé pour encoder chacune d'entre elle). Les séquences d'atomes et de liaisons prennent en considération les deux types d'éléments et peuvent ainsi être interprétées comme de véritables fragments à part entière. Pour chacun des trois types de descripteurs, les nombres minimum et maximum d'atomes peuvent être définis lors de la génération des descripteurs.
- **Les atomes unis** encodent quant à eux les atomes et leur environnement proche (au sens topologique). De manière analogue aux séquences, trois sous-types de descripteurs existent, suivant l'information qui est encodée au voisinage : atomes, liaisons, ou les deux. Ces descripteurs présentent l'avantage d'encoder de manière précise les motifs présents dans chaque molécule et d'être facilement interprétables, une caractéristique essentielle pour l'utilisation de descripteurs dans le cadre d'études QSAR.

3.3.7 Schémas de simplification du graphe moléculaire

L'ensemble des méthodes de description de molécules ont toutes un but commun : la transformation d'une certaine représentation d'une molécule en une autre. Nous avons jusqu'ici décrit des méthodes permettant d'encoder les structures sous la forme de valeurs numériques mais il convient également de mentionner d'autres méthodes permettant d'en obtenir une représentation simplifiée.

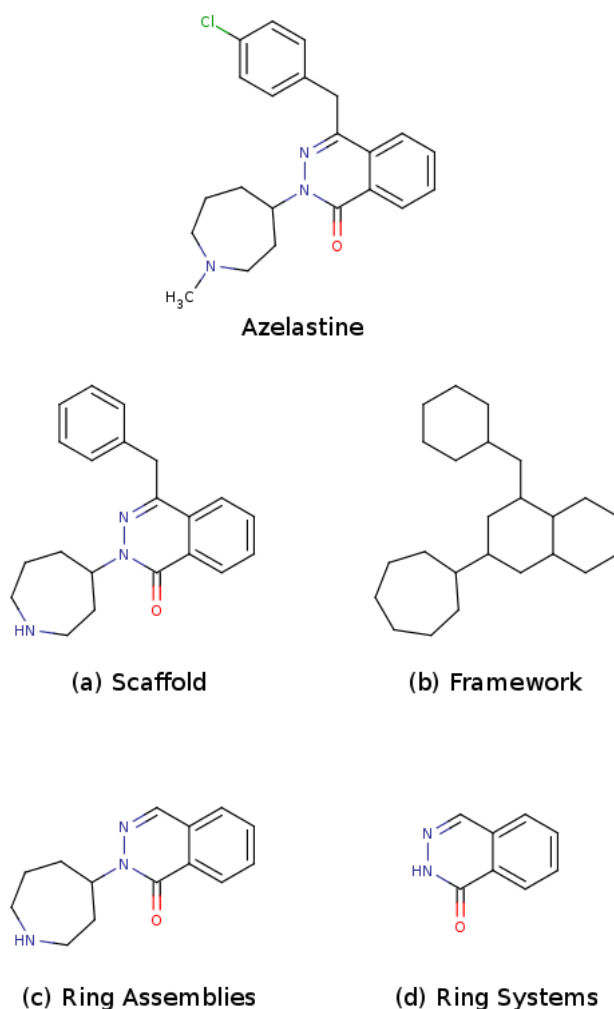


Figure I.4 – Les différentes règles de réduction de molécules pour en générer le squelette présentées pour la molécule d'Azelastine (antihistaminique) : (a) le scaffold, tel que défini par Bemis et Murcko, contenant l'ensemble des systèmes cycliques sans les chaînes latérales ; (b) le framework, équivalent du scaffold avec les atomes transformés en carbone et les liaisons en liaisons simples ; (c) les assemblages de cycles, comme étant l'ensemble des cycles fusionnés ou connectés par une seule liaison ; (d) les systèmes cycliques, comme étant le plus grand sous-ensemble de cycles fusionnés. Notons que les hydrogènes sont systématiquement supprimés avant l'application du schéma de réduction et qu'à l'exception du framework les doubles liaisons exo-cycliques sont systématiquement conservées.

Les squelettes moléculaires

La notion de squelette moléculaire (scaffold) est essentielle en chimie médicinale et en chémoinformatique. Ce type de représentation est en effet beaucoup utilisé par les chimistes médicaux et c'est à travers celle-ci que sont généralement regroupées les molécules, que ce soit pour identifier des sous-structures actives ou plus simplement pour regrouper les molécules en séries lors des étapes de sélection et d'optimisation de *leads*. Le scaffold d'une molécule est représenté par un sous-ensemble connecté d'atomes et de liaisons qui forment le "cœur" de celle-ci. C'est une notion qui reste assez subjective, bien qu'il existe des règles établies qui sont beaucoup utilisées aujourd'hui. La plus connue est celle proposée par de Bemis et Murcko [149], qui définissent deux types de règles permettant de simplifier le graphe moléculaire : (1) le scaffold qui est constitué de l'ensemble des systèmes cycliques dont on aura retiré les chaînes latérales (à l'exception des doubles liaisons exocycliques) et (2) le framework qui n'est autre que le scaffold dans lequel l'ensemble des atomes auront été remplacés par des carbones et l'ensemble des liaisons transformées en liaisons simples. D'autres règles existent pour définir les scaffolds et la figure I.4 en présente les plus connues.

Les graphes réduits

Les graphes réduits [150, 151] représentent un autre moyen de simplifier le graphe d'une molécule. L'objectif de cette méthode est de regrouper, *via* un schéma de réduction pré-déterminé, des atomes en groupements fonctionnels qui formeront les noeuds (typés) du graphe réduit, tout en conservant la connectivité entre ces groupements. Chaque noeud du graphe représente donc un sous-ensemble d'atomes et de liaisons initialement connectés dans la molécule et possède un type décrivant les caractéristiques du groupement qu'il encode. Le type de chaque noeud est déterminé par le schéma de réduction utilisé et il en existe un grand nombre. On peut ainsi ne différencier que les noeuds représentant les cycles, les noeuds représentant les chaînes latérales et ceux encodant les atomes liant les cycles. On peut également utiliser des schémas plus poussés qui encoderont le type pharmacophorique des groupements (accepteur, donneur, cyclique, aromatique, etc.), soit *via* l'utilisation d'une règle permettant de prioriser les types, soit *via* une description plus fine permettant de combiner les différents types pharmacophoriques. Des descripteurs / fingerprints peuvent alors être générés en se basant sur une telle représentation (liste de triplets / quadruplets de noeuds, etc.) et ont notamment été utilisés pour des études de regroupement et de recherche par similarité en criblage virtuel [152], pour l'analyse SAR et l'optimisation multi-objective de molécules représentées sous forme de graphes réduits [153, 154] ou encore pour l'encodage de bioisostères [155].

3.4 Espaces chimiques

Le notion d'espace chimique est essentielle en chimoinformatique. C'est sur elle que s'appuient les fondations théoriques de la plupart des méthodes chimoinformatiques. Une définition du domaine reposant sur cette notion a d'ailleurs été énoncée par le Pr. Varnek lors des premières journées nationales françaises de chimoinformatique : "la chimoinformatique est le domaine qui considère les molécules ou les réactions comme des objets (graphe, vecteur) dans un espace chimique". Dans un premier temps, nous définirons plus en détail cette notion et nous verrons les méthodes existantes qui permettent de créer et d'exploiter les espaces chimiques afin de décrire des chimiothèques.

3.4.1 Définition générale

"L'espace chimique englobe toutes les molécules organiques possibles, y compris les molécules présentes dans les systèmes biologiques" [156]. C'est ainsi que Christopher M. Dobson introduit la notion dans une série d'articles [157–161] y étant consacrés. Dans un article de revue, Hann et Oprea [55] différencient quant à eux quatre types d'espaces chimiques :

Espace chimique virtuel

Il regroupe tous les composés qu'il serait théoriquement possible de synthétiser. En prenant un maximum par molécule de 30 atomes lourds communément trouvés dans les molécules de synthèse actuelles, on estime à 10^{60} [162] le nombre de molécules qui pourraient être synthétisées. Pour mieux se rendre compte du sens d'une telle grandeur, on peut la comparer aux 10^{17} secondes écoulées depuis le big bang (à quelques secondes près) ou encore à la masse de l'univers qui est estimée à environ 10^{54} kg.

Espace chimique tangible

Il regroupe toutes les molécules synthétisables avec les moyens actuels. Leur nombre a été estimé entre 10^{20} et 10^{24} [163].

Espace chimique global

Il regroupe tous les composés ayant déjà été synthétisés. Les molécules (organiques et inorganiques) référencées dans la base de données CAS sont au nombre d'environ 69 millions. Il reste néanmoins difficile d'estimer le nombre réel de molécules ayant déjà été synthétisées étant donné que chaque laboratoire ou entreprise pharmaceutique conserve parfois les données confidentielles.

Espace chimique réel

Il correspond à tous les composés possédés par un organisme (entreprise pharmaceutique, institut de recherche, etc.). La chimiothèque nationale contient aujourd’hui plus de 50 000 produits, et l’ICOA en possède de l’ordre de 10 000. Les plus gros catalogues de fournisseurs proposent plusieurs dizaines de millions de molécules, tandis que les chimiothèques propriétaires issues de l’industrie pharmaceutique en contiennent bien souvent plusieurs millions.

Etant donné son immensité, l’analogie avec l’espace au sens astronomique du terme est souvent retrouvée dans la littérature [161] où les molécules sont associées aux étoiles positionnées dans l’espace. De même que l’espace est essentiellement composé de vide, l’espace chimique est lui aussi composé de molécules qui n’ont pas forcément d’intérêt par rapport à une activité biologique donnée ou qui ne possèdent pas les propriétés adéquates pour en faire de bons médicaments. L’un des plus grand défi de la chémoinformatique et plus généralement de la recherche de médicaments, consiste en la recherche des zones de l’espace chimique dans lesquelles on trouvera les molécules biologiquement actives et qui pourront devenir de bons médicaments. Cette notion est donc utilisée en permanence afin de représenter et comparer des ensembles de molécules et nous allons détailler dans les sections suivantes les moyens permettant d’atteindre ce but.

3.4.2 Coordonnées et métriques

Les espaces chimiques que nous avons décrits dans le paragraphe précédent sont issus de définitions assez générales. Lorsque l’on fait l’analogie entre les molécules et les planètes, on imagine alors que chaque molécule possède également un ensemble de coordonnées dans l’espace chimique. Ainsi, naviguer dans un espace chimique nécessite l’utilisation d’un repère et de métriques afin de positionner les molécules et quantifier leurs distances respectives. Les descripteurs moléculaires sont alors le moyen le plus naturel d’assigner des coordonnées à une molécule. Puisqu’il encode numériquement les nombreuses propriétés, un descripteur permet de positionner des molécules dans un espace chimique à une distance reflétant leur similarité dans cet espace de descripteurs. En chémoinformatique, un espace chimique se définit presque toujours comme étant un espace multidimensionnel, le nombre de dimensions N étant donné par le nombre de descripteurs utilisés. Chaque produit aura alors une position déterminée par ses coordonnées obtenues à partir des N descripteurs calculés.

Une fois positionnées, il est bien souvent utile de pouvoir quantifier la distance entre deux

I.3 L'apport de la chémoinformatique

molécules dans un espace chimique donné. A cette fin, on utilise une métrique adaptée au type de descripteurs utilisés, qui renverra une valeur numérique quantifiant la distance (ou la similarité, suivant le problème) entre deux molécules. La table I.1 liste les métriques les plus utilisées.

Métrique	Variables continues	Variables binaires
Distance de Hamming	$\sum_{i=0}^N xA_i - x_i^B $	$a + b - 2c$
Distance Euclidienne	$\sqrt{\sum_{i=0}^N (xA_i - xB_i)^2}$	$\sqrt{a + b - 2c}$
Distance de Soergel	$\frac{\sum_{i=0}^N xA_i - xB_i }{\sum_{i=0}^N \max(xA_i - xB_i)}$	$\frac{a + b - 2c}{a + b - c}$
Coefficient de Tanimoto	$\frac{2 \sum_{i=0}^N xA_i xB_i}{\sum_{i=0}^N xA_i^2 + \sum_{i=0}^N xB_i^2 + \sum_{i=0}^N xA_i xB_i}$	$\frac{c}{a + b - c}$
Coefficient de Dice	$\frac{\sum_{i=0}^N xA_i xB_i}{\sum_{i=0}^N xA_i^2 + \sum_{i=0}^N xB_i^2}$	$\frac{2c}{a + b}$
Coefficient Cosinus	$\frac{\sum_{i=0}^N xA_i xB_i}{\sqrt{\sum_{i=0}^N xA_i^2 + \sum_{i=0}^N xB_i^2}}$	$\frac{c}{\sqrt{ab}}$

Tableau I.1 – Métriques utilisées en chémoinformatique pour comparer des vecteurs numériques. N représente le nombre de variables composant chaque vecteur. xA_i et xB_i représentent la valeur de la variable i pour les molécules A et B, respectivement. Pour les fingerprints binaires (seconde colonne), a représente le nombre de bits à 1 dans le fingerprint A, b le nombre de bits à 1 dans le fingerprint B et c le nombre de bits à 1 communs à A et B.

3.4.3 Visualisation d’espaces chimiques

L’utilisation d’un espace multi-dimensionnel permet un positionnement précis des molécules à travers la description de multiples propriétés. Il présente ainsi l’avantage de capturer une quantité importante d’informations. Néanmoins, l’analyse d’un espace multi-dimensionnel est rendue difficile par l’impossibilité d’en obtenir une représentation visuelle. La visualisation de données est un moyen efficace et intuitif permettant d’analyser et d’extraire de l’information sur une grande quantité de données et de ce fait beaucoup utilisée dans la recherche pharmaceutique [164–168], au même titre que n’importe quel autre domaine. Les espaces chimiques sont alors souvent décrits à l’aide d’espace réduits condensant l’information contenue dans un espace multi-dimensionnel en un nombre limité de coordonnées. Ces espaces sont construits en utilisant des méthodes permettant de projeter les données dans un espace de plus faible dimensionnalité, généralement 2 ou 3, tout en conservant le maximum d’informations présentes dans l’espace d’origine. Nous décrirons ici les méthodes permettant d’atteindre ce but, ainsi que leur utilisation dans le contexte de la description de chimiothèques.

Analyse en composantes principales L’analyse en composantes principales (ACP) est une méthode linéaire permettant de réduire la dimension d’un jeu de données contenant des objets décrits à l’aide de variables numériques. Le principe est de transformer un nombre de N variables en $M < N$ variables décorréliées entre elles. A l’aide d’une transformation linéaire, de nouveaux axes (les composantes principales) sont déterminés afin qu’ils maximisent la dispersion (i.e. la variance) des données projetées sur ceux-ci. Ces axes sont en général ordonnés de telle sorte qu’ils contiennent une quantité d’information décroissante et l’on en garde en général que 2 ou 3 afin de pouvoir projeter les données dans un espace pouvant être visualisé sur un écran, définissant ainsi un nouveau repère orthogonal.

En pratique l’ACP est assez simple à mettre en oeuvre. Les composantes principales peuvent en effet être déterminées en calculant les vecteurs propres de la matrice de covariance calculée sur le jeu de données généralement centré-réduit. Chaque axe (i.e. vecteur propre) ainsi obtenu se présente sous la forme d’une combinaison linéaire des données originales, à partir de laquelle on peut calculer les coordonnées dans le nouvel espace réduit. Pour obtenir la coordonnée sur l’axe i , on utilise donc une simple formule :

$$Coord_i = \sum_{k=0}^N x_k v_k \quad (\text{I.1})$$

Où x_k correspond à la valeur (éventuellement centrée-réduite) de la variable k et v_k correspond à l'élément k du vecteur propre i . Par ailleurs, la part de variance expliquée par chaque axe peut être quantifiée par le rapport entre la valeur propre correspondant au vecteur et la somme de toutes les valeurs propres obtenues, ce qui permet ainsi d'identifier les axes les plus représentatifs et de quantifier la quantité totale d'informations portée par un modèle ACP à N composantes.

L'ACP est probablement la méthode ayant été la plus utilisée pour la définition d'espaces chimiques et pour analyser et comparer des ensembles de produits. L'une des études les plus citées est celle d'Oprea et al. [169, 170], où le ChemGPS (Chemical Global Positioning System) fut introduit comme étant l'un des premiers systèmes permettant de positionner des molécules dans un espace multi-dimensionnel de référence. Traditionnellement, l'ACP était réalisée de manière ponctuelle sur de petits ensembles de molécules. Ici, en se basant sur 72 descripteurs physico-chimiques, les auteurs définissent un modèle de référence (en conservant les 9 premières composantes principales) à partir d'une chimiothèque contenant à la fois des molécules se voulant représentatives de l'espace chimique drug-like, mais également des molécules dites "Satellites" ayant des propriétés extrêmes. Ces satellites ont été introduits volontairement afin d'étendre au maximum le domaine d'applicabilité du modèle. Les auteurs suggéreront par la suite le terme de "Chemography" [170] comme étant l'art de naviguer dans les espaces chimiques.

Le ChemGPS a été beaucoup dérivé depuis sa première publication. Oprea et al. [171] définissent ainsi un nouvel espace de référence ADME en utilisant cette fois des descripteurs physico-chimiques issus du programme Volsurf [172]. Ils parviennent ainsi à corrélérer les propriétés issues des composantes principales à des propriétés telles que la perméabilité et la solubilité. Plus tard, Larsson et al. [173] observent que l'espace de référence du ChemGPS n'est pas représentatif des produits naturels, et reconstruisent alors un nouvel espace de référence pour ce type de produits. Ils introduiront ensuite le ChemGPS-NP [174] permettant de naviguer dans l'espace chimique des produits naturels et ainsi de déterminer la couverture de cet espace pour des bibliothèques destinées au criblage. Plus tard, un outil web de navigation en ligne permettant d'utiliser le ChemGPS-NP sera proposé [175]. Les auteurs illustreront son utilisation en comparant la distribution de produits naturels à celle de produits issus de la chimie médicinale [176]. En utilisant les 8 premières dimensions de l'espace de référence et la distance euclidienne, ils montreront que certains produits naturels voisins de molécules issues de Wombat [177] possèdent une activité biologique similaire. Cet espace de référence a été à nouveau utilisé récemment [178] afin de comparer la distribution de produits naturels à des chimiothèques dédiées au criblage

ainsi qu'à des librairies de médicaments.

L'ACP a également été utilisée afin de générer un espace chimique à deux dimensions dans le but de comparer la distribution de trois types de produits : des produits naturels, des médicaments et une sélection aléatoire de molécules issues d'une chimiothèque combinatoire [179]. Seuls dix descripteurs physico-chimiques ont été utilisés afin de décrire les molécules et les auteurs ont pu mettre en évidence des différences notables dans la couverture (et donc la diversité) de l'espace chimique ainsi généré. Ils concluent que les molécules issues de librairies combinatoires présentent une diversité moindre comparée aux composés naturels ou aux médicaments mais qu'ils sont présents dans une zone de l'espace chimique peu peuplée par les produits naturels. Dans une étude similaire, l'ACP a également été utilisée conjointement à d'autres méthodes afin de comparer plusieurs classes de composés [180]. Les auteurs mettent à nouveau en évidence la présence de composés issus de la chimie combinatoire dans des zones de l'espace chimique peu peuplées par les médicaments.

Dans une étude focalisée sur des molécules utilisées en oncologie, l'ACP a été utilisée afin d'analyser l'espace chimique 3D (les 3 premières composantes) de molécules actives et inactives sur certains cancers, ainsi que des molécules *drug-like* [181]. Les auteurs mettent en évidence une distribution significativement différente, les molécules actives couvrant un espace bien plus important malgré un certain degré de recouvrement. Ils mettent également en évidence des groupes de molécules localisés dans une zone très clairsemée, mais suffisamment proche d'une zone *drug-like* pour pouvoir suggérer la possibilité d'optimiser ces produits en les amenant vers une zone plus *drug-like*.

Plus récemment, une chimiothèque virtuelle (base GDB - the chemical universe Generated DataBase) composée de 26,4 millions de produits a été générée par le groupe de Raymond [182] en énumérant exhaustivement toutes les molécules de 11 atomes lourds restreints à un ensemble réduit d'atomes organiques courants (C, F, O, N). Ils ont créé un modèle ACP sur la base de 6 descripteurs physicochimiques. En comparant la distribution des produits virtuels à un ensemble de molécules de référence provenant de différentes bases publiques, ils ont pu décrire les zones de l'espace chimique dans lesquelles on ne trouvait que des molécules virtuelles et notamment des molécules contenant plusieurs groupes polaires. Ils ont ainsi pu mettre en évidence les zones qui n'ont pas ou peu été explorées par la chimie de synthèse. Ils ont également démontré l'efficacité de leur procédure de sélection des molécules les plus "réalistes" en comparant la distribution de GDB à celle d'un ensemble de plus d'un milliard de molécules

générées sans prendre en considération la stabilité et la faisabilité chimique et en montrant que la GDB ne couvre que très peu les zones de l'espace chimique dans lesquelles on retrouve des molécules considérées comme aberrantes.

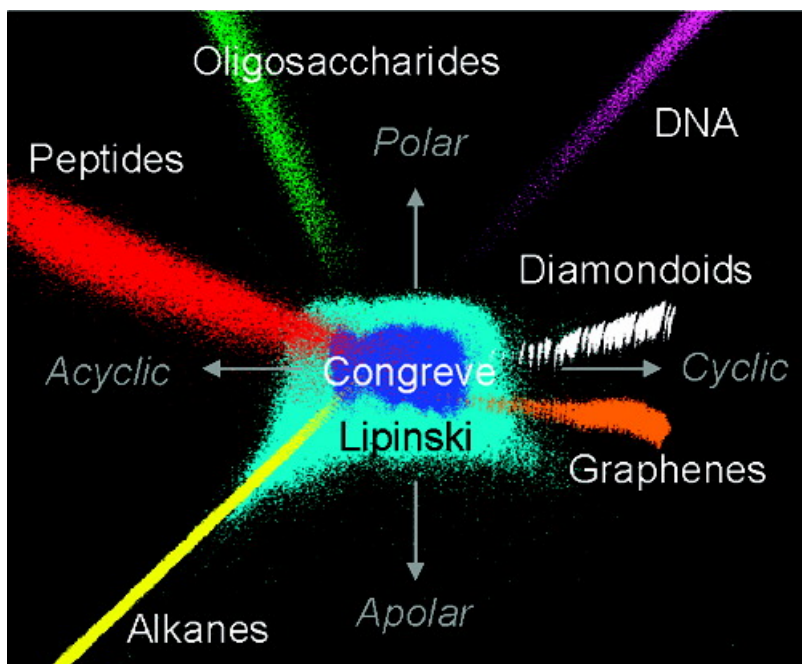


Figure I.5 – Une représentation de la base de données Pubchem *via* un modèle ACP calculé sur 42 descripteurs (les MQNs). Figure extraite de Van Deursen et al. [183]

Le même groupe a également utilisé l'ACP [183] afin d'analyser la base de données Pubchem [91]. Un ensemble de 42 descripteurs (les MQNs - Molecular Quantum Numbers) comptant différents éléments (nombre d'atomes, de liaisons, de groupes polaires, etc.) a été calculé sur l'ensemble des molécules de la base et un modèle ACP a été généré. Les trois premières composantes principales ont ensuite été utilisées afin de représenter graphiquement les projections de différents groupes de molécules sur ces trois axes. Ils noteront la présence de formes coniques (Figure I.5) correspondant à certaines classes de composés (oligonucléotides, polymères, etc.). Plus tard, le même groupe utilisera à nouveau l'ACP afin de décrire une nouvelle base de molécules virtuelles de près de 970 millions de produits contenant jusqu'à 13 atomes lourds [184, 185]. Ils feront de même avec la base de données Pubchem afin cette fois de mettre en évidence les différences de couverture de l'espace chimique entre les molécules *drug-like*, *lead-like* et *fragment-like* [186].

Cartes auto-organisatrices de Kohonen Les cartes auto-organisatrices sont un ensemble de méthodes non-linéaires de réduction de dimension utilisant un réseau de neurones non-supervisé. Elles ont été introduites en 1982 par Teuvo Kohonen [187]. Une carte de Kohonen est modélisée par un ensemble de N vecteurs (les neurones ou modèles) constitués de M éléments qui représentent les descripteurs utilisés pour décrire les objets à représenter, les molécules dans notre cas. Chaque neurone est associé à une position dans l'espace d'apprentissage, à travers une valeur numérique pour chacun des M éléments constituant le vecteur. Ils sont généralement répartis uniformément sur une grille en deux dimensions (cases rectangles ou hexagonales), chaque neurone possédant une position unique sur celle-ci. L'ensemble des objets est alors successivement présenté aux neurones et chaque vecteur est mis à jour au cours de l'apprentissage, de telle sorte qu'à la fin de celui-ci, deux neurones proches sur la carte représentent deux régions proches dans l'espace d'apprentissage.

Concrètement, l'algorithme d'apprentissage se déroule de manière itérative. Au départ, chaque neurone se voit assigner des valeurs aléatoires pour chacun des éléments de son vecteur (les valeurs sont généralement bornées afin de ne pas s'éloigner de l'espace de descripteurs original). Lors de chaque itération, l'ensemble (ou un sous-ensemble tiré aléatoirement) des molécules est présenté à la carte et pour chacune des molécules le neurone "vainqueur", celui ayant la distance la plus faible par rapport à celle-ci, est déterminé. Le vecteur représentant le neurone vainqueur sera alors mis à jour avec une certaine intensité décroissante avec le nombre d'itération, afin de le rapprocher du vecteur de la molécule et le rendre ainsi plus représentatif. Durant cette phase, les vecteurs voisins seront également mis à jour, mais dans une moindre mesure. L'étendue du voisinage concerné par ce changement est généralement déterminée par une fonction linéaire ou gaussienne qui diminue avec la distance entre chaque neurone sur la grille. De la même façon, l'ampleur des modifications effectuées pour les neurones concernés sera également proportionnelle à la distance au neurone vainqueur, de telle sorte que plus un neurone voisin est proche du vainqueur, plus il sera adapté pour se rapprocher de celui-ci. Finalement, la règle de modification d'un neurone v_i modèle à l'itération t de présentation de la molécule m est donné par :

$$v_i(t+1) = v_i(t) + \alpha(t)N_v(t)(m(t) - v_i(t)) \quad (\text{I.2})$$

où $\alpha(t)$ représente la vitesse à laquelle chaque neurone vainqueur est modifié durant l'itération t (cette vitesse diminue au fur et à mesure du déroulement de l'algorithme, généralement *via* une méthode linéaire décroissante en fonction de t), $N_v(t)$ permet de déterminer l'ampleur

des modifications au voisinage et est généralement représentée par une fonction linéaire ou gaussienne diminuant avec la distance entre les neurones, $m(t)$ est le vecteur de la molécule présentée et $v_i(t)$ est le vecteur du neurone gagnant.

Les cartes de Kohonen ont beaucoup été utilisées en chémoinformatique [188]. De part leur nature non-linéaire, elles sont bien adaptées à des études de regroupement où l'on souhaite mettre en avant des corrélations entre structure et propriétés (descripteurs). Elles ont d'ailleurs été utilisées assez tôt, dès les années 90, dans cette optique [189, 190] et sont depuis surtout utilisées dans des études de regroupement où l'on cherche à séparer des molécules actives sur une cible / une classe de cibles, des molécules inactives. Elles ont ainsi, par exemple, été utilisées afin d'identifier de nouveaux inhibiteurs de protéines G [191, 192], des inhibiteurs du CYP 3A4 [193] ou encore des inhibiteurs du canal ionique hERG [194–196].

Les cartes de Kohonen ont également été utilisées dans l'optique d'analyser et comparer le contenu de bibliothèques. Lee et al. [197] analysent ainsi les squelettes moléculaires d'un ensemble de produits naturels et les comparent à ceux contenus dans une base de médicaments. À l'aide d'une carte de Kohonen générée sur 150 descripteurs topologiques et physico-chimiques, ils mettent en évidence les zones de l'espace chimique contenant des classes de scaffolds que l'on ne retrouve pas dans les médicaments. Une étude assez similaire a été réalisée quelques années plus tard [198], toujours par G. Schneider et al.

En 2003, Schneider et Schneider proposent une nouvelle bibliothèque de référence (COBRA - Collection of Bioactive Reference Analogues) contenant 4 236 médicaments ou candidats médicaments [199]. Ils réalisent une carte de Kohonen de 100 neurones sur l'ensemble de ces molécules en utilisant des descripteurs topologiques quantifiant la distribution de paires de points pharmacophoriques (méthode CATS [200]). Ils montrent que certains types de molécules (par exemple, les inhibiteurs de l'enzyme ACE - Angiotensin Converting Enzyme) forment des ensembles bien regroupés dans certaines zones de l'espace chimique.

Ertl et al. [201] ont de leur côté généré environ 600 000 squelettes hétéroaromatiques et ont créé une carte de Kohonen de 10 000 neurones en utilisant différents descripteurs topologiques. L'espace chimique ainsi obtenu leur permettra d'analyser la couverture de composés biologiquement actifs et d'identifier 6 îlots contenant des squelettes majoritairement associés à des molécules actives. Ils notent également que l'espace couvert par les squelettes actifs est nettement moins important que l'espace couvert par des squelettes générés virtuellement, ce qui

suggère à nouveau que l'espace chimique à explorer est d'autant plus vaste.

Dans une étude présentée dans la section précédente, le groupe de Reymond utilise les cartes de Kohonen conjointement à l'ACP afin de décrire la base virtuelle GDB [182] à l'aide de descripteurs d'autocorrélation. Une carte de 200x200 neurones a été créée et entraînée avec 1 million de molécules sélectionnées aléatoirement parmi la GDB. En colorant les molécules par classe chimique (aromatiques, hétéroaromatiques, acycliques, etc.), ils obtiennent une représentation globale de l'espace chimique GDB (Figure I.6) et mettent ainsi en évidence les différences de répartition entre plusieurs classes de molécules.

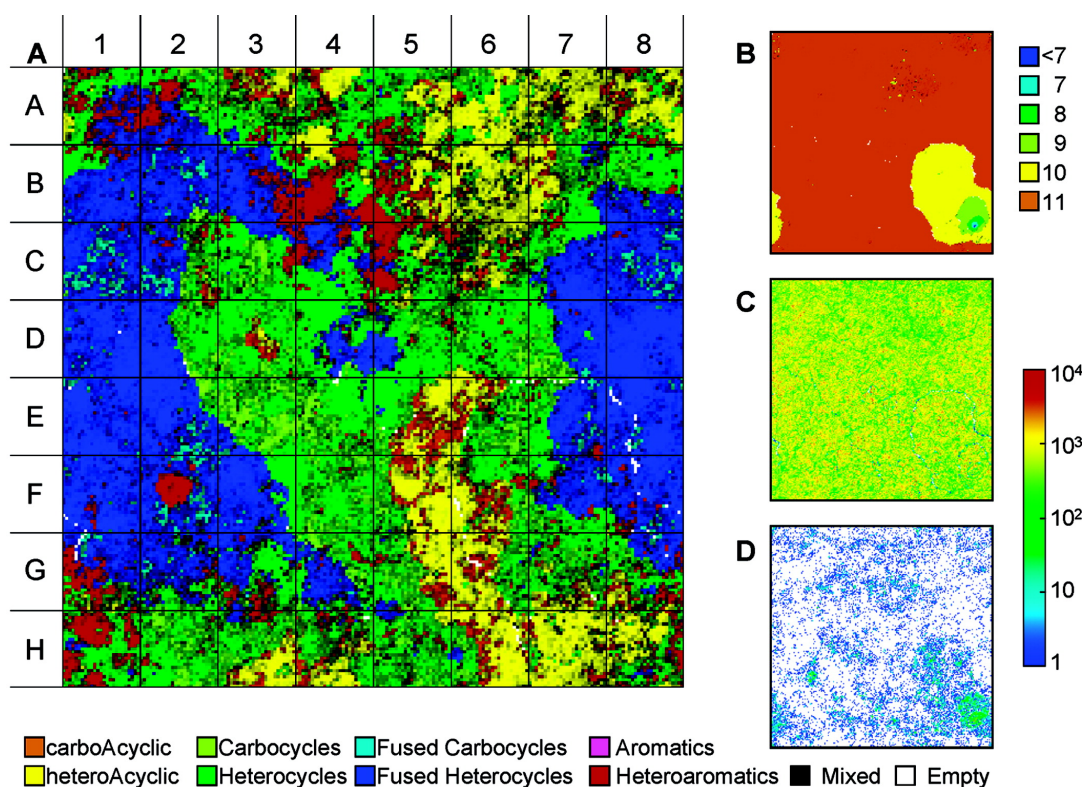


Figure I.6 – Une représentation de l'espace chimique peuplé par la base GDB. Les couleurs représentent les différents types de molécules, classées selon leur composition chimique. Figure extraite de Fink et al. [182]

Generative Topographic Maps (GTM) Les GTM ont été introduites à la fin des années 90 [202, 203] comme étant une alternative probabiliste aux cartes de Kohonen. Malgré sa robustesse théorique, on dénombre encore assez peu d'applications de la méthode à l'analyse de chimiothèques.

L'une des premières applications de cette méthode en chémoinformatique est celle de Maniyar et al. en 2006 [164]. Ils ont utilisé les GTM afin de visualiser et analyser des données propriétaires issues de 5 campagnes de criblage haut débit réalisées sur 5 cibles chez Pfizer. En comparant les projections obtenues avec les cartes de Kohonen, l'ACP et Neuroscale [204] (une implémentation des projections de Sammon [205] basée sur des réseaux de neurones), les auteurs montrent que les GTM permettent de mieux regrouper les molécules actives retenues pour cette étude. Ils présentent également une version hiérarchique de la méthode, qui permet d'analyser plus en détails les différents groupes de molécules présents dans des zones très denses et discontinues.

Owen et al. [206] ont eux comparé les GTM (ainsi que deux autres méthodes dérivées et adaptées aux variables binaires) à NeuroScale et à l'ACP dans le but de visualiser des chimiothèques décrites à l'aide de fingerprints binaires. En projetant différentes librairies (combinatoires, molécules biologiquement actives), ils mettront en évidence la capacité des GTM, et notamment des versions adaptées aux variables binaires, à produire des groupes de molécules plus purs (en termes de similarité) comparés à ceux obtenus avec les autres méthodes.

Enfin, dans une revue récente présentant la méthode et ses applications potentielles en chémoinformatique [207], les GTM ont également été utilisées afin de visualiser les librairies disponibles dans le Directory of Usefull Decoys (DUD) [208], un jeu de données de référence utilisé pour la validation de méthodes de criblage virtuel. La méthode a été comparée aux cartes de Kohonen et à l'ACP pour la visualisation de librairies et les auteurs ont ainsi montré que les GTM permettent de mieux regrouper les différentes classes de molécules présentes dans le DUD. Parmi les atouts mis en avant lors de cette étude, la nature probabiliste de la méthode a été présentée comme un avantage important permettant de quantifier plus précisément le degré de recouvrement entre deux librairies.

Stochastic Proximity Embedding La méthode SPE (Stochastic Proximity Embedding) [209–211] a été introduite par Agrafiotis en 2002 [209]. Cette méthode permet de générer des coordonnées dans un espace de faible dimension (2 ou 3) à partir d'une matrice de distances dans un espace donné. A partir d'un ensemble de N objets, d'une matrice de distances entre ces objets et d'un ensemble de points localisés dans l'espace à m dimensions $x_i, i = 1, 2, \dots, N; x_i \in \mathbb{R}(m)$, la méthode procède de manière itérative en essayant de placer les différents points dans $\mathbb{R}(m)$ de telle sorte que la distance euclidienne entre ces points soit la plus représentative possible de la distance donnée en entrée. Les points dans $\mathbb{R}(m)$ sont tout d'abord initialisés de manière aléatoire et l'algorithme procède ensuite de manière itérative en sélectionnant au hasard des

paires de points et en ajustant leurs coordonnées de telle sorte que la nouvelle distance obtenue dans $\mathfrak{R}(m)$ soit la plus représentative de la distance originale. L'intensité de la modification des coordonnées est proportionnelle à un coefficient dit de disparité, qui, de manière similaire aux cartes de Kohonen, décroît avec le nombre d'itérations réalisées.

Bien que la méthode ait été introduite dans le cadre de la visualisation de très grosses chimiothèques [211], elle a été relativement peu utilisée dans ce contexte depuis. Suite à une modification de l'algorithme original [212], elle a surtout été illustrée dans le cadre d'exploration d'espaces conformationnels [213–215] et fut comparée à d'autres programmes et validée en 2007 [216]. Elle a également été utilisée pour comparer différents modèles QSAR en les projetant dans un espace 2D généré par la méthode SPE [217]. De manière plus surprenante, Agrafiotis indique que la méthode a aussi été appliquée à la prédiction de poses pour le docking, ainsi qu'à la depiction de molécules [211].

L'un des principaux avantages de cette méthode est sa rapidité (la complexité de l'algorithme étant linéaire), mais l'inconvénient majeur est qu'elle ne permet pas de créer d'espace de référence dans lequel on pourrait projeter de nouveaux objets une fois l'algorithme terminé (de la même manière que les méthodes similaires, telles que le Multi-dimensional Scaling ou les projections de Sammon [205]).

Cartes de fusion par similarité En se basant sur des mesures de similarité, les cartes de fusion par similarité (Fusion Similarity Maps [218]) permettent de comparer une ou plusieurs chimiothèques à un ensemble de molécules de référence. Cette méthode diffère des autres méthodes présentées jusqu'ici dans la mesure où l'idée n'est pas de représenter sur la carte 2D la proximité des molécules entre elles, mais plutôt leur similarité par rapport à la chimiothèque de référence. A partir d'une certaine représentation, la matrice de similarité entre les molécules de référence et les molécules des librairies à traiter est tout d'abord calculée. Une visualisation 2D est ensuite obtenue en traçant, pour chaque molécule des librairies à traiter, les valeurs de similarité moyenne par rapport aux molécules de référence en abscisse et les valeurs de similarité maximale en ordonnée. Deux points proches sur la représentation ainsi obtenue représenteront donc deux composés ayant des valeurs similarité moyenne et maximale proches par rapport à la librairie de référence, sans que ces molécules ne soient nécessairement similaires entre elles.

L'avantage de cette méthode est qu'elle permet d'identifier aisément les molécules redondantes (similarité maximale et moyenne élevée), les molécules similaires à certaines des molécules de

référence (similarité maximale élevée) mais globalement diverses (similarité moyenne faible), ou les molécules apportant une réelle diversité (similarité maximale et moyenne faibles). L'inconvénient majeur est sa complexité qui augmente de manière quadratique avec le nombre de composés dans la librairie de référence et dans les librairies à traiter.

Outre l'article présentant la méthode, dans lequel des librairies supposées diverses (NCIdiv, Drugbank) ont été comparées, la méthode a également été utilisée conjointement à l'ACP pour comparer des librairies combinatoires à des ensembles de produits naturels et de médicaments. Les auteurs ont ainsi pu montrer que leur chimiothèque combinatoire contient de nombreux produits localisés dans des régions peu peuplées de l'espace des produits naturels et des médicaments. Elle a également été utilisée dans la même optique [219] pour à nouveau mettre en avant la diversité de chimiothèques combinatoires.

3.4.4 Conclusion

Les méthodes présentées ici permettent d'obtenir une représentation visuelle de chimiothèques et facilitent ainsi l'analyse et la comparaison de plusieurs chimiothèques par rapport à un espace chimique donné. Elles ont toutes leurs qualités et leurs défauts propres et sont généralement utilisées pour traiter des problèmes différents. Ainsi, les méthodes linéaires comme l'ACP sont généralement utilisées à des fins descriptives pour comparer la couverture globale de plusieurs chimiothèques. Elles présentent l'avantage de pouvoir être utilisées sur de gros volumes de données, tout en restant stables et robustes lorsqu'elles sont bien utilisées. Les méthodes non-linéaires quant à elles, sont plutôt utilisées dans des études où l'on cherche clairement à séparer des groupes de molécules, typiquement actives *vs* inactives. De part leur nature non-linéaire, elles sont capables de capturer une information qui se veut souvent complémentaire à celle renvoyée par les méthodes linéaires telle que l'ACP.

3.5 La Diversité

La diversité est devenue partie intégrante des programmes de recherche modernes [80, 133, 220–223]. Comme nous avons pu le voir, le nombre de molécules disponibles pour le criblage augmente de façon rapide et constante. Puisqu'il n'est pas possible de tester l'ensemble des produits disponibles sur le marché et qu'il est important de rationaliser et optimiser le contenu des chimiothèques à tester, il devient naturellement nécessaire de sélectionner un sous-ensemble de composés destiné au criblage. Nous avons vu précédemment qu'il existe un premier ensemble de méthodes permettant de sélectionner les molécules afin de minimiser le nombre de faux positifs

ou se restreindre aux molécules les plus *drug-like*. La diversité se trouve être un moyen supplémentaire permettant de s'assurer que l'on couvre le mieux possible un espace chimique donné. Elle est en général appliquée en phase amont de la recherche lorsqu'il y a peu d'informations disponibles sur la cible et sur les molécules éventuellement connues, lors de l'acquisition de nouveaux composés dans le cadre d'un programme d'enrichissement de chimiothèque ou dans la création de librairies combinatoires [224, 225].

La notion de diversité repose sur une idée simple : minimiser la redondance tout en maximisant la couverture de l'espace chimique à explorer. Elle trouve ses fondations dans le principe de similarité qui suggère que des molécules similaires auront des propriétés biologiques similaires. Plusieurs analyses ont ainsi montré que pour deux molécules ayant un coefficient de similarité (Tanimoto) supérieur à 0,85 et si l'une d'entre elle démontre une certaine activité, il existe une probabilité de 80% que la deuxième molécule possède la même activité [66, 226]. D'autres analyses rétrospectives présentent des résultats plus contrastés, avec entre 20 et 40 % de chances qu'un composé ayant un coefficient de Tanimoto supérieur à 0,85 par rapport à une molécule active soit actif lui-même [19, 227, 228]. La validation du principe de similarité n'est pas chose aisée, comme le démontrent les résultats discordants présentés dans ce paragraphe. Les résultats obtenus sont en effet grandement dépendants des cibles étudiées, du mode de description des molécules, ainsi que de la métrique permettant de les comparer.

Ce principe reste néanmoins généralement accepté, que ce soit par les chémoinformaticiens ou par les chimistes médicaux, bien que de nombreuses exceptions aient été constatées [227], comme l'illustre la figure I.7. En témoigne l'existence de nombreuses librairies dédiées au criblage ayant été créées à partir d'une recherche de molécules similaires à des molécules actives connues. Ce type de situation est idéal pour démarrer un projet de recherche. Cependant, lorsque l'on ne dispose que de peu d'informations sur la cible ou si aucune molécule active n'est connue, il devient nécessaire d'utiliser des chimiothèques destinées à un criblage "aveugle" ne faisant aucune hypothèse sur la zone de l'espace chimique à explorer.

Typiquement, une librairie diverse devra alors présenter le moins de redondance possible tout en explorant au mieux l'espace chimique afin de maximiser les chances de trouver une molécule active. Elle offre en outre la possibilité d'obtenir des points de départ différents pour la phase d'optimisation et ainsi de ne pas se focaliser sur une seule zone de l'espace chimique actif pour la cible étudiée. Un processus itératif est généralement utilisé dans de tels cas, en criblant des chimiothèques diverses lors d'une première itération, puis en criblant des chimiothèques focalisées créées à partir des *hits* ainsi obtenus.

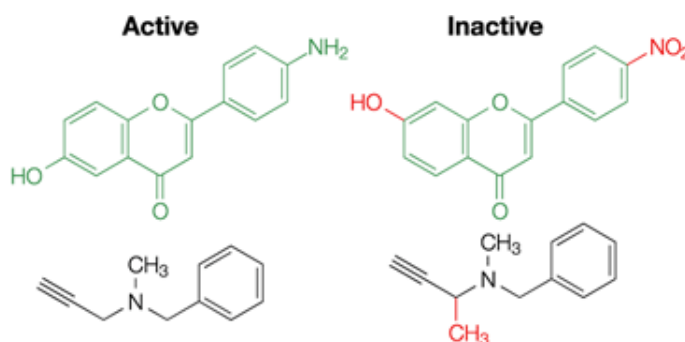


Figure I.7 – Un exemple de molécules très similaires mais ayant une activité biologique différente. En vert, des molécules ciblant une tyrosine kinase et en noir, des molécules ciblant une monoamine oxydase. Figure extraite de [229]

De nombreuses études ont tenté de démontrer l'apport de la diversité pour le criblage haut débit. Dans un article opposant les deux approches [230], Hajduk présente le criblage sur fragments comme étant une alternative plus efficace à l'utilisation d'une librairie diverse. *A contrario*, Galloway et Spring suggèrent que les bibliothèques diverses permettent d'explorer l'espace chimique de manière plus exhaustive et moins biaisée et apportent plus de nouveauté quant aux types de molécules testées. Les deux points de vues se rejoignent néanmoins sur une chose : l'approche la plus efficace dépend bien entendu du problème et notamment des informations dont on dispose sur la cible et sur des molécules actives connues. Il est également intéressant de noter que l'un des arguments de Hajduk en faveur du criblage sur fragments est qu'il permet justement d'accéder à plus de diversité mais avec moins de molécules, étant donné que l'espace chimique des fragments est par définition plus restreint que l'espace chimique global.

Récemment, une étude rétrospective de Novartis a pu démontrer que l'utilisation de bibliothèques diverses augmente significativement le taux de molécules actives détectées par rapport à une librairie dont les molécules ont été sélectionnées aléatoirement [77]. En utilisant une méthode simple de regroupement basée sur un fingerprint, ils ont sélectionné un sous ensemble divers de 250 000 molécules (20% de l'ensemble de la chimiothèque utilisée). Les auteurs ont alors constaté une augmentation significative du taux de hits pour 86 % des campagnes de criblages. De nombreuses autres études ont également mis en avant l'avantage de la diversité (et notamment dans le cadre de criblage sur des bibliothèques issues de la chimie combinatoire) à travers la découverte de molécules actives suite à des projets de criblages utilisant des chimiothèques diverses [73–76, 78–80].

Concrètement, la diversité trouve deux applications principales dans l’analyse et la conception de chimiothèques : la description et la quantification de la diversité d’un ensemble de produits et la sélection de sous-ensembles divers au sein d’une chimiothèque de référence. Nous présenterons ici ces deux aspects en décrivant les méthodes les plus utilisées dans la littérature.

3.5.1 Méthodes de quantification de la diversité

La diversité n’a de sens que lorsqu’elle est définie pour un ensemble de molécules. Les méthodes permettant de quantifier la diversité d’une chimiothèque permettent ainsi de statuer sur la qualité (en termes de diversité) de celle-ci ou bien de comparer entre elles plusieurs chimiothèques. Cependant, la définition (et donc la quantification) de la diversité d’un ensemble de molécules est, comme souvent, une chose suggestive. La plupart des méthodes existantes nécessitent en général au moins trois composantes essentielles : (1) un mode de représentation des molécules permettant de les décrire de manière numérique, (2) une métrique permettant de quantifier la similarité (ou la dissimilarité) entre molécules et (3) une fonction permettant de quantifier la diversité elle-même. Etant donné le nombre de descripteurs et de métriques existants, le choix de ces trois composantes s’avère bien souvent délicat. Bien sûr, dans le cas où l’on souhaite créer une chimiothèque dédiée au criblage sur une cible bien particulière, il devient indispensable de valider les modes de description utilisés afin de confirmer qu’ils permettent effectivement une bonne séparation entre molécules actives et molécules inactives. Ce type d’étude permet alors, en cas de succès, d’obtenir un jeu de descripteurs qui rend compte des propriétés capturant suffisamment d’informations pour rendre compte de l’activité des molécules et ainsi de restreindre la zone de l’espace chimique à étudier. Pour une chimiothèque destinée à être utilisée en criblage aveugle, le choix du mode de description est moins critique puisqu’on ne cherche pas à explorer une zone de l’espace chimique étant spécifique d’une ou plusieurs cibles en particulier.

Quoi qu’il en soit, les méthodes existantes peuvent être classées dans trois catégories principales :

- Les méthodes basées sur les distances, qui quantifient la diversité en utilisant des mesures de similarité ou de dissimilarité entre les molécules.
- Les méthodes basées sur un espace chimique multi-dimensionnel partitionné, qui permettent de quantifier la couverture d’une librairie projetée dans cet espace en comptant le nombre de zones (ou cellules) occupées.
- Les méthodes basées sur les structures, qui tentent de quantifier la diversité en termes

de sous-structures présentes dans une librairie donnée.

Méthodes basées sur la dissimilarité

Ces méthodes ont pour objectif de quantifier la dissimilarité entre les molécules présentes dans une chimiothèque donnée. De ce fait, elles nécessitent de calculer l'ensemble des distances entre les molécules à l'aide d'une métrique adaptée au type de description utilisée. Elle sont donc généralement coûteuses en temps de calcul, et plutôt appliquées pour évaluer la diversité sur des chimiothèques de taille raisonnable.

Parmi les plus utilisées, on trouve la somme des dissimilarités entre molécules, la valeur moyenne des dissimilarités entre molécules, la distance minimale observée sur l'ensemble des distances inter-molécules, ou encore la distance minimale moyenne observée entre paires de molécules [231].

Ce type de mesure est par ailleurs souvent utilisé dans le cadre de l'optimisation de sous-ensembles divers sélectionnés à partir d'une chimiothèque de référence car nombre d'entre elles ne sont adaptées que pour la comparaison de chimiothèques de même taille (par exemple, toutes les métriques faisant la somme des distances, dont la valeur augmente naturellement avec le nombre de produits étudiés). Une étude récente a par ailleurs évalué et comparé certaines de ces métriques ainsi qu'une nouvelle permettant de prendre en compte l'activité des molécules lorsque celles-ci sont disponibles [232]. Les expériences réalisées se sont limitées à des chimiothèques de faible taille, entre 800 et 13 000 produits, montrant ainsi la difficulté d'utiliser ce type de métrique sur de grosses chimiothèques, d'autant plus lorsqu'elles sont utilisées comme critère d'optimisation de la diversité.

Bien qu'indirectement basées sur le calcul des distances, l'ensemble des méthodes de regroupement (Jarvis-Patrick, K-mean, etc.) peuvent également être utilisées afin de quantifier la diversité d'une chimiothèque. A l'issue d'une étape de regroupement, la diversité se quantifie alors en général à partir du nombre de groupes (et de leur composition) ainsi obtenus.

Méthodes basées sur le partitionnement d'espace chimique

Les méthodes de partitionnement [132, 233] reposent sur la définition d'un espace chimique de faible dimension (typiquement 10 au maximum), que l'on décomposera en plusieurs zones (ou cellules) formant un ensemble d'hypercubes dans l'espace de référence. Pour chaque propriété utilisée, l'intervalle contenant l'ensemble des valeurs est divisé en un certain nombre de

bins, suivant le nombre de molécules à sélectionner. Une fois le partitionnement obtenu, chaque molécule est associée à une cellule à partir de ses coordonnées dans l'espace de référence. La diversité d'une librairie peut alors être évaluée de deux manières principales :

- En quantifiant la proportion de cellules contenant des molécules
- En évaluant l'uniformité de la distribution des cellules remplies

Le premier cas est trivial et peut se calculer facilement et rapidement. La valeur obtenue permet ainsi de quantifier la proportion de l'espace couvert par une librairie donnée. En revanche, elle ne permet pas en soit de caractériser l'uniformité de cette couverture. Une librairie ayant 50% de cellules occupées peut ainsi ne couvrir qu'une partie restreinte de l'espace chimique avec une forte densité de composés, tandis qu'une autre librairie avec la même proportion de cellules remplies couvrira elle l'ensemble de l'espace mais de manière plus uniforme. Du point de vue de la diversité, le deuxième cas est bien sûr préférable. Le deuxième type de mesure permet alors de mieux caractériser l'uniformité de la couverture de l'espace partitionné. Shammugasundaram et Maggiora ont par exemple utilisé l'entropie de Shannon afin de quantifier d'une part, l'uniformité de répartition sur l'ensemble des cellules de l'espace partitionné et d'autre part, cette uniformité en prenant en compte chaque axe de l'espace chimique [234]. Agrafiotis utilise lui la statistique de Kolmogorov Smirnov [235, 236], qui permet de comparer une distribution observée à une distribution théorique et d'estimer ainsi la correspondance. En l'occurrence, une chimiothèque diverse devra présenter une répartition uniforme des produits dans un espace chimique donné.

Le principal avantage de ces méthodes est qu'elles permettent une évaluation très rapide de la diversité, tout en étant intuitives et bien adaptées à des espace chimiques reposant sur des descripteurs moléculaires. La difficulté principale réside en la définition des intervalles définissant les partitions sur chaque dimension et la prise en compte des molécules exotiques qui ont une influence importante dans la définition des intervalles de valeurs [237]. Des solutions ont été proposées pour minimiser les artefacts liés à ce type de méthodes [238, 239] et nous présenterons notre propre contribution basée sur les DRCS dans le chapitre 3.

Méthodes basées sur les structures

Ces méthodes reposent sur l'idée que la diversité d'une chimiothèque peut également se définir par la quantité et la distribution du nombre de fragments qu'elle contient. A nouveau, le terme "fragment" reste subjectif, étant donné qu'il existe de nombreux moyens de décomposer

une molécule en un ensemble de fragments. Des méthodes d'énumérations tels que les descripteurs fragmentaux présentés dans les sections précédentes, permettent par exemple d'obtenir une liste exhaustive de fragments trouvés au sein d'une molécule, et donc d'en établir la liste complète pour une chimiothèque donnée. Il existe également d'autres méthodes qui, plutôt que d'énumérer tous les fragments présents, appliquent en amont du processus de fragmentation une méthode de réduction du graphe moléculaire (telle que les graphes réduits) afin de regrouper les atomes en groupements fonctionnels. La fragmentation consiste alors à énumérer les fragments présents dans la molécule. Cette énumération peut se faire essentiellement de deux manières : soit de manière exhaustive, où l'ensemble des combinaisons de fragments connectés est générée, soit de manière plus simple, où l'on énumère les groupements obtenus lors de la simplification du graphe moléculaire. D'autres méthodes ont également été développées selon le même principe, en utilisant cette fois des indices encodant les atomes et leur environnement [26].

En se basant sur l'ensemble des fragments obtenus par une méthode donnée, il est ainsi possible de comparer la diversité de deux chimiothèques en utilisant différentes mesures. La plus simple d'entre elle est un comptage du nombre de fragments uniques au sein de la chimiothèque (éventuellement en classant les fragments par groupes : cycliques, aromatiques, etc.), qui permet ainsi d'obtenir une première idée de la diversité. Le rapport entre le nombre de fragments et le nombre de molécules fournit une information similaire, mais permet d'appliquer la méthode pour comparer des chimiothèques de tailles différentes.

De manière assez similaire, la diversité d'une chimiothèque peut également se définir à partir d'une analyse par scaffolds. Nous avons vu qu'il existe différents schémas de réduction pour obtenir le scaffold d'une molécule et l'on peut alors analyser la diversité à partir de cette représentation de manière analogue à l'utilisation de fragments. En 2006, Krier et al. [28] ont utilisé les scaffolds afin d'analyser 17 collections provenant de différents fournisseurs. La diversité fut évaluée par le nombre de scaffolds regroupant 50% des composés initialement présents dans les librairies, une mesure se voulant indépendante de la taille des librairies. Medina-Franco et al. ont eux utilisé une méthode basée sur l'entropie de Shannon afin de quantifier la diversité de chimiothèques à partir d'une simple énumération des squelettes moléculaires [240]. L'entropie de Shannon a permis ici de quantifier l'uniformité de la distribution du nombre de molécules associées à chaque scaffold. Une étude récente [241] a utilisé le Scaffold Tree [242, 243], ainsi qu'une analyse basée sur la définition de Bemis et Murcko afin d'étudier la diversité de chimiothèques. Le Scaffold Tree est une méthode permettant d'obtenir une hiérarchie de scaffolds présents au sein d'une chimiothèque. La méthode décompose les molécules de manière itérative

en supprimant les cycles au fur et à mesure à l'aide de règles complexes qui permettent de sélectionner les cycles à supprimer lors de chaque itération. Dans cette étude, le nombre de scaffolds uniques, le rapport entre le nombre de scaffolds et le nombre de molécules, la proportion de scaffolds représentant une certaine proportion de la base initiale (mesure similaire à celle utilisée par Krier et al. [28]) ou encore le rapport entre le nombre de scaffolds singletons (n'étant associé qu'à une seule molécule) et le nombre total de molécules ont été utilisés afin de comparer les librairies.

Autres méthodes

On notera enfin l'existence de méthodes alternatives, peu utilisées aujourd'hui, mais utilisant un mode de description assez original. Notamment, Martin et al. [244] utilisent un fingerprint global pour une librairie donnée en calculant les fingerprints sur l'ensemble des molécules de la librairie et en faisant un OU logique sur ceux-ci. Le fingerprint ainsi obtenu peut être comparé à d'autres fingerprints générés sur des librairies en utilisant une métrique standard pour la comparaison de vecteurs de bits.

3.5.2 Méthodes de sélection par diversité

Les mesures de diversité sont utilisées afin de comparer plusieurs chimiothèques et sélectionner celles qui contiennent le moins de redondances en vue d'un criblage ou d'une étude d'enrichissement. Lorsque les capacités de criblage ne permettent pas de tester l'ensemble des composés d'une chimiothèque, on utilise alors des méthodes permettant de sélectionner un sous-ensemble de molécules à cribler. Outre les différents filtres présentés dans la première partie de cette section, les méthodes de sélection par diversité [133, 223, 245] sont typiquement utilisées afin de restreindre le nombre de produits à tester, tout en essayant de couvrir au maximum un espace chimique donné.

Méthodes basées sur les études de regroupement Ce type de sélection implique de regrouper les molécules d'une chimiothèque en sous-ensembles homogènes dans lesquels les molécules possèdent un degré important de similarité, tout en maximisant la dissimilarité entre les différents groupes obtenus. Elles ont beaucoup été utilisées en chimie [246] et notamment dans les études de diversité [133, 247] ou regroupement pour l'identification de séries actives [226]. Une fois les groupes obtenus, la sélection consiste généralement en l'ajout dans la chimiothèque diverse d'une ou plusieurs molécules par groupe.

Les méthodes de regroupement sont généralement classées en deux catégories principales :

les méthodes hiérarchiques et les méthodes non-hiérarchiques. Dans le premier cas, les groupes sont obtenus en combinant (regroupement agglomératif) ou en divisant (regroupement divisif) de manière itérative d'autres groupes. La méthode de Ward [248] est certainement la plus connue dans cette catégorie. Dans les méthodes non-hiérarchiques, les groupes sont obtenus sans prendre en compte les relations entre les différents sous-ensembles. Les méthodes de K-mean [249] (dont beaucoup d'algorithmes ont été dérivés et appliqués aux études de regroupement et de diversité sur de gros jeux de données [250–252]) ou de Jarvis-Patrick en sont les plus connues. Cette dernière a notamment été mise en avant pour sa capacité à être appliquée à de grands jeux de données [253], bien qu'elle ait tendance à générer un faible nombre de groupes contenant de nombreuses molécules [250].

Méthodes basées sur le partitionnement d'espaces chimiques

De la même manière que pour l'évaluation de la diversité, le partitionnement d'espaces chimiques peut être utilisé afin d'extraire d'une chimiothèque un sous-ensemble divers [132, 133, 223, 233, 254–256]. A partir d'un partitionnement donné et une fois que les molécules ont été positionnées dans les cellules, la sélection se fait en choisissant une ou plusieurs molécules contenues dans chacune des cellules. A nouveau, l'avantage est la simplicité et la vitesse de calcul, tandis que le problème du choix du schéma de partitionnement se pose de la même manière que pour les méthodes de quantification de la diversité. Parmi les méthodes les plus récentes, JEDA [256] utilise le partitionnement d'un espace chimique combiné à l'entropie de Shannon afin de sélectionner non seulement des molécules diverses mais également des molécules reflétant la densité de produits dans une zone donnée de l'espace chimique. Le type de bibliothèques ainsi obtenus représente alors un bon compromis entre diversité et représentativité.

Méthodes basées sur la dissimilarité

Ce type de méthodes a pour but de maximiser la dissimilarité entre les molécules contenues dans le sous-ensemble cible. La plupart des méthodes fonctionnent sur le même principe :

1. Sélectionner un premier composé et le placer dans la bibliothèque diverse
2. Calculer la dissimilarité entre chaque composé de la bibliothèque de référence et l'ensemble des composés de la bibliothèque diverse
3. Ajouter à la bibliothèque diverse le composé qui est le plus dissimilaire aux composés de l'ensemble de référence

4. Retourner à l'étape 2 si le nombre de composés sélectionnés est inférieur au nombre de composés attendus.

Il existe de nombreuses variantes de cet algorithme. Elles diffèrent principalement sur la manière dont la sélection du premier composé se fait (au hasard, en sélectionnant le composé le plus dissimilaire aux autres ou encore le composé le plus représentatif). Pour l'étape 3, l'évaluation de la dissimilarité entre un produit et un ensemble de produits peut également se faire de plusieurs manières. On peut ainsi la quantifier en calculant la somme des dissimilarités entre le produit et les molécules de référence (MaxSum) ou bien en prenant la valeur minimale de l'ensemble de ces valeurs (MaxMin). Cette dernière mesure a notamment été mise en avant pour sa capacité à sélectionner des molécules réparties de manière plus uniforme dans l'espace chimique [231, 257, 258].

Parmi les algorithmes permettant une telle sélection, la méthode MaxMin est de loin la plus utilisée. Il a notamment été montré maintes fois qu'elle donnait des résultats supérieurs aux autres méthodes de sélection [257, 259–261] et notamment aux méthodes de regroupement telles que le K-mean [228].

Reposant sur le même principe que la méthode décrite précédemment, les algorithmes d'exclusion de sphère (Sphere exclusion) ajoutent un critère supplémentaire pour sélectionner les molécules lors de l'étape 3 [132, 258]. A travers la définition d'une valeur de dissimilarité minimale par rapport aux molécules sélectionnées, certaines molécules peuvent se retrouver exclues du processus. De ce fait, ce type de méthodes ne permet pas de définir le nombre de composés à extraire.

Méthodes d'optimisation de la diversité

Ce type de méthodes repose sur l'optimisation d'une mesure de la diversité, telles que celles décrites dans la section précédente. Elles permettent théoriquement de sélectionner, parmi l'ensemble des groupes de molécules de tailles N (N étant le nombre de composés à sélectionner), le groupe qui optimisera au mieux la mesure de diversité utilisée. En pratique, puisque l'énumération exhaustive de l'ensemble des sous-groupes possibles n'est pas réalisable, des méthodes stochastiques sont utilisées afin d'approximer la solution optimale. Typiquement, les algorithmes génétiques ou le recuit simulé sont souvent utilisés à cette fin [232, 262–264].

3.6 Conclusion

Les méthodes de chimoinformatique et de modélisation moléculaire font partie des techniques récentes aujourd'hui appliquées en routine dans les phases recherche et notamment pour la gestion et l'analyse de chimiothèques. Elles suscitent certainement autant d'espoir que de critiques mais constituent néanmoins un axe de recherche très actif durant ces deux dernières décennies. Nous avons présenté dans ce chapitre la plupart des méthodes utilisées en chimoinformatique dans le cadre de l'analyse de chimiothèques. L'ensemble de ces notions ont été utilisées au cours de cette thèse afin d'une part, d'apporter notre propre contribution à travers la création de nouvelles méthodes permettant de caractériser les chimiothèques et d'autre part, de développer un logiciel permettant de mettre en oeuvre certaines de ces méthodes.

Références bibliographiques

- [1] C. ADAMS AND V. BRANTNER. *Estimating the cost of new drug development : is it really 802 million dollars ?* Health Aff (Millwood) **25**(2), 420–428 (2006). [9](#)
- [2] J. DIMASI. *The value of improving the productivity of the drug development process : faster times and better decisions.* Pharmacoeconomics **20**(3), 1–10 (2002). [9](#)
- [3] J. DIMASI, R. HANSEN, AND H. GRABOWSKI. *The price of innovation : new estimates of drug development costs.* J. Health Econ. **22**(2), 151–185 (2003). [9](#)
- [4] C. ADAMS AND V. BRANTNER. *Spending on new drug development.* Health Economics **19**(2), 130–141 (2010). [9](#)
- [5] ROBERT M. RYDZEWSKI. *Real World Drug Discovery : A Chemist's Guide to Biotech and Pharmaceutical Research.* Elsevier Science, 1 edition (2008). [9](#), [11](#), [12](#), [13](#), [14](#)
- [6] ISMAIL KOLA AND JOHN LANDIS. *Can the pharmaceutical industry reduce attrition rates ?* Nature Reviews Drug Discovery **3**(8), 711–716 (2004). [9](#)
- [7] HÉCTOR HERNÁNDEZ, FERNANDO HERVÁS, AND ALEXANDER TÜBKE. *The 2011 eu industrial r&d investment scoreboard.* publications office of the european union., (2011). [9](#)
- [8] SERGEI MASLOV AND I. ISPOLATOV. *Propagation of large concentration changes in reversible protein-binding networks.* Proceedings of the National Academy of Sciences **104**(34), 13655–13660 (2007). [10](#)
- [9] HIROAKI KITANO. *Towards a theory of biological robustness.* Molecular Systems Biology **3**(1) (2007). [10](#)
- [10] ANDREW L HOPKINS. *Network pharmacology : the next paradigm in drug discovery.* Nature chemical biology **4**(11), 682–690 (2008). [10](#)
- [11] RICARDO MACARRON, MARTYN N. BANKS, DEJAN BOJANIC, DAVID J. BURNS, DRAGAN A. CIROVIC, TINA GARYANTES, DARREN V. S. GREEN, ROBERT P. HERTZBERG, WILLIAM P. JANZEN, JEFF W. PASLAY, ULRICH SCHOPFER, AND G. SITTA SITTAMPALAM. *Impact of*

Références bibliographiques

- high-throughput screening in biomedical research*. *Nature Review Drug Discovery* **10**(3), 188–195 (2011). [vii](#), [11](#), [13](#), [14](#)
- [12] J NOAH. *New developments and emerging trends in high-throughput screening methods for lead compound identification*. *International Journal of High Throughput Screening* pages 141–149 (2010). [11](#)
- [13] LORENZ M MAYR AND DEJAN BOJANIC. *Novel trends in high-throughput screening*. *Current Opinion in Pharmacology* **9**(5), 580–588 (2009). [11](#), [13](#)
- [14] DAVID C. SWINNEY AND JASON ANTHONY. *How were new medicines discovered?* *Nature Reviews Drug Discovery* **10**(7), 507–519 (2011). [11](#), [13](#)
- [15] ADAM GOLEBIEWSKI, SEAN R KLOPFENSTEIN, AND DAVID E PORTLOCK. *Lead compounds discovered from libraries : Part 2*. *Current Opinion in Chemical Biology* **7**(3), 308–325 (2003). [11](#)
- [16] ADAM GOLEBIEWSKI, SEAN R KLOPFENSTEIN, AND DAVID E PORTLOCK. *Lead compounds discovered from libraries*. *Current Opinion in Chemical Biology* **5**(3), 273–284 (2001). [11](#)
- [17] ANN JACQUELINE HUNTER. *Have animal models of disease helped or hindered the drug discovery process ?* *Annals of the New York Academy of Sciences* **1245**(1), 1–2 (2011). [11](#)
- [18] JOANNE KOTZ. *Phenotypic screening, take two*. *SciBX : Science-Business eXchange* **5**(15) (2012). [11](#)
- [19] STEPHEN D. PICKETT. *Multi-objective approaches to screening collection design and analysis of hts data*. (2005). [12](#), [46](#)
- [20] SEAN EKINS, J DANA HONEYCUTT, AND JAMES T METZ. *Evolving molecules using multi-objective optimization : applying to ADME/Tox*. *Drug Discovery Today* **15**(11-12), 451–460 (2010). [12](#)
- [21] SCOTT J. LUSHER, ROSS MCGUIRE, RITA AZEVEDO, JAN-WILLEM BOITEN, RENE C VAN SCHAIK, AND JACOB DE Vlieg. *A molecular informatics view on best practice in multi-parameter compound optimization*. *Drug Discovery Today* **16**, 555–568 (2011). [12](#)
- [22] R P HERTZBERG AND A J POPE. *High-throughput screening : new technology for the 21st century*. *Current Opinion in Chemical Biology* **4**(4), 445–451 (2000). [13](#)
- [23] JENS SCHAMBERGER, MICHAEL GRIMM, ANDREAS STEINMEYER, AND ALEXANDER HILLISCH. *Rendezvous in chemical space ? comparing the small molecule compound libraries of bayer and schering*. *Drug Discovery Today* **16**(13–14), 636–641 (2011). [13](#)
- [24] JOHANNES H. VOIGT, BRUNO BIENFAIT, SHAOMENG WANG, AND MARC C. NICKLAUS. *Comparison of the NCI open database with seven large chemical structural databases*. *Journal of Chemical Information and Computer Sciences* **41**(3), 702–712 (2001). [13](#)
- [25] HERMAN J VERHEIJ. *Leadlikeness and structural diversity of synthetic screening libraries*. *Molecular diversity* **10**(3), 377–388 (2006). [13](#)

- [26] SUZANNE FERGUS, ANDREAS BENDER, AND DAVID R SPRING. *Assessment of structural diversity in combinatorial synthesis*. *Current opinion in chemical biology* **9**(3), 304–309 (2005). [13](#), [51](#)
- [27] AURÉLIEN MONGE, ALBAN ARRAULT, CHRISTOPHE MAROT, AND LUC MORIN-ALLORY. *Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers*. *Molecular Diversity* **10**(3), 389–403 (2006). [13](#)
- [28] MIREILLE KRIER, GUILLAUME BRET, AND DIDIER ROGNAN. *Assessing the scaffold diversity of screening libraries*. *Journal of Chemical Information and Modeling* **46**(2), 512–524 (2006). [13](#), [51](#), [52](#)
- [29] MARY P BRADLEY. *An overview of the diversity represented in commercially-available databases*. *Molecular diversity* **5**(4), 175–183 (2002). [13](#)
- [30] ALEXANDER CHUPRINA, OLEG LUKIN, ROBERT DEMOISEAUX, ALEXANDER BUZKO, AND ALEXANDER SHIVANYUK. *Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers*. *Journal of Chemical Information and Modeling* **50**(4), 470–479 (2010). [13](#)
- [31] FABIAN LOPEZ-VALLEJO, MARC A GIULIANOTTI, RICHARD A HOUGHTEN, AND JOSÉ L MEDINA-FRANCO. *Expanding the medicinally relevant chemical space with compound libraries*. *Drug Discovery Today* **17**(13-14), 718–726 (2012). [13](#)
- [32] CHRISTOPHER NEWTON. *Molecular diversity in drug design. application to high-speed synthesis and high-throughput screening*. In PHILIP DEAN AND RICHARD LEWIS, editors, *Molecular Diversity in Drug Design*, pages 23–42. Springer Netherlands (2002). [14](#)
- [33] DAVID J. PAYNE, MICHAEL N. GWYNN, DAVID J. HOLMES, AND DAVID L. POMPLIANO. *Drugs for bad bugs : confronting the challenges of antibacterial discovery*. *Nature Reviews Drug Discovery* **6**(1), 29–40 (2006). [15](#)
- [34] MATTHEW A SILLS, DONNA WEISS, QUYNHCHI PHAM, ROBERT SCHWEITZER, XIANG WU, AND JINZI J WU. *Comparison of assay technologies for a tyrosine kinase assay generates different results in high throughput screening*. *Journal of Biomolecular Screening* **7**(3), 191–214 (2002). [15](#)
- [35] GAVIN HARPER AND STEPHEN D PICKETT. *Methods for mining HTS data*. *Drug Discovery Today* **11**(15-16), 694–699 (2006). [15](#)
- [36] PHILIP GRIBBON AND ANDREAS SEWING. *Fluorescence readouts in HTS : no gain without pain ?* *Drug Discovery Today* **8**(22), 1035–1043 (2003). [15](#)
- [37] GILBERT M RISHTON. *Nonleadlikeness and leadlikeness in biochemical screening*. *Drug Discovery Today* **8**(2), 86–96 (2003). [15](#), [16](#)
- [38] W PATRICK WALTERS AND MARK NAMCHUK. *Designing screens : how to make your hits a hit*. *Nature Reviews. Drug Discovery* **2**(4), 259–266 (2003). [15](#)

Références bibliographiques

- [39] RISHTON G.M. *Reactive compounds and in vitro false positives in HTS*. Drug Discovery Today **2**(9), 382–384 (1997). [15](#)
- [40] SUSAN L. MCGOVERN, EMILIA CASELLI, NIKOLAUS GRIGORIEFF, AND BRIAN K. SHOICHET. *A common mechanism underlying promiscuous inhibitors from virtual and High-Throughput screening*. Journal of Medicinal Chemistry **45**(8), 1712–1722 (2002). [16](#)
- [41] JAMES SEIDLER, SUSAN L. MCGOVERN, THOMPSON N. DOMAN, AND BRIAN K. SHOICHET. *Identification and prediction of promiscuous aggregating inhibitors among known drugs*. Journal of Medicinal Chemistry **46**(21), 4477–4486 (2003). [16](#)
- [42] AJIT JADHAV, RAFAELA S. FERREIRA, CARLEEN KLUMPP, BRYAN T. MOTT, CHRISTOPHER P. AUSTIN, JAMES INGLESE, CRAIG J. THOMAS, DAVID J. MALONEY, BRIAN K. SHOICHET, AND ANTON SIMEONOV. *Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease*. Journal of Medicinal Chemistry **53**(1), 37–51 (2009). [16](#)
- [43] BRIAN Y. FENG, ANTON SIMEONOV, AJIT JADHAV, KERIM BABAOGLU, JAMES INGLESE, BRIAN K. SHOICHET, AND CHRISTOPHER P. AUSTIN. *A High-Throughput screen for Aggregation-Based inhibition in a large compound library*. Journal of Medicinal Chemistry **50**(10), 2385–2390 (2007). [16](#)
- [44] KERIM BABAOGLU, ANTON SIMEONOV, JOHN J. IRWIN, MICHAEL E. NELSON, BRIAN FENG, CRAIG J. THOMAS, LAURA CANCIAN, M. PAOLA COSTI, DAVID A. MALTBY, AJIT JADHAV, JAMES INGLESE, CHRISTOPHER P. AUSTIN, AND BRIAN K. SHOICHET. *Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase*. Journal of Medicinal Chemistry **51**(8), 2502–2511 (2008). [16](#)
- [45] JONATHAN B BAELL AND GEORGINA A HOLLOWAY. *New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays*. Journal of Medicinal Chemistry **53**(7), 2719–2740 (2010). [16](#)
- [46] MICHAEL S LAJINESS, MICHAL VIETH, AND JON ERICKSON. *Molecular properties that influence oral drug-like behavior*. Current Opinion in Drug Discovery & Development **7**(4), 470–477 (2004). [17](#)
- [47] INGO MUEGGE. *Selection criteria for drug-like compounds*. Medicinal Research Reviews **23**(3), 302–321 (2003). [17](#)
- [48] GIULIO VISTOLI, ALESSANDRO PEDRETTI, AND BERNARD TESTA. *Assessing drug-likeness—what are we missing?* Drug Discovery Today **13**(7-8), 285–294 (2008). [17](#)
- [49] MING-QIANG ZHANG AND BARRIE WILKINSON. *Drug discovery beyond the 'rule-of-five'*. Current Opinion in Biotechnology **18**(6), 478–488 (2007). [17](#)
- [50] PAUL D LEESON AND BRIAN SPRINGTHORPE. *The influence of drug-like concepts on decision-making in medicinal chemistry*. Nature Reviews. Drug Discovery **6**(11), 881–890 (2007). [17](#)

- [51] C A LIPINSKI, F LOMBARDO, B W DOMINY, AND P J FEENEY. *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*. Advanced Drug Delivery Reviews **46**(1-3), 3–26 (2001). [17](#)
- [52] DANIEL F. VEBER, STEPHEN R. JOHNSON, HUNG-YUAN CHENG, BRIAN R. SMITH, KEITH W. WARD, AND KENNETH D. KOPPLE. *Molecular properties that influence the oral bioavailability of drug candidates*. Journal of Medicinal Chemistry **45**(12), 2615–2623 (2002). [17](#)
- [53] MICHAEL M. HANN, ANDREW R. LEACH, AND GAVIN HARPER. *Molecular complexity and its impact on the probability of finding leads for drug discovery*. Journal of Chemical Information and Computer Sciences **41**(3), 856–864 (2001). [17](#)
- [54] T I OPREA, A M DAVIS, S J TEAGUE, AND P D LEESON. *Is there a difference between leads and drugs ? a historical perspective*. Journal of Chemical Information and Computer Sciences **41**(5), 1308–1315 (2001). [17](#)
- [55] MIKE M HANN AND TUDOR I OPREA. *Pursuing the leadlikeness concept in pharmaceutical research*. Current opinion in chemical biology **8**(3), 255–63 (2004). [17](#), [33](#)
- [56] GYÖRGY M. KESERÜ AND GERGELY M. MAKARA. *The influence of lead discovery strategies on the properties of drug candidates*. Nature Reviews Drug Discovery **8**(3), 203–212 (2009). [17](#)
- [57] MILES CONGREVE, ROBIN CARR, CHRIS MURRAY, AND HARREN JHOTI. *A 'rule of three' for fragment-based lead discovery ?* Drug Discovery Today **8**(19), 876–877 (2003). [18](#)
- [58] XAVIER MORELLI, RAPHAEL BOURGEAS, AND PHILIPPE ROCHE. *Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I)*. Current Opinion in Chemical Biology **15**(4), 475–481 (2011). [18](#)
- [59] ANDREW L. HOPKINS, COLIN R. GROOM, AND ALEXANDER ALEX. *Ligand efficiency : a useful metric for lead selection*. Drug Discovery Today **9**(10), 430–431 (2004). [18](#)
- [60] CELE ABAD-ZAPATERO AND JAMES T METZ. *Ligand efficiency indices as guideposts for drug discovery*. Drug Discovery Today **10**(7), 464–469 (2005). [18](#)
- [61] CHARLES H REYNOLDS, SCOTT D BEMBENEK, AND BRETT A TOUNGE. *The role of molecular size in ligand efficiency*. Bioorganic & Medicinal Chemistry letters **17**(15), 4258–4261 (2007). [18](#)
- [62] MASAYA ORITA, KAZUKI OHNO, AND TATSUYA NIIMI. *Two 'Golden ratio' indices in fragment-based drug discovery*. Drug Discovery Today **14**(5-6), 321–328 (2009). [18](#)
- [63] J WILLEM M NISSINK. *Simple size-independent measure of ligand efficiency*. Journal of Chemical Information and Modeling **49**(6), 1617–1622 (2009). [18](#)
- [64] PAUL N MORTENSON AND CHRISTOPHER W MURRAY. *Assessing the lipophilicity of fragments and early hits*. Journal of Computer-Aided Molecular Design **25**(7), 663–667 (2011). [18](#)
- [65] MARK A. JOHNSON AND GERALD M. MAGGIORA, editors. *Concepts and Applications of Molecular Similarity*. Wiley-Interscience, 1 edition (1990). [19](#)

Références bibliographiques

- [66] D E PATTERSON, R D CRAMER, A M FERGUSON, R D CLARK, AND L E WEINBERGER. *Neighborhood behavior : a useful concept for validation of "molecular diversity" descriptors*. Journal of Medicinal Chemistry **39**(16), 3049–3059 (1996). [19](#), [46](#)
- [67] DRAGOS HORVATH AND BORYEU MAO. *Neighborhood behavior. fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity*. QSAR and Combinatorial Science **22**(5), 498–509 (2003). [19](#)
- [68] YVONNE C. MARTIN, JAMES L. KOFRON, AND LINDA M. TRAPHAGEN. *Do structurally similar molecules have similar biological activity?* Journal of Medicinal Chemistry **45**(19), 4350–4358 (2002). [19](#)
- [69] ANDREAS BENDER AND ROBERT C GLEN. *Molecular similarity : a key technique in molecular informatics*. Org. Biomol. Chem. **2**(22), 3204–3218 (2004). [19](#)
- [70] GEORGE PAPADATOS, ANTHONY W J COOPER, VISAKAN KADIRKAMANATHAN, SIMON J F MACDONALD, IAIN M MCLAY, STEPHEN D PICKETT, JOHN M PRITCHARD, PETER WILLETT, AND VALERIE J GILLET. *Analysis of neighborhood behavior in lead optimization and array design*. Journal of Chemical Information and Modeling **49**(2), 195–208 (2009). [19](#)
- [71] PETER WILLETT. *Similarity searching using 2D structural fingerprints*. Methods in Molecular Biology **672**, 133–158 (2011). [19](#)
- [72] P. WILLETT, J. M. BARNARD, AND G. M. DOWNS. *Chemical similarity searching*. Journal of Chemical Information and Computer Sciences **38**, 983–996 (1998). [19](#)
- [73] THORSTEN POTTER AND HANS MATTER. *Random or rational design ? evaluation of diverse compound subsets from chemical structure databases*. Journal of Medicinal Chemistry **41**(4), 478–488 (1998). [19](#), [47](#)
- [74] SIEW KUEN YEAP, ROSALIND J WALLEY, MIKE SNAREY, WILLEM P VAN HOORN, AND JONATHAN S MASON. *Designing compound subsets : comparison of random and rational approaches using statistical simulation*. Journal of Chemical Information and Modeling **47**(6), 2149–2158 (2007). [19](#), [47](#)
- [75] THOMAS J CRISMAN, JEREMY L JENKINS, CHRISTIAN N PARKER, W ADAM G HILL, ANDREAS BENDER, ZHAN DENG, JAMES H NETTLES, JOHN W DAVIES, AND MEIR GLICK. *"Plate cherry picking" : a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection*. Journal of biomolecular screening **12**(3), 320–327 (2007). [19](#), [47](#)
- [76] RICHARD J SPANDL, ANDREAS BENDER, AND DAVID R SPRING. *Diversity-oriented synthesis ; a spectrum of approaches and results*. Organic & biomolecular chemistry **6**(7), 1149–1158 (2008). [19](#), [47](#)
- [77] SAI CHETAN K SUKURU, JEREMY L JENKINS, ROHAN E J BECKWITH, JOSEF SCHEIBER, ANDREAS BENDER, DMITRI MIKHAILOV, JOHN W DAVIES, AND MEIR GLICK. *Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity*. Journal of biomolecular screening **14**(6), 690–699 (2009). [19](#), [47](#)

- [78] SIMONE DI MICCO, ROMINA VITALE, MAURIZIO PELLECCIA, MICHELE F REGA, RENATA RIVA, ANDREA BASSO, AND GIUSEPPE BIFULCO. *Identification of lead compounds as antagonists of protein bcl-xL with a diversity-oriented multidisciplinary approach*. Journal of Medicinal Chemistry **52**(23), 7856–7867 (2009). [19](#), [47](#)
- [79] WARREN R J D GALLOWAY, ANDREAS BENDER, MARTIN WELCH, AND DAVID R SPRING. *The discovery of antibacterial agents using diversity-oriented synthesis*. Chemical communications (Cambridge, England) (18), 2446–2462 (2009). [19](#), [47](#)
- [80] WARREN R J D GALLOWAY, ALBERT ISIDRO-LLOBET, AND DAVID R SPRING. *Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules*. Nature communications **1**, 80 (2010). [19](#), [45](#), [47](#)
- [81] ROSS L STEIN. *High-throughput screening in academia : the harvard experience*. Journal of Biomolecular Screening **8**(6), 615–619 (2003). PMID : 14711386. [20](#)
- [82] JULIE A FREARSON AND IAIN T COLLIE. *HTS and hit finding in academia—from chemical genomics to drug discovery*. Drug Discovery Today **14**(23-24), 1150–1158 (2009). [20](#)
- [83] PHILIP GRIBBON. *High-throughput hit finding and compound-profiling technologies for academic drug discovery*. Drug Discovery Today **5**(1), e3–e7 (2009). [20](#)
- [84] MARCEL F. HIBERT. *French/European academic compound library initiative*. Drug Discovery Today **14**(15–16), 723–725 (2009). [20](#)
- [85] ANURADHA ROY, PETER R McDONALD, SITTA SITTAMPALAM, AND RATHNAM CHAGUTURU. *Open access high throughput drug discovery in the public domain : a mount everest in the making*. Current Pharmaceutical Biotechnology **11**(7), 764–778 (2010). [20](#)
- [86] A list of existing screening centers. <http://www.slas.org/screeningFacilities/facilityList.cfm>. [20](#)
- [87] RACHEL L CLARK, BLAIR F JOHNSTON, SIMON P MACKAY, CATHERINE J BRESLIN, MURRAY N ROBERTSON, AND ALAN L HARVEY. *The drug discovery portal : a resource to enhance drug discovery from academia*. Drug Discovery Today **15**(15-16), 679–683 (2010). [20](#)
- [88] The queensland compound library was established to synergise interactions between australasian chemists, biomedical researchers, and their international colleagues. <http://www.griffith.edu.au/science-aviation/queensland-compound-library>. [20](#)
- [89] CHRISTOPHER P AUSTIN, LINDA S BRADY, THOMAS R INSEL, AND FRANCIS S COLLINS. *NIH molecular libraries initiative*. Science (New York, N.Y.) **306**(5699), 1138–1139 (2004). [20](#)
- [90] JEFFREY H. TONEY. *Big payoffs possible for small molecule screening*. Science **322**(5898), 46–46 (2008). [20](#)
- [91] EVAN E. BOLTON, YANLI WANG, PAUL A. THIESSEN, AND STEPHEN H. BRYANT. PubChem : Integrated platform of small molecules and biological activities. , **4**, pages 217–241. Elsevier (2008). [20](#), [24](#), [39](#)

Références bibliographiques

- [92] JOCELYN KAISER. *Industrial-style screening meets academic biology*. Science **321**(5890), 764–766 (2008). [21](#)
- [93] JOCELYN KAISER. *Drug-screening program looking for a home*. Science **334**(6054), 299–299 (2011). [21](#)
- [94] MICHAEL F. LYNCH, JUDITH M. HARRISON, AND WILLIAM G. TOWN. *Computer handling of chemical structure information*. Macdonald and Co. (1971). [21](#)
- [95] ANDREW R. LEACH AND VALERIE J. GILLET. *An Introduction to Chemoinformatics*. Springer (2003). [21](#)
- [96] JOHANN GASTEIGER, editor. *Handbook of Chemoinformatics : From Data to Knowledge (Representation of Molecular Structures)*. John Wiley & Sons, 1 edition (2003). [21](#), [22](#)
- [97] RAJARSHI GUHA AND ANDREAS BENDER. *Computational Approaches in Cheminformatics and Bioinformatics*. Wiley-Blackwell (an imprint of John Wiley & Sons Ltd) (2012). [21](#)
- [98] ALEXANDRE VARNEK AND ALEX TROPSHA. *Chemoinformatics approaches to virtual screening*. Royal Society of Chemistry (2008). [21](#)
- [99] JOHANN GASTEIGER AND THOMAS ENGEL. *Chemoinformatics : A Textbook*. John Wiley & Sons (2007). [21](#)
- [100] BARRY A. BUNIN. *Chemoinformatics : theory, practice, & products*. Springer (2007). [21](#)
- [101] TUDOR I. OPREA, RAIMUND MANNHOLD, HUGO KUBINYI, AND GERD FOLKERS. *Chemoinformatics in Drug Discovery*. John Wiley & Sons (2006). [21](#)
- [102] J. BAJORATH. *Chemoinformatics : concepts, methods, and tools for drug discovery*. Humana Press (2004). [21](#)
- [103] JOHANN GASTEIGER. *Chemoinformatics : a new field with a long tradition*. Analytical and Bioanalytical Chemistry **384**(1), 57–64 (2006). [21](#)
- [104] WILLIAM LINGRAN CHEN. *Chemoinformatics : Past, present, and future*. Journal of Chemical Information and Modeling **46**(6), 2230–2255 (2006). [21](#)
- [105] THOMAS ENGEL. *Basic overview of chemoinformatics*. Journal of Chemical Information and Modeling **46**(6), 2267–2277 (2006). [21](#)
- [106] DIMITRIS K AGRAFIOTIS, DEEPAK BANDYOPADHYAY, JÖRG K WEGNER, AND HERMAN VAN VLIJMEN. *Recent advances in chemoinformatics*. Journal of Chemical Information and Modeling **47**(4), 1279–1293 (2007). [21](#)
- [107] ALEXANDRE VARNEK AND IGOR I. BASKIN. *Chemoinformatics as a theoretical chemistry discipline*. Molecular Informatics **30**(1), 20–32 (2011). [21](#)
- [108] JÜRGEN BAJORATH. *Chemoinformatics : Recent advances at the interfaces between computer and chemical information sciences, chemistry, and drug discovery*. Bioorganic & Medicinal Chemistry **20**(18), 5316 (2012). [21](#)

- [109] MARTIN VOGT AND JÜRGEN BAJORATH. *Chemoinformatics : A view of the field and current trends in method development*. Bioorganic & Medicinal Chemistry **20**(18), 5317–5323 (2012). 21
- [110] ABRAHAM DJ, editor. *History of Quantitative Structure-Activity Relationships*. Wiley, 1 edition (2003). 22
- [111] CORWIN. HANSCH AND TOSHIO. FUJITA. $\rho - \sigma - \pi$ analysis. a method for the correlation of biological activity and chemical structure. Journal of the American Chemical Society **86**(8), 1616–1626 (1964). 22
- [112] ARKADIUSZ Z DUDEK, TOMASZ ARODZ, AND JORGE GÁLVEZ. *Computational methods in developing quantitative structure-activity relationships (QSAR) : a review*. Combinatorial Chemistry & High Throughput Screening **9**(3), 213–228 (2006). 22
- [113] FRANK K BROWN. Chemoinformatics what is it and how does it impact drug discovery. , **Volume 33**, chapter Chapter 35, pages 375–384. Academic Press (1998). 22
- [114] PARIS G. August 1999 meeting of the american chemical society, quoted by w. warr. 22
- [115] M HANN AND R GREEN. *Chemoinformatics—a new name for an old problem ?* Current opinion in chemical biology **3**(4), 379–383 (1999). 22
- [116] Définition de la chémoinformatique donnée lors du colloque Å obernais, juin 2006. <http://www.chemoinformatique.fr/modules/smartsection/item.php?itemid=4>. 22
- [117] JOHN KINNEY AND MARK HERMSMEIER. Validation and characterization of chemical structures derived from names and images in scientific documents, (2011). 23
- [118] LOUISA J BELLIS, RUTH AKHTAR, BISSAN AL-LAZIKANI, FRANCIS ATKINSON, A PATRICIA BENTO, JON CHAMBERS, MARK DAVIES, ANNA GAULTON, ANNE HERSEY, KAZUYOSHI IKEDA, FELIX A KRÜGER, YVONNE LIGHT, SHAUN MCGLINCHY, RITA SANTOS, BENJAMIN STAUCH, AND JOHN P OVERINGTON. *Collation and data-mining of literature bioactivity data for drug discovery*. Biochemical Society transactions **39**(5), 1365–1370 (2011). vii, 24, 26
- [119] DAVID WENINGER. *Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules*. Journal of Chemical Information and Computer Sciences **28**(1), 31–36 (1988). 24
- [120] An open standard for chemical structure representation - the iupac chemical identifier. <http://stage.iupac.org/inchi/Stein-2003-ref1.html>. 24
- [121] The iupac international chemical identifier (inchi). <http://www.iupac.org/home/publications/e-resources/inchi.html>. 24
- [122] STEPHEN HELLER AND ALAN MCNAUGHT. *The status of the InChI project and the InChI trust*. Journal of Cheminformatics **2**(Suppl 1), P2 (2010). 24
- [123] Daylight chemical information systems manual : Smarts - a language for describing molecular patterns. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. 25

Références bibliographiques

- [124] ARTHUR DALBY, JAMES G. NOURSE, W. DOUGLAS HOUNSHELL, ANN K. I. GUSHURST, DAVID L. GRIER, BURTON A. LELAND, AND JOHN LAUFER. *Description of several chemical structure file formats used by computer programs developed at molecular design limited*. Journal of Chemical Information and Computer Sciences **32**(3), 244–255 (1992). 25
- [125] Descriptif du format mol (sur demande). <http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php>. 25
- [126] ROBERTO TODESCHINI AND VIVIANA CONSONNI. *Handbook of Molecular Descriptors*. John Wiley & Sons (2008). 27
- [127] LOWELL H. HALL, LEMONT B. KIER, AND BRISCOE B. BROWN. *Molecular similarity based on novel atom-type electrotopological state indices*. Journal of Chemical Information and Computer Sciences **35**(6), 1074–1080 (1995). 27
- [128] LOWELL H. HALL AND LEMONT B. KIER. *Electrotopological state indices for atom types : A novel combination of electronic, topological, and valence state information*. Journal of Chemical Information and Computer Sciences **35**(6), 1039–1045 (1995). 27
- [129] RANDIC MILAN. *Characterization of molecular branching*. Journal of the American Chemical Society **97**(23), 6609–6615 (1975). 27
- [130] GONZALO CERRUELA GARCIA, IRENE LUQUE RUIZ, MIGUEL ANGEL GOMEZ-NIETO, JUAN ANTONIO CABRERO DONCEL, AND ANTONIO GUEVARA PLAZA. *From wiener index to molecules*. Journal of Chemical Information and Modeling **45**(2), 231–238 (2005). 27
- [131] JONATHAN S. MASON, ISABELLE MORIZE, PAUL R. MENARD, DANIEL L. CHENEY, CHRISTOPHER HULME, AND RICHARD F. LABAUDINIERE. *New 4-point pharmacophore method for molecular similarity and diversity applications : Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures*. Journal of Medicinal Chemistry **42**(17), 3251–3264 (1999). 27, 29
- [132] ROBERT S. PEARLMAN AND K.M. SMITH. *Novel software tools for chemical diversity*. Perspectives in Drug Discovery and Design **9-11**(0), 339–353 (1998). 28, 49, 53, 54
- [133] ALAIN-DOMINIQUE GORSE. *Diversity in medicinal chemistry space*. Current topics in medicinal chemistry **6**(1), 3–18 (2006). 28, 45, 52, 53
- [134] Mdl information system inc. <http://www.mdli.com/>. 29
- [135] JUSTIN KLEKOTA AND FREDERICK P ROTH. *Chemical substructures that enrich for biological activity*. Bioinformatics **24**(21), 2518–2525 (2008). 29
- [136] DAVID ROGERS AND MATHEW HAHN. *Extended-connectivity fingerprints*. Journal of Chemical Information and Modeling **50**(5), 742–754 (2010). 29
- [137] LI XING AND ROBERT C GLEN. *Novel methods for the prediction of logP, pK(a), and logD*. Journal of Chemical Information and Computer Sciences **42**(4), 796–805 (2002). 29

- [138] ANDREAS BENDER, HAMSE Y MUSSA, ROBERT C GLEN, AND STEPHAN REILING. *Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier*. Journal of Chemical Information and Computer Sciences **44**(1), 170–178 (2004). 29
- [139] ROBERT C GLEM, ANDREAS BENDER, CATRIN H ARNBY, LARS CARLSSON, SCOTT BOYER, AND JAMES SMITH. *Circular fingerprints : flexible molecular descriptors with applications from physical chemistry to ADME*. IDrugs : the Investigational Drugs Journal **9**(3), 199–204 (2006). 29
- [140] RAYMOND E. CARHART, DENNIS H. SMITH, AND R. VENKATARAGHAVAN. *Atom pairs as molecular features in structure-activity studies : definition and applications*. Journal of Chemical Information and Computer Sciences pages 64–73 (1985). 29
- [141] PIERRE MAHÉ, LIVA RALAIVOLA, VÉRONIQUE STOVEN, AND JEAN-PHILIPPE VERT. *The pharmacophore kernel for virtual screening with support vector machines*. Journal of Chemical Information and Modeling **46**(5), 2003–2014 (2006). 29
- [142] Moe, version 2009.10 ; chemical computing group (ccg) : Montreal, canada, 2009. <http://www.chemcomp.com/software.html>. 29
- [143] FANNY BONACHÉRA, BENJAMIN PARENT, FRÉDÉRIQUE BARBOSA, NICOLAS FROLOFF, AND DRAGOS HORVATH. *Fuzzy tricentric pharmacophore fingerprints. 1. topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes*. Journal of Chemical Information and Modeling **46**(6), 2457–2477 (2006). 29
- [144] FANNY BONACHÉRA AND DRAGOS HORVATH. *Fuzzy tricentric pharmacophore fingerprints. 2. application of topological fuzzy pharmacophore triplets in quantitative structure-activity relationships*. Journal of Chemical Information and Modeling **48**(2), 409–425 (2008). 29
- [145] A VARNEK, D FOURCHES, F HOONAKKER, AND V P SOLOV'EV. *Substructural fragments : an universal language to encode reactions, molecular and supramolecular structures*. Journal of Computer-Aided Molecular Design **19**(9-10), 693–703 (2005). 30
- [146] IGOR BASKIN AND ALEXANDRE VARNEK. *Building a chemical space based on fragment descriptors*. Combinatorial Chemistry & High Throughput Screening **11**(8), 661–668 (2008). 30
- [147] ALEXANDRE VARNEK. *Fragment descriptors in structure-property modeling and virtual screening*. Methods in Molecular Biology **672**, 213–243 (2011). 30
- [148] A VARNEK, D FOURCHES, D HORVATH, O KLIMCHUK, C GAUDIN, P VAYER, V SOLOV'EV, F HOONAKKER, IV TETKO, AND G MARCOU. *Isida - platform for virtual screening based on fragment and pharmacophoric descriptors*. Methods in Molecular Biology **4**(3), 191–198 (2008). 30
- [149] G W BEMIS AND M A MURCKO. *The properties of known drugs. 1. molecular frameworks*. Journal of medicinal chemistry **39**(15), 2887–2893 (1996). 32

Références bibliographiques

- [150] VALERIE J. GILLET, GEOFFREY M. DOWNS, JOHN D. HOLLIDAY, MICHAEL F. LYNCH, AND WINFRIED DETHLEFSEN. *Computer storage and retrieval of generic chemical structures in patents. 13. reduced graph generation*. Journal of Chemical Information and Computer Sciences **31**(2), 260–270 (1991). [32](#)
- [151] KRISTIAN BIRCHALL AND VALERIE J GILLET. *Reduced graphs and their applications in chemoinformatics*. Methods in Molecular Biology **672**, 197–212 (2011). [32](#)
- [152] G. HARPER, G. S. BRAVI, S. D. PICKETT, J. HUSSAIN, AND D. V. S. GREEN. *The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data*. Journal of Chemical Information and Computer Sciences **44**(6), 2145–2156 (2004). [32](#)
- [153] KRISTIAN BIRCHALL, VALERIE J GILLET, GAVIN HARPER, AND STEPHEN D PICKETT. *Evolving interpretable structure-activity relationship models. 2. using multiobjective optimization to derive multiple models*. Journal of Chemical Information and Modeling **48**(8), 1558–1570 (2008). [32](#)
- [154] KRISTIAN BIRCHALL, VALERIE J GILLET, GAVIN HARPER, AND STEPHEN D PICKETT. *Evolving interpretable structure-activity relationships. 1. reduced graph queries*. Journal of Chemical Information and Modeling **48**(8), 1543–1557 (2008). [32](#)
- [155] KRISTIAN BIRCHALL, VALERIE J GILLET, PETER WILLETT, PIERRE DUCROT, AND CLAUDE LUTTMANN. *Use of reduced graphs to encode bioisosterism for similarity-based virtual screening*. Journal of Chemical Information and Modeling **49**(6), 1330–1346 (2009). [32](#)
- [156] CHRISTOPHER M. DOBSON. *Chemical space and biology*. Nature **432**(7019), 824–828 (2004). [33](#)
- [157] CHRISTOPHER LIPINSKI AND ANDREW HOPKINS. *Navigating chemical space for biology and medicine*. Nature **432**(7019), 855–861 (2004). [33](#)
- [158] BRENT R. STOCKWELL. *Exploring biology with small organic molecules*. Nature **432**(7019), 846–854 (2004). [33](#)
- [159] RONALD R. BREAKER. *Natural and engineered nucleic acids as tools to explore biology*. Nature **432**(7019), 838–845 (2004). [33](#)
- [160] JON CLARDY AND CHRISTOPHER WALSH. *Lessons from natural molecules*. Nature **432**(7019), 829–837 (2004). [33](#)
- [161] PETER KIRKPATRICK AND CLARE ELLIS. *Chemical space*. Nature **432**(7019), 823–823 (2004). [33](#), [34](#)
- [162] REGINE S. BOHACEK, COLIN McMARTIN, AND WAYNE C. GUIDA. *The art and practice of structure-based drug design : A molecular modeling perspective*. Medicinal Research Reviews **16**(1), 3–50 (1996). [33](#)
- [163] PETER ERTL. *Cheminformatics analysis of organic substituents : identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups*. Journal of Chemical Information and Computer Sciences **43**(2), 374–380 (2003). [33](#)

- [164] DHARMESH M. MANIYAR, IAN T. NABNEY, BRUCE S. WILLIAMS, AND ANDREAS SEWING. *Data visualization during the early stages of drug discovery*. Journal of Chemical Information and Modeling **46**(4), 1806–1818 (2006). [36](#), [43](#)
- [165] TREVOR J HOWE, GUY MAHIEU, PATRICK MARICHAL, TOM TABRUYN, AND PIETER VUGTS. *Data reduction and representation in drug discovery*. Drug Discovery Today **12**(1-2), 45–53 (2007). [36](#)
- [166] JOSE L. MEDINA-FRANCO, KARINA MARTINEZ-MAYORGA, MARC A. GIULIANOTTI, RICHARD A. HOUGHTEN, AND CLEMENCIA PINILLA. *Visualization of the chemical space in drug discovery*. Current Computer Aided-Drug Design **4**(4), 322–333 (2008). [36](#)
- [167] PETER ERTL AND BERNHARD ROHDE. *The molecule cloud - compact visualization of large collections of molecules*. Journal of Cheminformatics **4**(1), 12 (2012). [36](#)
- [168] TIMOTHY J RITCHIE, PETER ERTL, AND RICHARD LEWIS. *The graphical representation of ADME-related molecule properties for medicinal chemists*. Drug Discovery Today **16**(1-2), 65–72 (2011). [36](#)
- [169] T I OPREA, J GOTTFRIES, V SHERBUKHIN, P SVENSSON, AND T C KÜHLER. *Chemical information management in drug discovery : optimizing the computational and combinatorial chemistry interfaces*. Journal of molecular graphics & modelling (4-5), 512–524, 541 (2000). [37](#)
- [170] T I OPREA AND J GOTTFRIES. *Chemography : the art of navigating in chemical space*. Journal of Combinatorial Chemistry **3**(2), 157–166 (2001). [37](#)
- [171] TUDOR I OPREA, ISMAEL ZAMORA, AND ANNA-LENA UNGELL. *Pharmacokinetically based mapping device for chemical space navigation*. Journal of Combinatorial Chemistry **4**(4), 258–266 (2002). [37](#)
- [172] GABRIELE CRUCIANI, MANUEL PASTOR, AND WOLFGANG GUBA. *VolSurf : a new tool for the pharmacokinetic optimization of lead compounds*. European Journal of Pharmaceutical Sciences **11**, Supplement **2**(0), S29–S39 (2000). [37](#)
- [173] JOSEFIN LARSSON, JOHAN GOTTFRIES, LARS BOHLIN, AND ANDERS BACKLUND. *Expanding the ChemGPS chemical space with natural products*. Journal of Natural Products **68**(7), 985–991 (2005). [37](#)
- [174] JOSEFIN LARSSON, JOHAN GOTTFRIES, SOREL MURESAN, AND ANDERS BACKLUND. *ChemGPS-NP : tuned for navigation in biologically relevant chemical space*. Journal of Natural Products **70**(5), 789–794 (2007). [37](#)
- [175] JOSEFIN ROSÉN, ANDERS LÖVGREN, THIERRY KOGEJ, SOREL MURESAN, JOHAN GOTTFRIES, AND ANDERS BACKLUND. *ChemGPS-NP(Web) : chemical space navigation online*. Journal of Computer-Aided Molecular Design **23**(4), 253–259 (2009). [37](#)
- [176] JOSEFIN ROSÉN, JOHAN GOTTFRIES, SOREL MURESAN, ANDERS BACKLUND, AND TUDOR I OPREA. *Novel chemical space exploration via natural products*. Journal of Medicinal Chemistry (7), 1953–1962 (2009). [37](#)

Références bibliographiques

- [177] M OLAH, R RAD, L OSTOPOVICI, A BORA, N HADARUGA, D HADARUGA, R MOLDOVAN, A FULIAS, M MRACEC, AND TI OPREA. WOMBAT and WOMBAT-PK : bioactive databases for lead and drug discovery. In SL SCHREIBER, TM KAPOOR, AND G WESS, editors, *Chemical Biology : From Small Molecules to Systems Biology and Drug Design*, pages 760–786. Wiley-VCH (2007). [37](#)
- [178] HUGO LACHANCE, STEFAN WETZEL, KAMAL KUMAR, AND HERBERT WALDMANN. *Charting, navigating, and populating natural product chemical space for drug discovery*. Journal of Medicinal Chemistry **55**(13), 5989–6001 (2012). [37](#)
- [179] MIKLOS FEHER AND JONATHAN M SCHMIDT. *Property distributions : differences between drugs, natural products, and molecules from combinatorial chemistry*. Journal of Chemical Information and Computer Sciences **43**(1), 218–227 (2003). [38](#)
- [180] NARENDER SINGH, RAJARSHI GUHA, MARC A GIULIANOTTI, CLEMENCIA PINILLA, RICHARD A HOUGHTEN, AND JOSE L MEDINA-FRANCO. *Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository*. Journal of Chemical Information and Modeling **49**(4), 1010–1024 (2009). [38](#)
- [181] DAVID G LLOYD, GEORGIA GOLFIS, ANDREW J S KNOX, DARREN FAYNE, MARY J MEEGAN, AND TUDOR I OPREA. *Oncology exploration : charting cancer medicinal chemistry space*. Drug discovery today **11**(3-4), 149–159 (2006). [38](#)
- [182] TOBIAS FINK AND JEAN-LOUIS REYMOND. *Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f : Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery*. Journal of Chemical Information and Modeling **47**(2), 342–353 (2007). [vii](#), [38](#), [42](#)
- [183] RUUD VAN DEURSEN, LORENZ C BLUM, AND JEAN-LOUIS REYMOND. *A searchable map of PubChem*. Journal of Chemical Information and Modeling **50**(11), 1924–1934 (2010). [vii](#), [39](#)
- [184] LORENZ C. BLUM AND JEAN-LOUIS REYMOND. *970 million druglike small molecules for virtual screening in the chemical universe database GDB-13*. Journal of the American Chemical Society **131**(25), 8732–8733 (2009). [39](#)
- [185] LORENZ C BLUM, RUUD VAN DEURSEN, AND JEAN-LOUIS REYMOND. *Visualisation and subsets of the chemical universe database GDB-13 for virtual screening*. Journal of Computer-Aided Molecular Design **25**(7), 637–647 (2011). [39](#)
- [186] RUUD VAN DEURSEN, LORENZ C BLUM, AND JEAN-LOUIS REYMOND. *Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem*. Journal of Computer-Aided Molecular Design **25**(7), 649–662 (2011). [39](#)
- [187] TEUVO KOHONEN. *Self-organized formation of topologically correct feature maps*. Biological Cybernetics **43**(1), 59–69 (1982). [40](#)
- [188] DANIELA DIGLES AND GERHARD F. ECKER. *Self-organizing maps for in silico screening and data visualization*. Molecular Informatics **30**(10), 838–846 (2011). [41](#)

- [189] SOHEILA ANZALI, JOHANN GASTEIGER, ULRIKE HOLZGRABE, JAROSLAW POLANSKI, JENS SADOWSKI, ANDREAS TECKENTRUP, AND MARKUS WAGENER. *The use of self-organizing neural networks in drug design*. Perspectives in Drug Discovery and Design **9-11**(0), 273–299 (1998). [41](#)
- [190] JURE ZUPAN AND JOHANN GASTEIGER. *Neural Networks in Chemistry and Drug Design : An Introduction*. John Wiley & Sons (1999). [41](#)
- [191] DOMINIK KAISER, LOTHAR TERFLOTH, STEPHAN KOPP, JAN SCHULZ, RANDOLF DE LAET, PETER CHIBA, GERHARD F ECKER, AND JOHANN GASTEIGER. *Self-organizing maps for identification of new inhibitors of p-glycoprotein*. Journal of Medicinal Chemistry **50**(7), 1698–1702 (2007). [41](#)
- [192] YONG-HUA WANG, YAN LI, SHENG-LI YANG, AND LING YANG. *Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach*. Journal of Chemical Information and Modeling **45**(3), 750–757 (2005). [41](#)
- [193] KONSTANTIN V BALAKIN, SEAN EKINS, ANDREY BUGRIM, YAN A IVANENKOV, DMITRY KOROLEV, YURI V NIKOLSKY, ANDREY V SKORENKO, ANDREY A IVASHCHENKO, NIKOLAY P SAVCHUK, AND TATIANA NIKOLSKAYA. *Kohonen maps for prediction of binding to human cytochrome p450 3A4*. Drug metabolism and disposition : the biological fate of chemicals **32**(10), 1183–1189 (2004). [41](#)
- [194] DMITRIY S CHEKMAREV, VLADYSLAV KHOLODOVYCH, KONSTANTIN V BALAKIN, YAN IVANENKOV, SEAN EKINS, AND WILLIAM J WELSH. *Shape signatures : new descriptors for predicting cardiotoxicity in silico*. Chemical research in toxicology **21**(6), 1304–1314 (2008). [41](#)
- [195] SEAN EKINS, KONSTANTIN V BALAKIN, NIKOLAY SAVCHUK, AND YAN IVANENKOV. *Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and kohonen and sammon mapping techniques*. Journal of Medicinal Chemistry **49**(17), 5059–5071 (2006). [41](#)
- [196] SHINNOSUKE HIDAKA, HIROYUKI YAMASAKI, YOSHIHIRO OHMAYU, AKIKO MATSUURA, KOSUKE OKAMOTO, NORIHITO KAWASHITA, AND TATSUYA TAKAGI. *Nonlinear classification of hERG channel inhibitory activity by unsupervised classification method*. The Journal of toxicological sciences **35**(3), 393–399 (2010). [41](#)
- [197] MAN-LING LEE AND GISBERT SCHNEIDER. *Scaffold architecture and pharmacophoric properties of natural products and trade drugs application in the design of natural product-based combinatorial libraries*. Journal of Combinatorial Chemistry **3**(3), 284–289 (2001). [41](#)
- [198] KRISTINA GRABOWSKI, KARL-HEINZ BARINGHAUS, AND GISBERT SCHNEIDER. *Scaffold diversity of natural products : inspiration for combinatorial library design*. Natural product reports **25**(5), 892–904 (2008). [41](#)
- [199] PETRA SCHNEIDER AND GISBERT SCHNEIDER. *Collection of bioactive reference compounds for focused library design*. QSAR & Combinatorial Science **22**(7), 713–718 (2003). [41](#)

Références bibliographiques

- [200] GISBERT SCHNEIDER, WERNER NEIDHART, THOMAS GILLER, AND GERARD SCHMID. *"Scaffold-Hopping" by topological pharmacophore search : A contribution to virtual screening*. Angewandte Chemie International Edition **38**(19), 2894–2896 (1999). [41](#)
- [201] PETER ERTL, STEPHEN JELFS, JÖRG MÜHLBACHER, ANSGAR SCHUFFENHAUER, AND PAUL SELZER. *Quest for the rings. in silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds*. Journal of Medicinal Chemistry **49**(15), 4568–4573 (2006). [41](#)
- [202] CHRISTOPHER M. BISHOP AND CHRISTOPHER K. I. WILLIAMS. GTM : a principled alternative to the self-organizing map. In *In Advances in Neural Information Processing Systems*, pages 354–360. Springer-Verlag (1997). [42](#)
- [203] CHRISTOPHER M. BISHOP, MARKUS SVENSÉN, AND CHRISTOPHER K. I. WILLIAMS. *GTM : the generative topographic mapping*. Neural Computation **10**(1), 215–234 (1998). [42](#)
- [204] DAVID LOWE AND MICHAEL E. TIPPING. *NeuroScale : Novel Topographic Feature Extraction using RBF Networks*. (1997). [43](#)
- [205] JR. SAMMON, J.W. *A nonlinear mapping for data structure analysis*. IEEE Transactions on Computers **C-18**(5), 401–409 (1969). [43](#), [44](#)
- [206] JOHN R. OWEN, IAN T. NABNEY, JOSEÁ½ L. MEDINA-FRANCO, AND FABIAN LOÁ½PEZ-VALLEJO. *Visualization of molecular fingerprints*. Journal of Chemical Information and Modeling **51**(7), 1552–1563 (2011). [43](#)
- [207] N. KIREEVA, I. I. BASKIN, H. A. GASPARD, D. HORVATH, G. MARCOU, AND A. VARNEK. *Generative topographic mapping (GTM) : universal tool for data visualization, structure-activity modeling and dataset comparison*. Molecular Informatics **31**(3-4), 301–312 (2012). [43](#)
- [208] NIU HUANG, BRIAN K. SHOICHET, AND JOHN J. IRWIN. *Benchmarking sets for molecular docking*. Journal of Medicinal Chemistry **49**(23), 6789–6801 (2006). [43](#)
- [209] DIMITRIS K. AGRAFIOTIS AND HUAFENG XU. *A self-organizing principle for learning nonlinear manifolds*. Proceedings of the National Academy of Sciences **99**(25), 15869–15872 (2002). [43](#)
- [210] DIMITRIS K. AGRAFIOTIS. *Stochastic proximity embedding*. Journal of Computational Chemistry **24**(10), 1215–1221 (2003). [43](#)
- [211] DIMITRIS K. AGRAFIOTIS, HUAFENG XU, FANGQIANG ZHU, DEEPAK BANDYOPADHYAY, AND PU LIU. *Stochastic proximity embedding : Methods and applications*. Molecular Informatics **29**(11), 758–770 (2010). [43](#), [44](#)
- [212] DMITRII N RASSOKHIN AND DIMITRIS K AGRAFIOTIS. *A modified update rule for stochastic proximity embedding*. Journal of molecular graphics & modelling **22**(2), 133–140 (2003). [44](#)
- [213] HUAFENG XU, SERGEI IZRAILEV, AND DIMITRIS K AGRAFIOTIS. *Conformational sampling by self-organization*. Journal of Chemical Information and Computer Sciences **43**(4), 1186–1191 (2003). [44](#)

- [214] SERGEI IZRAILEV, FANGQIANG ZHU, AND DIMITRIS K. AGRAFIOTIS. *A distance geometry heuristic for expanding the range of geometries sampled during conformational search*. Journal of Computational Chemistry **27**(16), 1962–1969 (2006). [44](#)
- [215] D. K. AGRAFIOTIS, A. GIBBS, F. ZHU, S. IZRAILEV, AND E. MARTIN. *Conformational boosting*. Australian Journal of Chemistry **59**(12), 874–878 (2006). [44](#)
- [216] DIMITRIS K AGRAFIOTIS, ALAN C GIBBS, FANGQIANG ZHU, SERGEI IZRAILEV, AND ERIC MARTIN. *Conformational sampling of bioactive molecules : a comparative study*. Journal of Chemical Information and Modeling **47**(3), 1067–1086 (2007). [44](#)
- [217] SERGEI IZRAILEV AND DIMITRIS K AGRAFIOTIS. *A method for quantifying and visualizing the diversity of QSAR models*. Journal of molecular graphics & modelling **22**, 275–284 (2004). [44](#)
- [218] JOSÉ L MEDINA-FRANCO, GERALD M MAGGIORA, MARC A GIULIANOTTI, CLEMENCIA PINILLA, AND RICHARD A HOUGHTEN. *A similarity-based data-fusion approach to the visual characterization and comparison of compound databases*. Chemical biology & drug design **70**(5), 393–412 (2007). [44](#)
- [219] FABIAN LOPEZ-VALLEJO, ADEL NEFZI, ANDREAS BENDER, JOHN R OWEN, IAN T NABNEY, RICHARD A HOUGHTEN, AND JOSÉ L MEDINA-FRANCO. *Increased diversity of libraries from libraries : chemoinformatic analysis of bis-diazacyclic libraries*. Chemical biology & drug design **77**(5), 328–342 (2011). [45](#)
- [220] DOMINIQUE GORSE, ANTHONY REES, MICHEL KACZOREK, AND ROGER LAHANA. *Molecular diversity and its analysis*. Drug Discovery Today **4**(6), 257–264 (1999). [45](#)
- [221] V.J. GILLET AND P. WILLETT. 4.08 - compound selection using measures of similarity and dissimilarity. In EDITORS IN CHIEF : JOHN B. TAYLOR AND DAVID J. TRIGGLE, editors, *Comprehensive Medicinal Chemistry II*, pages 167–192. Elsevier, Oxford (2007). [45](#)
- [222] VALERIE J GILLET. *New directions in library design and analysis*. Current Opinion in Chemical Biology **12**(3), 372–378 (2008). [45](#)
- [223] VARUN KHANNA AND SHOBA RANGANATHAN. *Molecular similarity and diversity approaches in chemoinformatics*. Drug Development Research **72**(1), 74–84 (2011). [45](#), [52](#), [53](#)
- [224] G HARPER, S D PICKETT, AND D V S GREEN. *Design of a compound screening collection for use in high throughput screening*. Combinatorial Chemistry & High Throughput Screening **7**(1), 63–70 (2004). [46](#)
- [225] LUC EBERHARDT, KAMAL KUMAR, AND HERBERT WALDMANN. *Exploring and exploiting biologically relevant chemical space*. Current drug targets **12**(11), 1531–1546 (2011). [46](#)
- [226] ROBERT D. BROWN AND YVONNE C. MARTIN. *Use of Structure-Activity data to compare structure-based clustering methods and descriptors for use in compound selection*. Journal of Chemical Information and Computer Sciences **36**(3), 572–584 (1996). [46](#), [52](#)

Références bibliographiques

- [227] YVONNE C MARTIN, JAMES L KOFRON, AND LINDA M TRAPHAGEN. *Do structurally similar molecules have similar biological activity?* Journal of Medicinal Chemistry **45**(19), 4350–4358 (2002). [46](#)
- [228] MARK ASHTON, JOHN BARNARD, FLORENCE CASSET, MICHAEL CHARLTON, GEOFFREY DOWNS, DOMINIQUE GORSE, JOHN HOLLIDAY, ROGER LAHANA, AND PETER WILLETT. *Identification of diverse database subsets using property-based and fragment-based molecular descriptions.* Quantitative Structure-Activity Relationships **21**(6), 598–604 (2002). [46](#), [54](#)
- [229] JÜRGEN BAJORATH. *Integration of virtual and high-throughput screening.* Nature Reviews Drug Discovery **1**(11), 882–894 (2002). [viii](#), [47](#)
- [230] PHILIP J. HAJDUK, WARREN R. J. D. GALLOWAY, AND DAVID R. SPRING. *Drug discovery : A question of library design.* Nature **470**(7332), 42–43 (2011). [47](#)
- [231] DIMITRIS K. AGRAFIOTIS AND VICTOR S. LOBANOV. *An efficient implementation of distance-based diversity measures based on k-d trees.* Journal of Chemical Information and Computer Sciences pages 51–58 (1999). [49](#), [54](#)
- [232] THORSTEN MEINL, CLAUDE OSTERMANN, AND MICHAEL R. BERTHOLD. *Maximum-Score diversity selection for early drug discovery.* Journal of Chemical Information and Modeling **51**(2), 237–247 (2011). [49](#), [54](#)
- [233] MASON J.S. AND PICKETT S.D. *Partition-based selection.* Perspectives in Drug Discovery and Design **7**/8(1), 85–114 (1997). [49](#), [53](#)
- [234] VEERABAHU SHANMUGASUNDARAM AND GERALD MAGGIORA. *Application of shannon-like diversity measures to cell-based chemistry spaces.* Journal of Mathematical Chemistry **49**(2), 342–355 (2011). [50](#)
- [235] DMITRII N. RASSOKHIN AND DIMITRIS K. AGRAFIOTIS. *Kolmogorov-smirnov statistic and its application in library design.* Journal of Molecular Graphics and Modelling **18**(4-5), 368–382 (2000). [50](#)
- [236] DIMITRIS K. AGRAFIOTIS. *A constant time algorithm for estimating the diversity of large chemical libraries.* Journal of Chemical Information and Computer Sciences **41**(1), 159–167 (2001). [50](#)
- [237] M J BAYLEY AND P WILLETT. *Binning schemes for partition-based compound selection.* Journal of molecular graphics & modelling **17**(1), 10–18 (1999). [50](#)
- [238] DIMITRIS K AGRAFIOTIS AND DMITRII N RASSOKHIN. *A fractal approach for selecting an appropriate bin size for cell-based diversity estimation.* Journal of Chemical Information and Computer Sciences **42**(1), 117–122 (2002). [50](#)
- [239] OBDULIA RABAL, ROSALIA PASCUAL, JOSÉ I. BORRELL, AND JORDI TEIXIDO. *Cell-integral-diversity criterion : A proposal for minimizing cluster artifact in cell-based selections.* Journal of Chemical Information and Modeling **47**(5), 1886–1896 (2007). [50](#)

- [240] JOSE L. MEDINA-FRANCO, KARINA MARTÁNEZ-MAYORGA, ANDREAS BENDER, AND THOMAS SCIOR. *Scaffold diversity analysis of compound data sets using an entropy-based measure*. QSAR & Combinatorial Science **28**(11-12), 1551–1560 (2009). [51](#)
- [241] SARAH R LANGDON, NATHAN BROWN, AND JULIAN BLAGG. *Scaffold diversity of exemplified medicinal chemistry space*. Journal of Chemical Information and Modeling **51**(9), 2174–2185 (2011). [51](#)
- [242] ANSGAR SCHUFFENHAUER, PETER ERTL, SILVIO ROGGO, STEFAN WETZEL, MARCUS A KOCH, AND HERBERT WALDMANN. *The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification*. Journal of Chemical Information and Modeling **47**(1), 47–58 (2007). [51](#)
- [243] PETER ERTL, ANSGAR SCHUFFENHAUER, AND STEFFEN RENNER. *The scaffold tree : an efficient navigation in the scaffold universe*. Methods in Molecular Biology **672**, 245–260 (2011). [51](#)
- [244] E J MARTIN, J M BLANEY, M A SIANI, D C SPELLMEYER, A K WONG, AND W H MOOS. *Measuring diversity : experimental design of combinatorial libraries for drug discovery*. Journal of Medicinal Chemistry **38**(9), 1431–1436 (1995). PMID : 7739001. [52](#)
- [245] PETER WILLETT. *Evaluation of molecular similarity and molecular diversity methods using biological activity data*. Methods in Molecular Biology **275**, 51–64 (2004). [52](#)
- [246] GEOFF M. DOWNS AND JOHN M. BARNARD. Clustering methods and their uses in computational chemistry. In KENNY B. LIPKOWITZ AND DONALD B. BOYD, editors, *Reviews in Computational Chemistry*, pages 1–40. John Wiley & Sons, Inc. (2003). [52](#)
- [247] MICHAEL F. M. ENGELS, ALAN C. GIBBS, EDWARD P. JAEGER, DANNY VERBINNEN, VICTOR S. LOBANOV, AND DIMITRIS K. AGRAFIOTIS. *A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition*. Journal of Chemical Information and Modeling **46**(6), 2651–2660 (2006). [52](#)
- [248] JOE H. WARD. *Hierarchical grouping to optimize an objective function*. Journal of the American Statistical Association **58**(301), 236–244 (1963). [53](#)
- [249] J. MACQUEEN. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I : Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif. (1967). [53](#)
- [250] JOHN D HOLLIDAY, SARAH L RODGERS, PETER WILLETT, MIN-YOU CHEN, MAHDI MAHFOUF, KEVIN LAWSON, AND GRAHAM MULLIER. *Clustering files of chemical structures using the fuzzy k-means clustering method*. Journal of Chemical Information and Computer Sciences **44**(3), 894–902 (2004). [53](#)
- [251] ALEXANDER BOCKER, SWETLANA DERKSEN, ELENA SCHMIDT, ANDREAS TECKENTRUP, AND GISBERT SCHNEIDER. *A hierarchical clustering approach for large compound libraries*. Journal of Chemical Information and Modeling **45**(4), 807–815 (2005). [53](#)

Références bibliographiques

- [252] ALEXANDER BOCKER. *Toward an improved clustering of large data sets using maximum common substructures and topological fingerprints*. Journal of Chemical Information and Modeling **48**(11), 2097–2107 (2008). [53](#)
- [253] PAUL R. MENARD, RICHARD A. LEWIS, AND JONATHAN S. MASON. *Rational screening set design and compound selection : Cascaded clustering*. Journal of Chemical Information and Computer Sciences **38**(3), 497–505 (1998). [53](#)
- [254] XIANG-QUN XIE AND JIAN-ZHONG CHEN. *Data mining a small molecule drug screening representative subset from NIH PubChem*. Journal of Chemical Information and Modeling **48**(3), 465–475 (2008). [53](#)
- [255] JEFFREY W. GODDEN, LING XUE, DOUGLAS B. KITCHEN, FLORENCE L. STAHURA, E. JAMES SCHERMERHORN, AND JÜRGEN BAJORATH. *Median partitioning : A novel method for the selection of representative subsets from large compound pools*. Journal of Chemical Information and Computer Sciences **42**(4), 885–893 (2002). [53](#)
- [256] MELISSA R LANDON AND SCOTT E SCHAUS. *JEDA : joint entropy diversity analysis. an information-theoretic method for choosing diverse and representative subsets from combinatorial libraries*. Molecular diversity **10**(3), 333–339 (2006). PMID : 17031536. [53](#)
- [257] M SNAREY, N K TERRETT, P WILLETT, AND D J WILTON. *Comparison of algorithms for dissimilarity-based compound selection*. Journal of molecular graphics & modelling **15**(6), 372–385 (1997). [54](#)
- [258] ROBERT D. CLARK. *OptiSim : an extended dissimilarity selection method for finding diverse representative subsets*. Journal of Chemical Information and Computer Sciences **37**(6), 1181–1188 (1997). [54](#)
- [259] M HASSAN, J P BIELAWSKI, J C HEMPEL, AND M WALDMAN. *Optimization and visualization of molecular diversity of combinatorial libraries*. Molecular diversity **2**(1-2), 64–74 (1996). PMID : 9238635. [54](#)
- [260] VICTOR S. LOBANOV AND DIMITRIS K. AGRAFIOTIS. *Stochastic similarity selections from large combinatorial libraries*. Journal of Chemical Information and Computer Sciences **40**(2), 460–470 (2000). [54](#)
- [261] ROSALIA PASCUAL, MARTA MATEU, JOHANN GASTEIGER, JOSÉ I. BORRELL, AND JORDI TEIXIDO. *Design and analysis of a combinatorial library of HEPT analogues : Comparison of selection methodologies and inspection of the actually covered chemical space*. Journal of Chemical Information and Computer Sciences **43**(1), 199–207 (2003). [54](#)
- [262] DIMITRIS K. AGRAFIOTIS. *Stochastic algorithms for maximizing molecular diversity*. Journal of Chemical Information and Computer Sciences **37**(5), 841–851 (1997). [54](#)
- [263] R P SHERIDAN, S G SANFELICIANO, AND S K KEARSLEY. *Designing targeted libraries with genetic algorithms*. Journal of molecular graphics & modelling **18**(4-5), 320–334 (2000). [54](#)
- [264] D K AGRAFIOTIS AND D N RASSOKHIN. *Design and prioritization of plates for high-throughput screening*. Journal of Chemical Information and Computer Sciences **41**(3), 798–805 (2001). [54](#)

Chapitre II

Délimitation d'espaces chimiques réduits

1 Introduction

Dans ce chapitre, une nouvelle méthodologie permettant de délimiter des espaces réduits obtenus à l'aide de l'analyse en composantes principales (les DRCS - Delimited Reference Chemical Space) sera décrite. L'origine de ce projet tient au constat qu'il n'existe actuellement aucune méthode robuste et surtout reproductible permettant de délimiter visuellement des sous-espaces au sein d'un espace de référence. Nous nous intéresserons ici à des espaces à deux dimensions obtenus à l'aide de l'ACP, bien que cette méthode soit en théorie applicable à d'autres types de représentation 2D (les limitations de la méthode seront discutées plus loin).

L'intérêt d'une telle délimitation est double :

1. Visuellement, elle permet de définir des limites concrètes afin de baliser une zone de l'espace chimique couverte par un ensemble de molécules. De ce point de vue, elle représente une aide à l'interprétation de graphes projetant les molécules dans un espace 2D. Elle facilite par ailleurs l'échange et le partage d'espaces chimiques de référence, comme nous le décrirons dans l'article qui suit et dans le chapitre consacré à Screening Assistant 2.
2. Numériquement, elle permet au delà de l'interprétation visuelle, le calcul d'indices de diversité, comme décrit dans les chapitres 4 et 5. L'une des problématiques importantes dans ce domaine réside dans la difficulté à trouver une délimitation des espaces couverts par les ensembles de molécules à analyser. Typiquement, les méthodes se basant sur l'utilisation de grilles discrétisant l'espace chimique nécessitent par définition de connaître les limites de celui-ci. La méthodologie décrite dans les chapitres qui suivent apporte une solution originale pour l'évaluation numérique de la diversité sur des espaces réduits.

La méthodologie fonctionne comme suit :

Un modèle ACP est tout d'abord créé à partir d'un ensemble de descripteurs. Dans cette étude, nous avons évalué l'utilisation de modèles ACP moyennés afin de lisser les effets parfois négatifs de la présence de valeurs trop éloignées de la moyenne observée sur l'échantillon pour un descripteur donné.

Une enveloppe convexe moyennée est ensuite calculée à partir d'un échantillonnage des coordonnées des molécules projetées sur deux dimensions de l'ACP, typiquement les deux premières. L'enveloppe convexe d'un ensemble E_p est définie par un ensemble de points formant un polygone convexe qui encapsule l'ensemble des points de E_p . Lors de cette étape, N enveloppes convexes sont calculées sur des sous-ensembles de molécules de taille fixe. Pour chacun de ces sous-ensembles, les étapes suivantes sont appliquées :

- On conserve les molécules situées dans la zone la plus dense de l'espace chimique en supprimant les molécules isolées ou trop éloignées du centre du nuage de points.
- On calcule l'enveloppe convexe à partir du nouveau sous-ensemble ainsi généré.

Une fois que les N enveloppes convexes ont été obtenues, on obtient la délimitation finale en moyennant ces enveloppes. L'ensemble de ces étapes a pour but de rendre la délimitation représentative de la zone la plus dense de l'espace chimique définie par le sous ensemble de molécules considéré. Nous obtenons ainsi un moyen de délimiter plus précisément des sous-espaces peuplés par des molécules d'intérêt.

Dans l'étude présentée dans le chapitre suivant et publiée dans le *Journal of Chemical Information and Modeling*, la méthodologie utilisée pour la construction des DRCS a été décrite et analysée en détail et des espaces de référence pour les molécules dédiées au criblage haut débit (espaces HTS) ont été générés. Nous avons créé une base de données de plus de 6 millions de molécules uniques ayant été au préalable filtrées à l'aide de critères classiques utilisés en HTS, à partir de laquelle nous avons réalisé nos études. Nous avons ainsi pu déterminer des jeux de paramètres permettant d'obtenir des modèles et des enveloppes convexes relativement stables. L'utilisation des DRCS a été illustrée dans le cadre de la visualisation et de la comparaison de bibliothèques projetées dans les espaces HTS de référence définis dans

cet étude. L'ensemble des scripts utilisés pour générer les résultats décrits dans cet article a par ailleurs été mis à disposition de la communauté sous licence open-source, avec une documentation complète décrivant leur utilisation. Ils sont téléchargeables à l'adresse suivante : [http ://www.univ-orleans.fr/icoa/DRCS/](http://www.univ-orleans.fr/icoa/DRCS/).

2 Visual Characterization and Diversity Quantification of Chemical Libraries : 1. Creation of Delimited Reference Chemical Subspaces

Visual characterization and diversity quantification of chemical libraries.

1) Creation of Delimited Reference Chemical Subspaces.

Vincent Le Guilloux^a, Lionel Colliandre^a, Stéphane Bourg^b, Guillaume Guénégou^a, Julie Dubois-Chevalier^{a,c}, Luc Morin-Allory^{a}*

a) Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 6005

B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, France.

b) Fédération de Recherche, "Physique et Chimie du Vivant" Université d'Orléans-CNRS; FR 2708, Avenue Charles Sadron, 45071 Orléans Cedex 2, France.

c) Laboratoire d'Informatique Fondamentale d'Orléans (LIFO), Université d'Orléans, Rue de Chartres, 45067 Orléans Cedex 2, France.

AUTHOR EMAIL ADDRESS: luc.morin-allory@univ-orleans.fr

RECEIVED DATE

ABSTRACT

High Throughput Screening (HTS) is a well-established technology which can test up to several million compounds in a few weeks. Despite these appealing capabilities, available resources and high costs may limit the number of molecules screened, making diversity analysis a method of choice to design and prioritize screening libraries. With a constantly increasing number of molecules available for screening, chemical space has become a key concept for visualizing, analyzing and comparing chemical libraries. In this first article, we present a new method to build Delimited Reference Chemical Subspaces (DRCS). A set of 16 million screening compounds from 73 chemical providers has been gathered, resulting in a database of 6.63 million standardized and unique molecules. These molecules have been used to create three DRCS using three different sets of chemical descriptors. A robust Principal Component Analysis (PCA) model for each space has been obtained, whereby molecules are projected in a reduced 2D viewable space. The specificity of our approach is that each reduced space has been delimited by a representative contour encompassing a very large proportion of molecules, and reflecting its overall shape. The methodology is illustrated by mapping and comparing various chemical libraries. Several tools used in these studies are made freely available, thus enabling any user to compute DRCS matching specific requirements.

INTRODUCTION

"Chemical space — which encompasses all possible small organic molecules, including those present in biological systems — is vast. So vast, in fact, that so far only a tiny fraction of it has been explored" observed Dobson in a reference paper in 2004.¹ He provided the following definition of chemical space: "Chemicals can be characterized by a wide range of 'descriptors', such as their molecular mass, lipophilicity (...) and topological features. 'Chemical space' is a term often used in place of 'multi-dimensional descriptor space': it is a region defined by a particular choice of descriptors and the limits placed on them. (...) chemical space is defined as the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created".

The notion of chemical space has been used in drug discovery for over ten years now²⁻⁸ and is still an active field of research.⁹⁻¹⁵ In a rich review, Medina-Franco *et al*¹⁶ quoted another definition which is "the set of all possible molecular structures"¹⁷. This definition is however too trivial and lacks essential characteristics, namely descriptions and rules. Without descriptions and rules, objects cannot be compared to each other. In fact, both in the mathematical meaning and in the common-sense usage (e.g. our 3D universe), spaces are defined by objects having several properties and related by mathematical rules. Thus, a chemical space must be defined by a set (finite or infinite) of compounds and by rules defining their relations (e.g. positions in a multidimensional descriptor space, similarity or dissimilarity metrics, a graphical representation with specific rules etc.).

While Dobson's definition fulfills these requirements and gives a complete and appropriate definition of chemical space, it has a limited practical use since the potential number of compounds and descriptors to be calculated is way too large to be mined. Therefore, chemoinformaticians have in practice used a "restricted" number of compounds and, for each one, a restricted number of descriptors to represent chemical spaces.

This number of compounds varies from thousands to hundreds of millions.^{3, 18-26} Virtual and Tangible chemical spaces as defined by Hann and Oprea²⁷ are a good illustration of this. Virtual space is defined using all the imaginable molecules. In such a space, the number of molecules is almost infinite; even when restricted to small druglike molecules, it is estimated at over 10^{60} ^{4, 28}. Despite the increasing power of recent computers, navigating through such a large set of molecules is impossible with tools available nowadays. Tangible space corresponds to compounds that can reliably be made (many approaches have been tested for the creation of such real or virtual chemical libraries e.g.²⁹); it shows similar limitations. Besides the difficulty of defining what is (and will be) a compound that “can reliably be made”, the potential number of molecules involved is still out of reach. Restricting the space to smaller molecules (up to 11 or 13 heavy atoms^{18, 19}) is possible and yields more usable sets of compounds but, obviously, these spaces represent only part of the real space.

Similarly, obtaining the "total descriptor space" would require the calculation of all the molecular descriptors available. In their reference book "Molecular descriptors for Chemoinformatics"³⁰ Tosdeschini and Consonni reference more than 3 000 descriptors, some of which are fingerprints, graphs or arrays of numerical values. It is therefore possible to characterize each compound by much more than 10 000 values. Working in a space of such a huge dimensionality is practically impossible. One has to reduce the number of descriptors by selecting those which are the most appropriate for a given problem and/or to use data analysis approaches to reduce the apparent dimensionality.^{2, 31-43}

To this end, various multivariate methods have already been used to reduce this dimensionality, and facilitate the interpretation and navigation through chemical spaces. Non linear methods such as Self-Organizing Map (SOM)^{16, 19, 44}, Generative Topographic Map⁴⁵, or multi-fusion similarity approaches^{16, 46} have been successfully used to mine and visualize chemical spaces. On the other hand, Principal Component Analysis (PCA) is a gold standard linear method used since several decades now for dimensionality reduction problems. It was used by Oprea *et al* to create the ChemGPS chemical space navigation system³; they suggested the term of “chemography” as the “art of navigating chemical

spaces". This original approach was subsequently extended toward pharmacokinetic properties⁴⁷ and natural products^{48, 49}, allowing to position chemical entities in different reference spaces. PCA has also been used to analyze active cancer medicinal chemistry compounds, showing that active molecules cover a different chemical space compared to non-active and hit-like molecules⁵⁰. Various "HTS-like" screening libraries (Lipinski molecules, Natural products, Fragments...) were also described and compared by Shelat and Guy⁵¹ using PCA. Reymond *et al* recently used both PCA and SOM to represent a virtual chemical space built using more than 26.4 million virtual molecules described by standard physico-chemical and auto-correlation descriptors¹⁹. The same group subsequently used PCA to determine the chemical space repartition of PubChem compounds which were described using Molecular Quantum Numbers¹⁵. The visualization of binding site-centric chemical space using PCA was also reported by Macchiarulo *et al*⁵². Recently, Singh *et al*⁵³ compared combinatorial libraries, natural products and molecules from Molecular Libraries Small Molecule Repository using various molecular descriptions, including multi-fusion similarity maps and PCA analysis to compare their chemical space coverage. They further highlighted the importance of comparing chemical libraries from different points of view. The wide usage of PCA can be explained by the simplicity of the underlying methodology, the absence of parameters, and the possibility to apply it on a very large datasets without much effort, which is generally more difficult to perform using more complex methods such as those cited previously.

In the framework of our research dealing with chemical libraries^{22, 54} we intend to render the use of chemical space and the notion itself more accessible. The notion of chemical space has two main applications in drug discovery: the comparison of chemical libraries and the quantification of the overall diversity of a given library^{16, 22, 55-57}. The role of chemical diversity in drug discovery and more precisely in the HTS of a chemical library has been the focus of debate for many years^{12, 13, 58-65}. In a recent retrospective study, Sukuru *et al*⁶⁶ close the debate about diversity and HTS, showing clearly the influence of diversity on the ratio of hits.

In this series of two papers, we propose a new way to create and use viewable chemical subspaces, referred to hereafter as Delimited Reference Chemical Subspaces (DRCS). Each DRCS is defined by a set of normalized molecules used to build the space and a set of descriptors encoding each compound. Molecules are then represented in a 2D reduced space using PCA. Furthermore, each DRCS is delimited by a visual contour encompassing a given proportion of molecules and representing the overall shape of the chemical space.

The present paper describes the basic methodology used to build a DRCS, illustrated by its creation for HTS molecules based on a set of more than 16 million available compounds. The second paper will describe the application of DRCS to compare the relative molecular diversity of chemical libraries using DRCS based diversity indices which are independent of the size of the library. The delimitation of chemical spaces makes it possible to use numerous mathematical methods to compute this diversity. Various chemical libraries will be profiled and characterized in terms of diversity.

The practical use of the method, along with DRCS defined in this paper, is made possible for everyone through the release under open-source licenses of several in-house tools used in these studies.⁶⁷

DATA PREPARATION

To illustrate the methodology described herein, three DRCS have been defined for a large set of available HTS compounds. This section describes the gathering and preparation of the data used for this purpose.

Collections of compounds

Collections of compounds from 73 chemical providers have been retrieved, resulting in a set of around 16 million non-unique compounds. The detailed list of providers and the number of molecules for each of them is given in Tables S1 and S2 in Supporting Information. Each provider usually proposes a unique dataset of compounds. Sometimes however, various types of collections are made

available: screening compounds, building-blocks, fragments, target-focused compounds, etc. In such a case, explicit building-blocks libraries were skipped for reactivity reasons. We have chosen to create our own internal database rather than to use publicly available databases because specific standardization and filtering procedures may have been applied in their development (e.g. the Zinc Database²⁶).

Data standardization

Data preparation is a crucial step in chemoinformatics,⁶⁸ especially when chemical descriptors have to be computed. Two identical molecules represented with different ionization states will have different descriptor values, e.g. the number of hydrogen bond acceptors and donors, the formal charge... Moreover, *in silico* representation of molecules is prone to various types of errors, often leading to inconsistent data. A recent publication by Fourches *et al.*⁶⁹ shed light on various practical issues related to data consistency in chemical libraries, and provided several guidelines to minimize the risk of errors and obtain clean datasets. To this end we have developed and applied a 11-step Pipeline Pilot⁷⁰ protocol to obtain the final standardized 3D structures. The main steps of this protocol are summarized in Table 1; a complete description is given in Supporting Information. All molecules considered in this paper were standardized using this protocol. We insist on the fact that using the DRCS defined in this article raises the need to strictly apply this protocol (or an equivalent implementation) to any chemical library subsequently mapped in order to obtain interpretable results.

Duplicates removal

Chemical libraries often contain duplicated molecules, and the standardization process may also create some additional redundancy: some compounds initially present with different counter-ions will be identical after the standardization procedure. To obtain a unique and non-redundant chemical library, an in-house InChITM⁷¹ based script was applied to remove duplicate compounds for each provider's library. The high value of the mean percentage of duplicates for each provider (5.89 %) can be explained by the fact that the libraries of several providers contain up to 50 % of duplicates (Table S2 Supporting Information). This is due to the redundancy existing between multiple SD files for the same provider; e.g. different SD files corresponding to different quantities of the same list of products.

All the standardized libraries were finally merged to obtain a final screening database of 15.10 million unique compounds. As duplicates may also exist between providers, the InChI™ based script was again applied on the whole dataset. Finally, an SD file containing 7.14 million unique and normalized compounds was obtained. This step allows us to compute the “originality” of each provider among this set of providers, defined as the proportion of molecules that are present only in the provider’s library⁷² (Table S2 of Supporting Information).

Reactive compounds removal

An additional filtering step was applied to remove reactive compounds (7.22 %), obviously not meant to be included in an HTS chemical space. The *sdfilter* batch program from MOE⁷³ was used for this purpose. The reactive patterns used are based on those defined by Oprea⁷⁴. A database of 6.63 million unique, normalized and non-reactive molecules was finally obtained.

An overview of all preparation steps is given in **Table 1**.

Steps of the standardization protocol	Removed compounds ^a		Remaining compounds
	#		#
Initial collection from 73 providers	/		16 068 877
File reading	10	(0.6 ppm)	16 068 867
SDF entry without coordinates	7 105	(442 ppm)	16 061 762
Compounds with isotopic atoms	4 855	(302 ppm)	16 056 907
Compounds containing “alias” atoms	1 554	(97 ppm)	16 055 353
Kekulization	21	(1.3 ppm)	16 055 332
Compounds with exotic atoms	3 496	(218 ppm)	16 051 836
Compounds containing atoms with bad valence	621	(39 ppm)	16 051 215
Corina 3D conformation calculation	7 158	(446 ppm)	16 044 057
InChI calculation	5	(0.3 ppm)	16 044 052
Internal duplicates removal	945 622	(58 939 ppm)	15 098 430

Global duplicates removal	7 954 015	(473 189 ppm)	7 144 415
Reactive compounds filtering	515 476	(72 151 ppm)	6 628 939
Final collection from 73 providers	/		6 628 939

Table 1: number of compounds rejected at each step of the overall protocol. (a) Proportions in ppm are calculated with reference to the compounds remaining at the previous step.

CALCULATION OF DESCRIPTORS

Molecular description is the central part of any methodology seeking to represent and compare molecules. In this article, we aim to illustrate a new methodology intended to be coupled with the PCA dimensionality reduction method in order to describe and visualize chemical spaces. Therefore, we do not address the issue of selecting the appropriate descriptors for a specific problem. Since the PCA method is suited to continuous variables, we do not consider fingerprint-based descriptors. Although structural fingerprints have shown their usefulness in many applications of chemoinformatics (e.g. retrieving bioactive compounds⁷⁵), the nature of these descriptors would require the use of an appropriate multivariate method⁷⁶⁻⁷⁸. A similar work, as presented in this article, could be performed in this direction but we will focus here on the widely accepted PCA method. Three sets of descriptors were subsequently used: two sets of 2D descriptors and one set of 3D descriptors (Table 2).

Descriptors	Initial Count	Manually Removed	Null variance	Final Count	Composition for 2D descriptors		
					Physico-chemical	Constitutional	Topological
MOE 2D	185	9	2	174	47.0 %	26.0 %	28.0 %
CDK 2D	215	1	35	179	3.0 %	50.0 %	47.0 %
MOE 3D	148	31	0	117	-	-	-

Table 2: the three sets of descriptors used to compute each corresponding DRCS. Manually removed descriptors correspond to various meta-descriptors such as Lipinski's "drug-like" flag.

The choice of MOE 2D descriptors is justified by its good coverage of various standard types of 2D descriptors.⁷⁹ The free and open source CDK 2D molecular descriptors^{80, 81} were selected so as to allow anyone to use and navigate through the DRCS defined in this study, and for comparison purposes. Only a few descriptors were removed for statistical reasons (see Table S3 in Supporting Information for details).

Using a simple classification scheme, Table 2 shows that CDK descriptors are highly biased toward constitutional and topological descriptors, while MOE descriptors have a much more balanced distribution. Although the assignment of a descriptor to a given category may be subject to discussion, the differences are notable enough to support this observation.

In addition to these two pools of descriptors, MOE 3D descriptors were also used to assess eventual differences in chemical libraries mapping between 2D and 3D descriptors. Only a few descriptors were removed for various reasons (see Table S3 in Supporting Information for details).

The *sddesc* program from MOE was used to calculate 2D and 3D descriptors. Prior to descriptor calculation, partial charges were computed using the MMFF94 forcefield⁸². CDK descriptors were calculated using an in-house JAVA program made freely available with the other tools used in this study.

DRCS DEFINITION

Based on the previous data, the three chemical spaces are basically represented by all the descriptors, one space for each descriptor set. Their large dimensionality (one dimension per chemical descriptor: 174, 179 and 117 dimensions) makes visual analysis impossible. A reduction to a lower dimensionality (e.g. 2 or 3) is thus required to obtain visual and intuitive representations. This can be easily performed using PCA.

Briefly, PCA is a popular linear projection method used to transform an N-dimensional space into an M-dimensional one ($M < N$) created by M uncorrelated vectors called Principal Components (PCs).

Principal components are actually defined by the eigenvectors of the variance-covariance matrix of the input matrix. In our case, the input matrix is centered and scaled to unit variance prior to any calculation. Principal components are then calculated from the variance-covariance matrix of the scaled original matrix. When ordered by decreasing eigenvalues, the first components correspond to the largest eigenvalues and explain the largest amount of the total original variance. Finally, each PCA model is defined by a set of principal components, and a set of means and standard deviations used to scale and center each descriptor.

Graphical representation of the resulting chemical space can then be obtained by projecting the original molecules (N-dimensions) onto a reduced PCA space using the first two or three components. We are aware that the transformation of such a high dimensional space into 2D or 3D space leads to a substantial loss of information. But priority has been given to intuitive and easy visualization and interpretation of each DRCS, and despite the dimension reduction required by the visual analysis, working on the N-dimensional DRCS remains possible. Practically, two proximal points on this representation can be distant in real space if they differ only by descriptors not represented in the first two components (*i.e.* having low weights). But, as a large part of the information is available in the first components, two distant points in the representation are expected to be rather distant in the real space. Yet, as some descriptors are not represented in the first components, these two distant points can still be rather closed when considering only other components. But their overall distance in the real space is, at least, their distance in the plane created by the two first components.

To compare the relative positions of chemical libraries within a reference chemical space, both molecules of interest and molecules of reference have to be plotted in the reduced PCA space if no boundaries have been defined. This raises two potential problems: (i) descriptor values of the reference molecules must be available and (ii) in our case, plotting several million molecules for each chemical library of interest cannot be done rapidly, and may generate graphs which are difficult to interpret. To address these issues, each chemical space will be delimited by a representative contour that can be used independently of the original dataset. This contour defines a zone of the space containing a given

proportion of molecules and represents a visual aid enabling chemical libraries to be compared and positioned in an intuitive way. A DRCS is finally defined by the following two components:

1. **A PCA model.** For visual mapping, only the first three components have been retained for each chemical space. Based on these eigenvectors, 2D and 3D space coordinates can easily be assigned to each compound. In the remainder of this paper, only 2D spaces will be considered, as visual interpretation is easier on a 2D medium, but the extension to three dimensions is of course possible.
2. **A contour defining the chemical subspace boundaries.** Each contour is computed using coordinates of molecules in the 2D space. This contour is based on a convex hull calculation⁸³. The convex hull of a set of points S is the minimal subset of points creating a convex polygon that completely encloses S. It is finally defined as a set of ordered points creating a polygon. An intuitive illustration of the concept may be to imagine an elastic band stretched open to encompass the given objects; when released, the elastic will enclose all the objects, reflecting its convex boundaries. If a 3D space is used, a 3D convex hull is created, generating a potato-shaped volume.

DRCS CONSTRUCTION

The most intuitive idea to build the DRCS based on a set of compounds would be to compute both PCA and contour on the whole set. This simple idea however is neither appropriate nor efficient for two main reasons:

1. Computing the PCA on the whole dataset may not be possible in a reasonable time scale for a large number of molecules. The cost of descriptor calculation (computation time, cost of the licenses) can significantly decrease our capacity to calculate a great number of descriptors. This is typically the case for the MOE 3D descriptors used in this study.

2. The convex hull is by nature very sensitive to the dataset composition, because it is defined by the objects located in the extreme zones of the space. Exotic molecules are typically mapped in these extreme zones of the reduced PCA space, and decrease the representativeness of the boundaries. Moreover, boundaries derived from a single convex hull are rather rough and located at the exact position of molecules defining the convex hull. A single exotic compound can substantially modify the final shape of the contour.

The procedure used to build each DRCS has therefore been broken down into two steps: PCA model definition and contour calculation. The following sections will describe these two steps and how their respective issues have been tackled. The final DRCS creation procedure will then be outlined.

PCA model definition

The limited number of licenses for commercial software decreases our capacity to calculate descriptors which require an extensive calculation time, e.g. MOE 3D descriptors. To overcome this problem, we tested the hypothesis that an equivalent PCA model could be obtained by averaging a limited number of PCA models computed on random subsets containing a limited number of molecules. The number of subsets and number of molecules per subset were determined in order to obtain PCA models that could be presumed to be very similar to those computed on the entire dataset. This approach was validated using the MOE 2D dataset, as the calculation of all descriptors for the whole dataset took only slightly over three days. The supplementary information details this procedure and the algorithm used to compute each averaged PCA model. Following this analysis, 30 random subsets of 20 000 molecules were extracted from the whole dataset. These subsets were subsequently used to compute average DRCS PCA models for each descriptors pool.

It should be emphasized that no molecules were filtered out during the model computation. The model is actually defined to be representative of the whole dataset, and no particular bias is introduced. We will see in the next section that this is not the case for contour calculation.

Contour calculation

Computing a single convex hull on the whole dataset is clearly not a satisfactory way to get a representative contour. One would obtain a rough contour enclosing all molecules, but no information describing their actual distribution and density. Our aim, in contrast, is to create a contour representing the part of the chemical space occupied by a very large proportion of molecules, excluding the outliers located in extreme or isolated zones of the space. This is obviously a simplified representation of a chemical space, highlighting the most important characteristics of the compounds' distribution. The obvious advantage is its ease of interpretation. It is conceptually similar to a caricature which exaggerates some characteristics and oversimplifies others, allowing faster recognition of a face than using the original photograph.⁸⁴

To build such a contour, the selected strategy is to compute several convex hulls on several subsets of compounds, with an additional step removing obvious outliers for each subset prior to each individual contour calculation. It should provide a way to: (1) make the boundaries representative of a given proportion of the original set by excluding isolated compounds, and (2) smooth the boundaries by averaging several contours, thus increasing the representativeness. The outlier removal procedure was defined so as to obtain a good trade-off between stability of the final shape and a small proportion of molecules located outside the final consensus envelope. For the sake of simplicity, we used the same 30 subsets as those used to compute each consensus model. For each subset, a given proportion of outliers was first removed; a convex hull was then computed using the remaining molecules.

Removing outliers raises the need to provide a definition of what constitutes an outlier. In the Oprea *et al.* study³ outliers (referred to as 'satellites') are defined both explicitly (e.g. benzene, cubane), or based on extreme property values. In another context, Baskin *et al.*⁸⁵ used one-class Support Vector Machine to surround chemical space regions having a high density of data points, making it possible to identify isolated compounds. Actually, there is no absolute definition of what an outlier is, and there could be many different and equally valid ones, depending on the context. In this study, the removal of outliers is needed to obtain a stable and representative envelope encompassing a given, preferably large proportion of molecules in the reduced space. Typically, outliers in the reduced space are located either

in extreme zones of the space, or isolated in a sparse region where the neighborhood density is very low. Thus, outliers have been defined based on their distance from the barycenter of the cloud (hereafter named barycenter method) and on their neighborhood density (hereafter named density method).

The following procedure was applied prior to the calculation of each contour for each subset:

1. Compute the reduced space coordinates for each molecule of the current subset.
2. Remove b % of molecules having the highest distance from the barycenter of the resulting cloud of points.
3. Remove d % of molecules having the lowest neighborhood density. For a given molecule, the neighborhood density is defined as the number of molecules located at a Euclidean distance D in the 2D reduced space.

The combination of these two different ways of removing outliers has been found to give the best results in terms of stability, representativeness and number of molecules excluded, a number which we try to minimize here. Both methods (and other in-house methods, not discussed here) were first tested separately with different values of parameters and numbers of molecules in the subset.

The consensus contour is finally obtained as the average of the contours of the 30 subsets. From the origin of the referential – which is located close to the center of the cloud of points due to the data centering procedure – 360 successive vectors are drawn making angles from 0° to 359° with the X axis. Each vector intersects the 30 convex hulls in 30 points. The average coordinates of these points are computed and represent the consensus point corresponding to each angle. These 360 average points finally form the consensus hull.

To investigate the differences between the two methods and the advantage of their combined use, both parameters b and d were successively set to 0 while the other varied from 0 to 5 % with a step of 0.025 %. Each consensus hull was subsequently calculated using the 3 DRCS models used in this study.

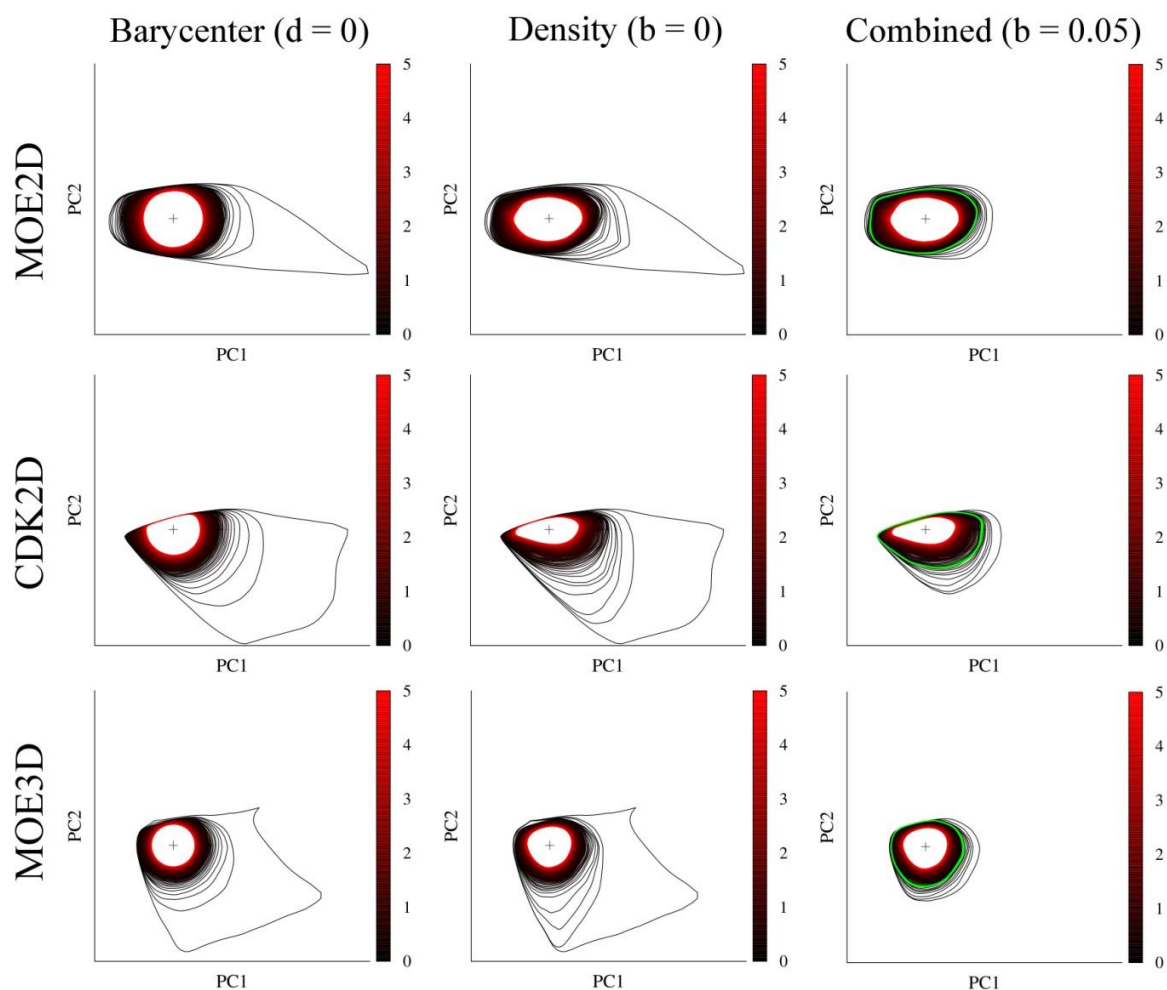


Figure 1: comparison of three outlier removal methods for the three DRCS. Consensus hulls are computed on the 30 subsets of 20 000 molecules using PC1 and PC2 coordinates with $D = 3$, and with a percentage of outliers removed ranging from 0 (black) to 5 (red) with a step of 0.025. **Left:** only the barycenter-based method is used ($d = 0$, b ranges from 0 to 5). **Middle:** only the density-based method is used ($b = 0$, d ranges from 0 to 5). **Right:** both methods are combined, with b fixed at 0.05 and d ranging from 0 to 5. The green consensus contours correspond to the final d value used to compute each DRCS contour. Axis boundaries have been scaled to better highlight the circled shape induced by the barycenter method.

Figure 1 shows the resulting consensus hulls obtained in all cases. The barycenter method used alone leads to consensus hulls showing a clear tendency to shift toward a circle shape, which obviously does not necessarily reflect the actual shape of a given chemical space. The density based method alone gives

much better results, reflecting the shape of each space based on the neighborhood density, but the number of molecules that need to be removed in order to obtain stable envelopes appears to be slightly higher. Figure 1 also demonstrates the importance of removing outliers. Chemical subspace boundaries reach stability (i.e. a stable shape which only becomes smaller when increasing the percentage of outliers removed) only when a given proportion of outliers has been removed. Based on empirical observations and determination of stability (see Supporting Information), b and d were set to 0.05 % and 0.20 %, respectively. We have verified that this procedure preserves the ratio of points inside and outside the final contour as defined by the two parameters b and d . Plotting the 600 000 molecules used to compute each DRCS yields 0.33 %, 0.30 % and 0.32 % of molecules outside the hull for MOE 2D, CDK 2D and MOE 3D subspaces, respectively. These ratios are very close to that defined for the creation of each single convex hull (0.25 %), the difference being explained by the averaging procedure.

Final procedure

The following final procedure was used to create each DRCS:

1. Pick 30 random subsets of 20 000 molecules in the HTS-Database,
2. Compute each subset PCA,
3. Compute the DRCS model by averaging PCA computed on each subset,
4. Find and filter out outliers for each subset,
5. Compute the convex hulls on the filtered subsets,
6. Compute the consensus contour by averaging these convex hulls,
7. Define the DRCS as being the PCA model and the consensus contour.

Of course, steps 1-3 could be replaced by a single PCA model computation procedure on the total library. The 30 reference subsets were finally projected in the three DRCS as shown in Figure 2. Although exotic compounds had been removed prior to calculating each convex hull, this figure also shows that averaging several contours is needed to obtain stability and representativeness. These contours encompass more than 99.6 % of the reference compounds, and provide useful visual

boundaries to describe and compare chemical libraries in this HTS space, as illustrated in the following section.

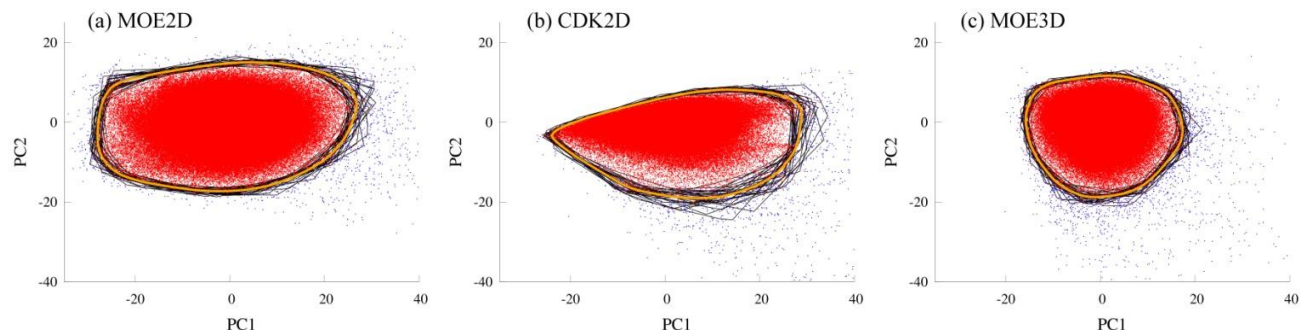


Figure 2: projection of the 30 subsets (for a total of 600 000 molecules) used to compute each DRCS (a) MOE 2D space, (b) CDK 2D space and (c) MOE 3D space. Blue dots represent molecules that have been excluded for at least one contour calculation. Black polygons represent convex hulls of each subset. The orange shape represents the final consensus contour.

USE OF DRCS FOR VISUALIZATION

Visualization is a rapid and intuitive way to explore and describe the content of a library. Selection of a screening library among all those proposed by the different providers raises the need to rapidly determine the chemical space coverage of each library and to compare these libraries in order to select the one(s) having the required chemical space coverage. Rapid identification of unexplored chemical space zones also greatly facilitates library enrichment with compounds having complementary properties regarding the chemical space under consideration.

By projecting chemical libraries using the described DRCS methodology, one is able to tackle these issues. Figure 3 illustrates the advantage of using the consensus contour. A chemical library (the Chembridge Diverset Library⁸⁶ – CDL) has been projected in various ways. In Figure 3a the library has been plotted in its own reduced space. It can be seen that it is very difficult to derive any useful

information in terms of chemical space coverage using this figure, as there is no reference to compare it with. When it is plotted in the MOE 2D reduced space against the 30 random subsets used to build the space (Figure 3b), it becomes much easier to compare both libraries and to assess the HTS chemical space coverage of the CDL: the figure clearly shows that the library has a limited coverage on this space. In Figure 3c these 30 subsets have been replaced by the consensus contour. One can see that a very similar interpretation can be derived, without the need to use the descriptor values of the original molecules which may be either unavailable or difficult to plot (especially when it contains millions of molecules). Moreover, with such a huge dataset, it is very difficult in (b) to view the isolated compounds of the library which are located in dense regions of the reference space.

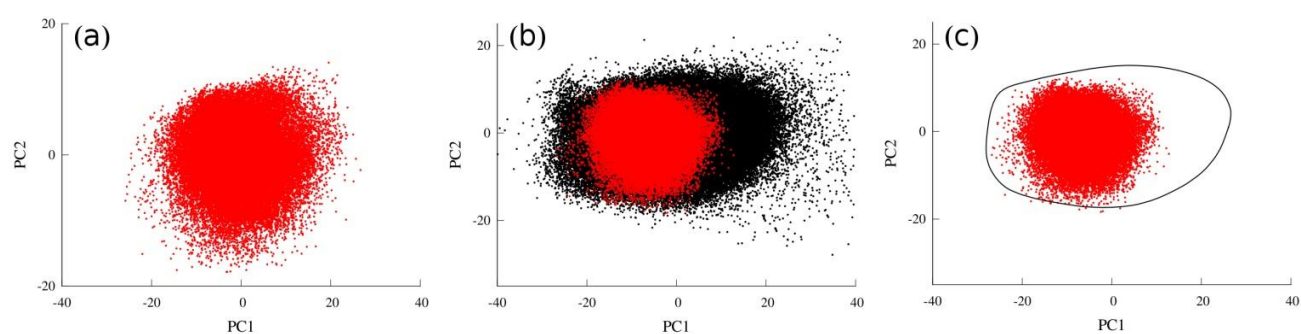


Figure 3: mapping of a chemical library (a) using its own MOE 2D PCA model, (b) in the MOE 2D space (red) with the 30 subsets (black) used to compute the MOE 2D model, and (c) in the MOE 2D DRCS plotted with the contour.

DRCS INTERPRETATION AND COMPARISON

Three DRCS have been obtained using the methodology described in the previous sections, one per descriptor set: DRCS-MOE 2D, DRCS-MOE 3D and DRCS-CDK 2D. The percentages of explained variances are shown in Table 3 for components 1 to 5 of each DRCS model.

	MOE 2D		CDK 2D		MOE 3D	
	(%)		(%)		(%)	
	<i>Component</i>	<i>Cumulative</i>	<i>Component</i>	<i>Cumulative</i>	<i>Component</i>	<i>Cumulative</i>
PC-1	31.31	31.31	25.84	25.84	19.01	19.01
PC-2	12.20	<u>43.51</u>	6.09	<u>31.93</u>	16.48	<u>35.49</u>
PC-3	10.99	54.50	5.07	37.01	10.33	45.82
PC-4	4.74	59.24	3.89	40.89	8.16	53.98
PC-5	3.84	63.08	3.49	44.39	5.98	59.97

Table 3: mean explained variance of the first five components of the PCA consensus model for each set of descriptors. Corresponding cumulative variances are presented in bold. The total explained variances for each 2D reduced space are underlined.

The percentage of explained variance varies depending on the descriptor set under consideration. It is important to realize that these DRCS cannot be compared based on their respective explained variance as the number and the nature of descriptors differ for each set. In fact, a large value for the total explained variance might rather suggest the presence of very strong inter-correlations between descriptors. Intuitively, one cannot expect to obtain a large proportion of explained variance by reducing the dimensionality e.g. from 174 to 2, unless strong redundancy exists within the original descriptor space. In other words, large explained variances are not necessarily synonymous with better chemical space representativeness.

For visualization purposes, only the first two PC were used to obtain viewable 2D DRCS, leading to a cumulated explained variance of 43.51 %, 31.93 % and 35.49 % for DRCS-MOE 2D, DRCS-CDK 2D and DRCS-MOE 3D, respectively. For MOE 2D and MOE 3D DRCS, PC-3 explains around 10 % of the variance, resulting in a 3D chemical space accounting for around 50 % of the total variance. Although the third component of the CDK 2D DRCS does not explain as much amount of variance, taking into account this component may lead to a substantial gain in information using a 3D visualization device.

The chemical interpretation for each axis can be derived by comparing the relative weights of descriptors for each PC (see Supporting Information for detailed figures showing PC weights), or using simple visual analysis of the compounds' distribution. PC-1 of DRCS-MOE 2D and DRCS-MOE 3D globally describes molecular size through large weights for descriptors related to volume, VDW surface area or molecular weight / atom counts. PC-1 of CDK 2D is more related to molecular complexity, probably explained by the larger number of topological descriptors present in the CDK library. PC-2 of both DRCS-MOE 2D and DRCS-MOE 3D spaces order compounds by increasing hydrophobicity. In PC-2 of DRCS-CDK 2D, the compounds are classified in a slightly different manner, ranging from aliphatic to aromatic compounds. Furthermore, the shape of the DRCS-CDK 2D (roughly triangular) is very similar to the shape of the space described by van Deursen *et al*,¹⁵ where the PCA was computed based on PubChem⁸⁷ compounds. The MQN descriptors used in their study are primarily constitutional and topological descriptors, which explains the similarity observed with our CDK 2D space. Finally, these tendencies show that different chemical descriptors lead to spaces closely related to weight = f(hydrophobicity) as it was previously found by Egan *et al*⁸⁸. However, each DRCS has its own specificities that will be highlighted in the next section.

LIBRARY MAPPING AND COMPARISON

In this section, the use of DRCS to view and compare libraries will be illustrated, focusing in particular on four different sets of molecules:

1. The Comprehensive Medicinal Chemistry (CMC), a database of pharmaceutical compounds.⁸⁹
2. The Prestwick Chemical Library® containing marketed drugs.⁹⁰
3. The Pyxis Discovery Smart Fragment Library,⁹¹ a fragment library based on scaffolds found in existing drugs.
4. Estrogen receptor (ER) agonists and antagonists provided by the DUD benchmark dataset.^{92, 93}

Prior to any library mapping to a DRCS, each library was standardized using the protocol described in a previous section and the three sets of descriptors calculated. The results are given in Figure 4.

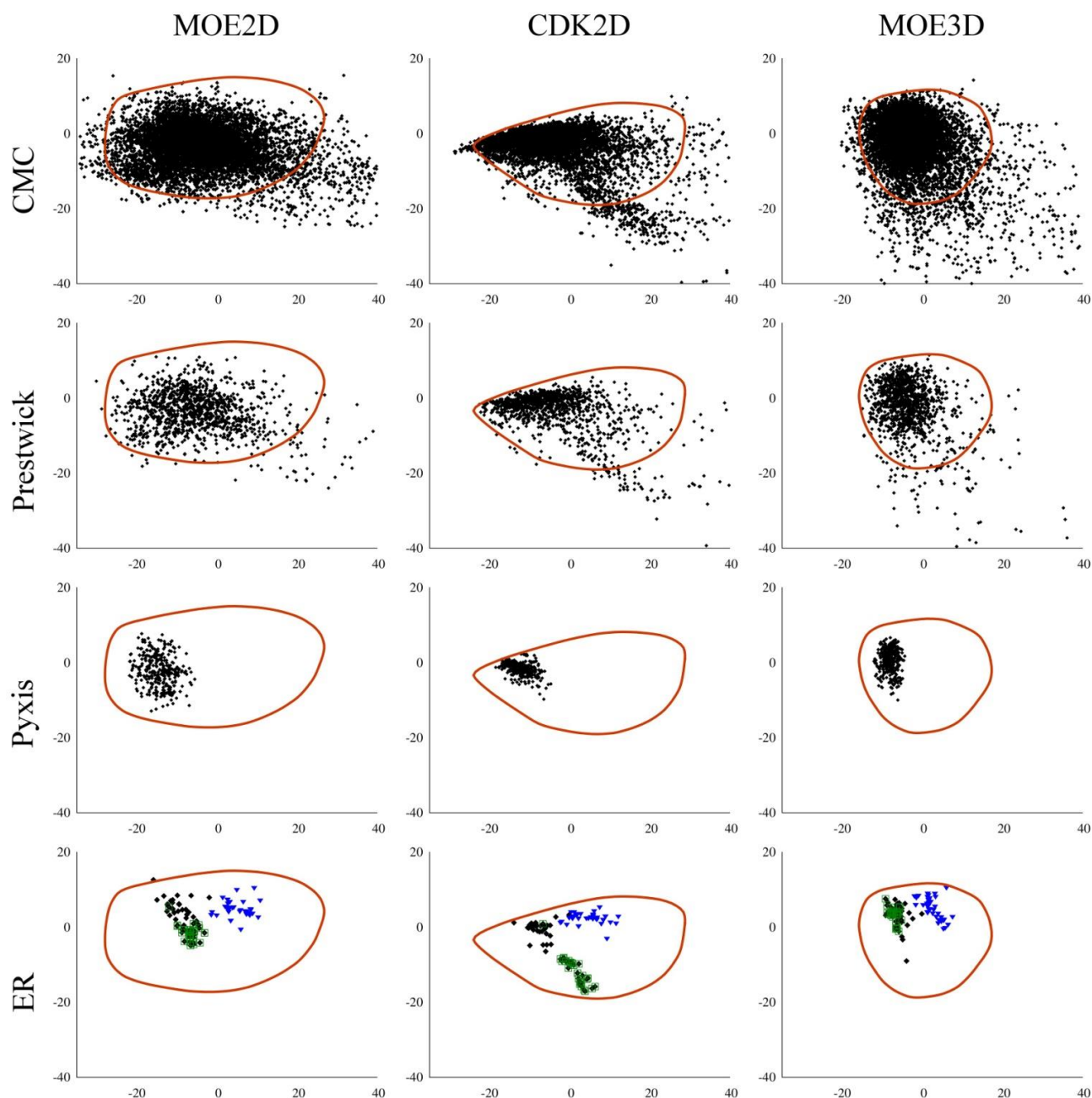


Figure 4: various libraries mapped in the DRCS described in this article, MOE 2D (left), CDK 2D (middle), and MOE 3D (right). The CMC, Prestwick, Pyxis and Estrogen receptor agonist (black) and antagonists (blue triangles) are shown from top to bottom, respectively. Steroid-like agonists are highlighted in green.

The 8 773 compounds of the CMC database cover a very large proportion of the three DRCS contours. Only one small region is not covered for each of these contours. It corresponds to large and very lipophilic compounds and could be associated to compounds having poor solubility and/or

bioavailability. Around 9.5 %, 8.1 % and 10.4 % of the compounds are located outside the contour in the MOE 2D, CDK 2D and MOE 3D DRCS, respectively. On the negative extreme part of PC-1, outside the hull, very small and volatile compounds can be found (eg. NO or cyclopropane which are gases used for anesthesia). On the extreme positive part, complex compounds such as cyclosporin A, a natural macrocycle, are typically present. These classes of compounds are unlikely to be exploited by the providers forming our initial HTS database, and it thus makes sense to find them outside the contour. The coverage of the DRCS contours and the large proportion of compounds outside these contours show that the chemical space of these pharmaceutical compounds has a significantly different distribution from that of commercial HTS compounds.

The Prestwick library, which is commercially available for screening purposes, shows similar space coverage to that of CMC. It contains 1 200 small molecules, 100 % being marketed drugs. Around 5.8 %, 7.3 % and 8.1 % of the compounds are located outside the contour in the MOE 2D, CDK 2D and MOE 3D DRCS, respectively. As the compounds are similar to those of the CMC (both are pharmaceutical compounds), it seems natural to observe the same distribution. With so few compounds, this library is clearly a good starting point for the creation of an HTS diverse library covering the current marketed drug space.

The projection of these two collections suggests some similarities between the HTS chemical space and the so-called “drug-like space” as defined using the Lipinski⁹⁴ rules for oral bioavailability. The "Lipinski" filter implemented by the MOE software was applied to the 6.6 M dataset, yielding around 6.2 million druglike (93.8 %) and 0.4 million non-druglike (6.2 %) molecules. Figure 5 shows the CMC plus Prestwick (CMC-P) set plotted in the MOE 2D subspace against the resulting druglike molecules (Figure 5a) and non-druglike molecules (Figure 5b). A rather good correspondence can be observed between the CMC-P set and the Lipinski druglike molecules space coverage, as shown in Figure 5a. An empty zone can be found in both sets on the top right zone of the DRCS, which corresponds to large lipophilic compounds, showing that the Lipinski filter clearly makes sense in this space area regarding pharmaceuticals and marketed drug compounds. Conversely, Figure 5b shows that this same empty

zone is filled by molecules not matching the Lipinski filter. On the other hand, a higher compounds density can be observed, outside the hull, on the bottom right part of the space for the CMC-P set, corresponding roughly to large hydrophilic molecules. The Lipinski druglike molecules are almost absent in this area (Figure 5a), while non-druglike molecules are much abundant (Figure 5b). It can also be seen that inside the contour, a significant overlap appears between zones covered by drug-like and non-druglike molecules, showing that pharmaceutical and marketed compounds can also be found in mixed (druglike + non-druglike) as well as in exclusively non-druglike zones of the space. All these trends highlight the advantages and limitations of the druglike filters.

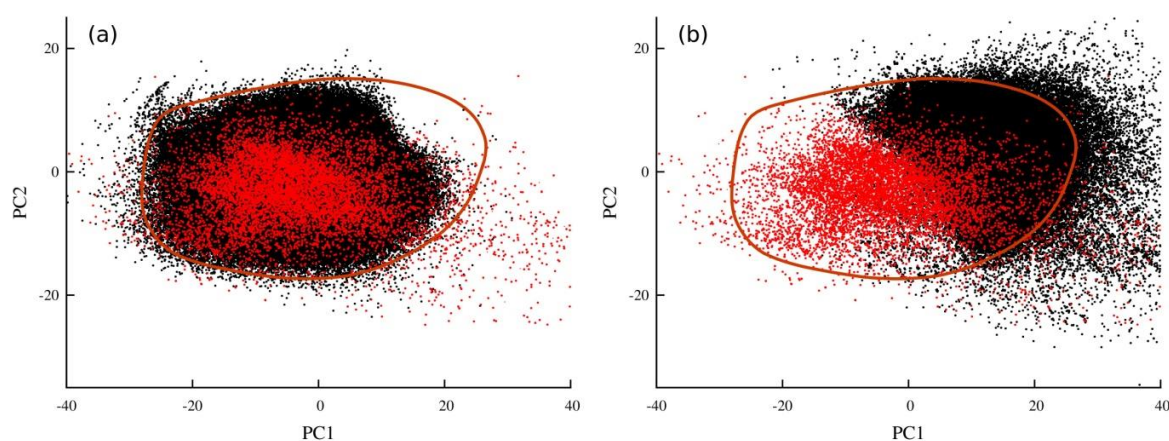


Figure 5: Prestwick and CMC (red dots) plotted in the DRCS-MOE 2D against (a) a random subset of 2 million "drug-like" molecules filtered using the MOE Lipinski filter, and (b) all non-druglike molecules (same filter) extracted from the entire dataset.

The Pyxis library is entirely located within the DRCS boundaries, in the left part of the subspace. The rule of three⁹⁵ used to define what a fragment is implicitly explains this observation. Both the lower and upper limits applied for the molecular mass and consensus properties defined by this rule explain the tight location of the library. It gives us an idea of where the fragment subspace of the HTS library is. The same simple idea could also be applied to other rules (e.g: lead-like) or subsets of molecules (target-specific), and focused sub-contours could also be drawn within the original space using the same consensus methodology described in this paper.

The ER agonists and antagonists are located in two distinct zones of each subspace. Global physicochemical trends separating the two classes of compounds can thus be easily captured using the three spaces axis interpretations. For the agonists however, each DRCS yields a slightly different compounds distribution when looking at the compounds' structures. Indeed, the CDK 2D contains more tight and isolated clusters, corresponding to different classes of molecules. In figure 4, steroid-like agonists have been manually highlighted in green. These compounds are clearly clustered in the CDK 2D subspace, and to a lesser extent in the MOE 2D subspace. In the CDK 2D space, they are found clearly isolated from other agonists. In contrast, no obvious specific cluster can be found in the MOE 3D space, and the steroids are found among the other agonists suggesting that 3D descriptors for both PC under consideration provide different information compared to CDK 2D and MOE 2D spaces.

Interestingly, the 3D information does not seem to provide any additional information in terms of space coverage compared to 2D descriptors. Although the CDK 2D space seems to give some topological information, the overall conclusion in terms of chemical space coverage remains quite similar despite the different nature of the descriptors used in each DRCS. This shows that, for a rapid overview of the diversity of a chemical library, any of these three DRCS could be used. The simplicity and speed of this methodology obviously ease its systematic application to the analysis of new chemical libraries.

CONCLUSION AND PERSPECTIVES

A new methodology has been described to create Delimited Reference Chemical Subspaces (DRCS). These so-called DRCS introduced a new way of associating each chemical space with boundaries that reflect their overall shape, encompassing a large majority of compounds in a reduced and viewable 2D subspace. Such a visual aid is of great interest for exploring and comparing chemical libraries by assessing their relative chemical space coverage, leading to more rational library design and selection. Furthermore, we showed that PCA models, very similar to that computed using the entire dataset, can be

obtained by averaging PCA models computed on several random subsets of molecules. This observation could be useful if the number of molecules to be processed is very large and if computational resources and / or software licenses are limiting factors, though care must be taken with respect to several issues highlighted in Supporting Information of this article. Besides the original purpose, data set cleaning and standardization has emerged as a critical issue. We believe that this step should never be neglected and must be systematically detailed in a modern chemoinformatics study, as results might be dramatically affected by inconsistencies and undetectable errors.

The crucial descriptor selection step has not been addressed here since it is usually context-dependant. Obviously, depending on the nature of the target and the amount of knowledge available, focused library design may require careful descriptors selection to create the appropriate DRCS and assess “target subspace” coverage. In contrast, the general-purpose HTS chemical spaces described in this study use a wide and diverse range of descriptors, without any particular assumption. They could be useful to assess the global diversity of a chemical library, typically for a blind HTS campaign where only little information is available on the targeted entity-ies. Yet, even with no particular descriptor selection, we showed that different information could still be captured depending on the set of descriptors used. Interestingly, the overall conclusion seems to remain quite similar in terms of chemical space coverage, regardless of which DRCS is under consideration. Rational selection of descriptors would probably provide more contrasted results for specific purposes.

The conclusions that can be derived using a chemical space based on principal components are obviously limited to the information available in the components under consideration. In the spaces described here, the first two components only explain around 40 % of the information available in the original descriptor space, and are related to global properties like size, complexity and lipophilicity. These properties are useful to determine zones of the space spanned by favorable physicochemical properties, and allow one to easily identify bias in chemical libraries. Further insight into chemical space coverage could nevertheless be obtained by computing DRCS using other components. This would provide additional information and complementary molecular representations, and one could

obtain a more complete picture of the chemical space, although at the expense of simplicity and ease of interpretation.

The contour associated with each DRCS provides a useful visual aid to compare chemical libraries. By encompassing the densest region and reflecting its overall shape, the most important characteristics of the reduced subspace can be highlighted. A further extension of this approach would be to define subspaces encompassing specific types of molecules, making it possible to locate more focused subspaces.

Besides a simple visual aid, the contour makes it much easier to obtain a more representative grid-based partition of each subspace. This way, grid-based diversity indices can be easily applied, and would reflect more accurately the coverage of the most explored regions, thereby making easier to determine an appropriate grid resolution.

From a more practical point of view (e.g. diversity indices calculation or diverse subset extraction), we suggest that the densest and more sparse regions of a chemical space should be treated separately. By creating a clear delimitation for the densest region, the focus is set on the widely explored part of a given chemical space, which a good general-purpose / prospective screening library should cover anyway. More exotic compounds that usually represent unexplored chemotypes, would require more attention and should thus be treated and sampled differently. The convex delimitation of each space provides a way to perform such analysis; a second paper will explore the creation of specific subspaces as well as various diversity indices adapted to the DRCS methodology.

Finally, the open source availability of the basic tools used in this study, both as standalone tools and as Screening Assistant⁹⁶ platform features, opens the way to easy and interactive DRCS navigation. Free, open and validated tools / data are still a missing piece in chemoinformatics, especially compared to the bioinformatics field. Fortunately, more and more valuable and very promising initiatives have been recently reported – CDK,⁹⁷ Bioclipse,⁹⁸ RDKit,⁹⁹ CDK-Taverna,¹⁰⁰ KNIME¹⁰¹ to cite but a few - and the work described herein is clearly striving to move in this promising direction.

ACKNOWLEDGMENT:

The authors thank Accelrys for providing, free of charge, the software "Pipeline Pilot Student edition", and the "Conseil Régional du Centre" for supporting this research. The authors also thank the referees and Peter Schmidtke for helpful comments on the manuscript. VLG thanks the "Conseil Général du Loiret" and JDC the "Conseil Régional du Centre" for funding their respective PhDs.

SUPPORTING INFORMATION PARAGRAPH

Tables S1 and S2: List of providers and quantitative data about the products,

Table S3: List of removed descriptors

Detailed preparation of chemical structures

Description of tools and computational performances

Methodology for the computation of the consensus PCA

Parameters of the contour calculation

The file Models.zip contains all the figures describing each PCA model in detail, and all the figures comparing the global and the average MOE 2D model.

This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

1. Dobson, C. M. Chemical space and biology. *Nature* **2004**, 432 (7019), 824-828.
2. Willett, P. Computational tools for the analysis of molecular diversity. *Perspect. Drug Discovery Des.* **1997**, 7/8, 1-11.
3. Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, 3 (2), 157-166.
4. Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, 16 (1), 3-50.
5. Lahana, R. How many leads from HTS? *Drug Discovery Today* **1999**, 4 (10), 447-448.
6. Gorse, D.; Rees, A.; Kaczorek, M.; Lahana, R. Molecular diversity and its analysis. *Drug Discovery Today* **1999**, 4 (6), 257-264.
7. Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem. Int. Ed. Engl.* **1999**, 38 (24), 3743-3748.
8. Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, 7/8, 31-49.
9. Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* **2010**, 50 (4), 470-479.
10. Gu, Q.; Xu, J.; Gu, L. Selecting Diversified Compounds to Build a Tangible Library for Biological and Biochemical Assays. *Molecules* **2010**, 15 (7), 5031-5044.
11. Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **2010**, 1, 30-38.
12. Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P. Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol.* **2009**, 27 (1), 18-26.
13. Schneider, P.; Tanrikulu, Y.; Schneider, G. Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. *Curr. Med. Chem.* **2009**, 16 (3), 258-266.
14. Varnek, A.; Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inf.* **2011**, 30, 20-32.
15. van Deursen, R.; Blum, L. C.; Reymond, J. L. A searchable map of PubChem. *J. Chem. Inf. Model.* **2010**, 50 (11), 1924-1934.
16. Medina-Franco, J. L. M.-M.; Karina; Giulianotti, Marc A.; Houghten, Richard A.; Pinilla, Clemencia Visualization of the Chemical Space in Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2008**, 4 (4), 322-333.
17. This definition is similar to the one given by Wikipedia, English version, http://en.wikipedia.org/wiki/Chemical_space (accessed January 15, 2011).
18. Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, 131 (25), 8732-8733.
19. Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, 47 (2), 342-353.
20. Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002**, 6 (3), 384-389.
21. Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Lead-like, drug-like or "Pub-like": how different are they? *J. Comput.-Aided Mol. Des.* **2007**, 21 (1-3), 113-119.
22. Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Diversity* **2006**, 10 (3), 389-403.

23. Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. USA* **2005**, *102* (48), 17272-17277.
24. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47* (1), 47-58.
25. Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5* (8), 581-583.
26. Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177-182.
27. Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8* (3), 255-263.
28. Gorse, A. D. Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* **2006**, *6* (1), 3-18.
29. Andrews, K.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43* (9), 1723-1740.
30. Todeschini, R.; Consonni, V., Molecular Descriptors for Chemoinformatics. In *Methods and Principles in Medicinal Chemistry* Mannhold, R.; Kubinyi, H.; Folkers, G., Eds. Wiley-VCH: Weinheim, 2009; Vol. 41.
31. Dunbar Jr, J. B. Cluster-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 51-63.
32. Godden, J. W.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 87-93.
33. Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (4), 1060-1066.
34. Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 796-800.
35. Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65-84.
36. Landon, M. R.; Schaus, S. E. JEDA: Joint entropy diversity analysis. An information-theoretic method for choosing diverse and representative subsets from combinatorial libraries. *Mol. Diversity* **2006**, *10* (3), 333-339.
37. Mason, J. S.; Pickett, S. D. Partition-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85-114.
38. Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (3), 550-558.
39. Vogt, I.; Bajorath, J. Design and exploration of target-selective chemical space representations. *J. Chem. Inf. Model.* **2008**, *48* (7), 1389-1395.
40. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983-996.
41. Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* **2000**, *3* (5), 363-372.
42. Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 801-809.
43. Xue, L.; Godden, J. W.; Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (5), 1227-1234.
44. Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing Similarity and Diversity of Combinatorial Libraries by Spatial Autocorrelation Functions and Neural Networks. *Angew. Chem. Int. Ed. Engl.* **1996**, *34* (23-24), 2674-2677.

45. Maniyar, D. M.; Nabney, I. T.; Williams, B. S.; Sewing, A. Data Visualization during the Early Stages of Drug Discovery. *J. Chem. Inf. Model.* **2006**, *46* (4), 1806-1818.
46. Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-based Data-fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chem. Biol. Drug Des.* **2007**, *70* (5), 393-412.
47. Oprea, T. I.; Zamora, I.; Ungell, A.-L. Pharmacokinetically Based Mapping Device for Chemical Space Navigation. *J. Comb. Chem.* **2002**, *4* (4), 258-266.
48. Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space. *J. Nat. Prod.* **2007**, *70* (5), 789-794.
49. Rosén, J.; Lövgren, A.; Kogej, T.; Muresan, S.; Gottfries, J.; Backlund, A. ChemGPS-NPWeb: chemical space navigation online. *J. Comput.-Aided Mol. Des.* **2009**, *23* (4), 253-259.
50. Lloyd, D. G.; Golfis, G.; Knox, A. J. S.; Fayne, D.; J., M. M.; Oprea, T. I. Oncology exploration: charting cancer medicinal chemistry space. *Drug Discovery Today* **2006**, *11* (3/4), 149-159.
51. Shelat, A. A.; Guy, R. K. The interdependence between screening methods and screening libraries. *Curr. Opin. Chem. Biol.* **2007**, *11* (3), 244-251.
52. Macchiarulo, A.; Pellicciari, R. Exploring the other side of biologically relevant chemical space: Insights into carboxylic, sulfonic and phosphonic acid bioisosteric relationships. *J. Mol. Graphics Modell.* **2007**, *26* (4), 728-739.
53. Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49* (4), 1010-1024.
54. Dubois, J.; Bourg, S.; Vrain, C.; Morin-Allory, L. Collections of Compounds - How to Deal with them? *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 156-168.
55. Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 643-651.
56. Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29* (1), 55-67.
57. Verheij, H. J. Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Diversity* **2006**, *10* (3), 377-388.
58. Potter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41* (4), 478-488.
59. Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J.; Jacoby, E. Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. High Throughput Screen.* **2004**, *7* (8), 771-781.
60. Jacoby, E.; Schuffenhauer, A.; Popov, M.; Azzaoui, K.; Havill, B.; Schopfer, U.; Engeloch, C.; Stanek, J.; Acklin, P.; Rigollier, P.; Stoll, F.; Koch, G.; Meier, P.; Orain, D.; Giger, R.; Hinrichs, J.; Malagu, K.; Zimmermann, J.; Roth, H. J. Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr. Top. Med. Chem.* **2005**, *5* (4), 397-411.
61. Crisman, T. J.; Jenkins, J. L.; Parker, C. N.; Hill, W. A.; Bender, A.; Deng, Z.; Nettles, J. H.; Davies, J. W.; Glick, M. "Plate cherry picking": a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J. Biomol. Screen.* **2007**, *12* (3), 320-327.
62. Valler, M. J.; Green, D. Diversity screening versus focussed screening in drug discovery. *Drug Discovery Today* **2000**, *5* (7), 286-293.
63. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862-865.
64. Hamprecht, F. A.; Thiel, W.; van Gunsteren, W. F. Chemical library subset selection algorithms: a unified derivation using spatial statistics. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (2), 414-428.
65. Willett, P. Chemoinformatics - similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **2000**, *11* (1), 85-88.

66. Sukuru, S. C.; Jenkins, J. L.; Beckwith, R. E.; Scheiber, J.; Bender, A.; Mikhailov, D.; Davies, J. W.; Glick, M. Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity. *J. Biomol. Screen.* **2009**, *14* (6), 690-699.
67. DRCS Tools; ICOA-CNRS: Orleans France. <http://www.univ-orleans.fr/icoa/DRCS/>; (accessed January 15, 2011)
68. Bologna, C. G.; Olah, M. M.; Oprea, T. I. Chemical database preparation for compound acquisition or virtual screening. *Methods Mol. Biol.* **2006**, *316*, 375-388.
69. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189-1204.
70. Pipeline Pilot, student edition, Accelrys, San Diego, CA, 2010.
71. InChI, 1.03, IUPAC, 2010. InChI. <http://www.iupac.org/inchi/> (accessed January 15, 2011)
72. Originality is a relative concept which depends both on the list of products of the provider but also on the list of compounds of the other providers. If the compounds of a provider are remarketed by another one of the list then the originality connected to these compounds is null for both providers. It is why the comparison with the results obtained in previous work is without object. It is also why this value cannot be used as a commercial argument (either positive or negative)
73. MOE, version 2009-10; Chemical Computing Group; Montreal, Quebec, Canada
74. Oprea, T. I. Property distribution of drug-related chemical databases*. *J. Comput.-Aided Mol. Des.* **2000**, *14* (3), 251-264.
75. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1177-1185.
76. Lee, S.; Huang, J. Z.; Hu, J. Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **2010**, *4* (3), 1579-1601.
77. Nikolaj, T. In *What is the Dimension of Your Binary Data?*, 6th IEEE International Conference on Data Mining (ICDM'06), Hong Kong, China, 2006; Taneli, M.; Aristides, G.; Heikki, M., Eds. IEEE Computer Society: Los Alamitos, CA, USA, 2006; 603-612.
78. Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **2001**, *22* (5), 488-500.
79. Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18* (4-5), 464-477.
80. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493-500.
81. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12* (17), 2111-2120.
82. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5-6), 490-519.
83. Graham, R. L. An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. *Inf. Process. Lett.* **1972**, *1*, 132-133.
84. Leopold, D. A.; Bondar, I. V.; Giese, M. A. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **2006**, *442* (7102), 572-575.
85. Baskin, I. I.; Kireeva, N.; Varnek, A. The One-Class Classification Approach to Data Description and to Models Applicability Domain. *Mol. Inf.* **2010**, *29* (8-9), 581-587.
86. Chembridge. <http://www.chembridge.com> (accessed January 15, 2011)
87. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W623-633.

88. Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43* (21), 3867-3877.
89. CMC. <http://www.akosgmbh.de/Symyx/software/databases/cmc-3d.htm> (accessed January 15, 2011)
90. Prestwick. <http://www.prestwickchemical.com/> (accessed January 15, 2011)
91. Pyxis. <https://www.chemonaut.com> (accessed January 15, 2011)
92. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789-6801.
93. DUD. <http://dud.docking.org/> (accessed January 15, 2011)
94. Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Del. Rev.* **1997**, *23*, 3-25.
95. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876-877.
96. ScreeningAssistant2; ICOA-CNRS: Orleans France. <http://www.univ-orleans.fr/icoa/modelisation/index.php?h=2;> (accessed June 15, 2011)
97. CDK. <http://sourceforge.net/projects/cdk/> (accessed January 15, 2011)
98. Spjuth, O.; Helmus, T.; Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* **2007**, *8*, 59.
99. RDKit. <http://rdkit.org/> (accessed January 15, 2011)
100. Kuhn, T.; Willighagen, E.; Zielesny, A.; Steinbeck, C. CDK-Taverna: an open workflow environment for cheminformatics. *BMC Bioinformatics*. **2010**, *11*, 159.
101. KNIME. <http://www.knime.org/> (accessed January 15, 2011)

Visual characterization and diversity quantification of chemical libraries.

1) Creation of Delimited Reference Chemical Subspaces

*Vincent Le Guilloux ^a, Lionel Colliandre ^a, Stéphane Bourg ^b, Guillaume Guénégou ^a, Julie Dubois-
Chevalier ^{a,c}, Luc Morin-Allory ^{a*}*

Supporting Information

N°	Provider's Name	Internet Site	Download	3 letter code
1	ACB Blocks, Ltd.	www.acbblocks.com	08-2010	ACB
2	Adesis, Inc.	www.adesisinc.com	09-2010	ADS
3	AF ChemPharm Limited	www.afchempharm.co.uk	08-2010	AFC
4	Albany Molecular Research, Inc.	www.amriglobal.com	01-2010	AMR
5	Alinda Chemical Limited	alinda.ru	01-2010	ALC
6	AllLab	www.albchemical.com	03-2010	ALB
7	Ambinter, SARL	www.ambinter.com	01-2010	AMB
8	AnalytiCon Discovery GmbH	www.ac-discovery.com	09-2009	ACD
9	Apollo Scientific, Ltd.	www.apolloscientific.co.uk	01-2010	APS
10	Arkive Database	www.ark.chem.ufl.edu	08-2010	ARK
11	Aronis	www.aronis.ru	01-2010	ARO
12	Arvi Co. Ltd.	www.ar-vi.com	03-2010	ARV
13	Asinex, Ltd.	www.asinex.com	09-2009	ASI
14	AsisChem, Inc.	www.asischem.com	09-2010	ASC
15	Azasynt	www.azasynt.com	01-2010	AZS
16	Bionet Database	www.keyorganics.ltd.uk	10-2009	BIN
17	Biosynt	www.biosynt.com	05-2010	BIS
18	Biotrend Chemikalien GmbH	www.biotrend.com	03-2010	BIT
19	Cayman Chemical Company	www.caymanchem.com	08-2010	CAY
20	ChemBridge Corporation	www.chembridge.com	08-2010	CBC
21	ChemDiv, Inc.	us.chemdiv.com	11-2009	CDV
22	Chemical Block, Ltd.	www.chemblock.com	08-2010	CBL
23	Chemik Co. Ltd.	www.chemik.com	09-2010	CKC
24	Chemivate Limited	chemivate.com	01-2010	CVT
25	Chemical Technologies & Investigations, Ltd.	www.chemti.com	08-2010	CTI
26	Chess GmbH	www.chess-chem.com	03-2010	CSS
27	Combi-Blocks, Inc.	www.combi-blocks.com	09-2010	CBK

28	EMC microcollections GmbH	www.microcollections.de	09-2009	EMC
29	Enamine, Ltd.	www.enamine.net	08-2010	ENA
30	Endeavour Speciality Chemicals, Ltd.	www.endeavourchem.co.uk	01-2010	ESC
31	Endotherm GmbH	www.endotherm-lsm.com	03-2010	EDT
32	Enzo Biochem Inc.	www.enzolifesciences.com	08-2010	EZB
33	Exclusive Chemistry, Ltd.	www.exchemistry.com	10-2009	EXC
34	Fluorochem	www.fluorochem.net	08-2010	FLU
35	Focus Synthesis LLC.	www.focussynthesis.com	01-2010	FOC
36	French National Library	chimiotheque-nationale.enscm.fr	08-2010	FNL
37	Frontier Scientific, Inc.	www.frontiersci.com	11-2009	FRO
38	Greenpharma S.A.	www.greenpharma.com	03-2010	GRE
39	InFarmatik	www.infarmatik.com	12-2009	INF
40	InterBioScreen, Ltd.	www.ibscreen.com	08-2010	IBS
41	Intermed, Ltd.	www.intermedchemicals.com	01-2010	INT
42	IS Chemical Technology Ltd.	www.ispharm.com	09-2010	ISC
43	KaïronKem	www.kaironkem.com	03-2010	KAK
44	Key Organics, Ltd.	www.keyorganics.ltd.uk	08-2010	KEO
45	LaboTest	www.labotest.com	09-2009	LBT
46	Life Chemicals, Inc.	www.lifechemicals.com	07-2010	LIF
47	Matrix Scientific	www.matrixscientific.com	11-2009	MAT
48	Maybridge	www.maybridge.com	09-2010	MAY
49	Menai Organics Limited	menaiorganics.co.uk	03-2010	MEN
50	MyriaScreen Collection	www.sigmaaldrich.com	08-2010	MYS
51	Nanosyn	www.nanosyn.com	09-2009	NAN
52	OakwoodProducts, Inc.	www.oakwoodchemical.com	09-2010	OAK
53	Otava, Ltd.	www.otavachemicals.com	09-2009	OTA
54	Peakdale Molecular, Ltd.	www.peakdale.com	09-2009	PEM
55	PepTech Corporation	www.peptechcorp.com	09-2010	PTC

56	Pharmeks	www.pharmeks.com	09-2009	PHA
57	Prestwick Chemical, Inc.	www.prestwickchemical.com	09-2009	PWK
58	Princeton BioMolecular Research, Inc.	www.princetonbio.com	09-2009	PBM
59	Pyxis Discovery	www.pyxis-discovery.com	08-2010	PYX
60	Scientific Exchange, Inc.	www.htscompounds.com	08-2010	SCE
61	Sequoia Research Products, Ltd.	www.seqchem.com	01-2010	SRP
62	Shanghai Sinofluoro Scientific Co., Ltd.	www.sinofluoro.com	09-2010	SFS
63	Sinova Inc.	www.sinovainc.com	09-2010	SIN
64	Specs	www.specs.net	08-2010	SPE
65	Spectrum Info, Ltd.	www.spectrum.kiev.ua	09-2009	SPI
66	SynChem, Inc.	www.synchem.com	11-2009	SYC
67	SynphaBase AG	www.synphabase.com	09-2009	SYB
68	Synthon-Lab, Ltd.	www.synthon-lab.com	05-2010	SYL
69	Szintekon Co, Ltd.	szintekon.hu	03-2010	SZI
70	TimTec, Inc.	www.timtec.net	08-2010	TIM
71	Toronto Research Chemicals, Inc.	www.trc-canada.com	09-2010	TRC
72	TOSLab	www.toslab.com	10-2009	TOS
73	Vitas-M Laboratory, Ltd	www.vitasmlab.com	08-2010	VML

Table S1 : list of providers from which chemical libraries were collected. Websites are provided as well as the dates at which each library was downloaded. A unique 3-letter code has been assigned to each provider and used in Table S2.

N°	3 letters code	Compounds	Treated and 3D converted compounds	Internal Duplicates		Unique compounds	Absolute Originality ¹
		#	#	#	(ppm)	#	(%)
1	ACB	3 250	3 238	6	(1 853)	3 232	1.39
2	ADS	1 175	1 174	1	(852)	1 173	28.99
3	AFC	646	622	5	(8 039)	617	74.07
4	AMR	510 372	509 234	254 623	(500 012)	254 611	1.09
5	ALC	254 120	253 870	3	(12)	253 867	2.43
6	ALB	10 011	10 011	13	(1 299)	9 998	0.07
7	AMB	5 339 256	5 333 597	129 282	(24 239)	5 204 314	29.41
8	ACD	26 603	26 534	107	(4 033)	26 427	0.01
9	APS	43 109	41 672	2 690	(64 552)	38 982	0.26
10	ARK	32 060	31 818	1 307	(41 077)	30 511	0.09
11	ARO	23 742	23 736	29	(1 222)	23 707	0.00
12	ARV	13 084	13 042	3 082	(236 313)	9 960	4.50
13	ASI	389 837	389 785	101	(259)	389 684	37.80
14	ASC	234 509	234 454	369	(1 574)	234 085	14.31
15	AZS	67	67	0	(0)	67	100.00
16	BIN	42 683	42 661	21	(492)	42 640	0.00
17	BIS	2 089	2 080	276	(132 692)	1 804	42.18
18	BIT	1 604	1 363	33	(24 211)	1 330	38.57
19	CAY	2 087	1 914	135	(70 533)	1 779	67.12
20	CBC	633 732	633 341	106	(167)	633 235	2.79
21	CDV	663 496	663 432	42	(63)	663 390	54.97
22	CBL	125 465	125 374	265	(2 114)	125 109	0.89
23	CKC	4 098	3 979	219	(55 039)	3 760	6.41
24	CVT	1 108	1 108	0	(0)	1 108	98.92
25	CTI	171 031	171 018	266	(1 555)	170 752	3.07

26	CSS	566	566	3	(5 300)	563	1.07
27	CBK	7 199	7 180	93	(12 953)	7 087	29.24
28	EMC	29 381	29 075	978	(33 637)	28 097	98.91
29	ENA	1 516 877	1 516 786	23	(15)	1 516 762	9.30
30	ESC	817	816	3	(3 676)	813	20.79
31	EDT	697	695	18	(25 899)	677	72.23
32	EZB	203	202	0	(0)	202	47.03
33	EXC	2 273	2 273	1	(440)	2 272	67.21
34	FLU	41 379	35 440	995	(28 076)	34 445	20.83
35	FOC	2 391	2 371	50	(21 088)	2 321	44.68
36	FNL	43 012	42 677	662	(15 512)	42 015	85.26
37	FRO	1 768	1 725	40	(23 188)	1 685	20.95
38	GRE	657	652	5	(7 669)	647	1.24
39	INF	1 385	1 384	11	(7 948)	1 373	0.00
40	IBS	474 811	473 953	10 557	(22 274)	463 396	1.41
41	INT	32 042	32 042	1	(31)	32 041	0.02
42	ISC	27 701	27 539	2 483	(90 163)	25 056	26.14
43	KAK	652	649	2	(3 082)	647	38.64
44	KEO	42 690	42 668	21	(492)	42 647	0.14
45	LBT	106 743	106 556	1 319	(12 378)	105 237	36.94
46	LIF	357 319	357 291	23 657	(66 212)	333 634	2.77
47	MAT	39 325	39 269	948	(24 141)	38 321	13.39
48	MAY	116 941	116 913	56 522	(483 454)	60 391	50.52
49	MEN	4 014	4 013	1	(249)	4 012	0.00
50	MYS	10 000	9 995	13	(1 301)	9 982	11.10
51	NAN	65 328	65 286	245	(3 753)	65 041	20.07
52	OAK	13 473	13 366	157	(11 746)	13 209	1.25
53	OTA	120 736	120 370	269	(2 235)	120 101	10.00
54	PEM	14 628	14 625	4	(274)	14 621	0.00

55	PTC	2 381	2 380	4	(1 681)	2 376	48.95
56	PHA	228 464	228 137	794	(3 480)	227 343	0.02
57	PWK	1 200	1 187	4	(3 370)	1 183	8.54
58	PBM	979 604	979 109	7 533	(7 694)	971 576	14.70
59	PYX	317	317	0	(0)	317	97.16
60	SCE	1 207 277	1 207 069	22 891	(18 964)	1 184 178	56.70
61	SRP	2 323	2 225	90	(40 449)	2 135	22.81
62	SFS	279	276	13	(47 101)	263	5.32
63	SIN	3 979	3 844	48	(12 487)	3 793	62.51
64	SPE	885 698	885 397	414 150	(467 756)	471 247	4.44
65	SPI	8 678	8 678	5	(576)	8 673	2.24
66	SYC	1 576	1 559	13	(8 339)	1 546	28.27
67	SYB	509	495	5	(10 101)	490	1.02
68	SYL	52 760	52 758	45	(853)	52 713	1.80
69	SZI	2 716	2 707	43	(15 885)	2 664	65.73
70	TIM	205 704	205 482	637	(3 100)	204 845	0.65
71	TRC	17 651	13 822	563	(40 732)	13 259	64.89
72	TOS	16 657	16 575	18	(1 086)	16 557	0.49
73	VML	846 862	846 539	6 704	(7 919)	839 835	0.15

Table S2: for each provider, number of compounds remaining after the main steps of the standardization protocol.

Set	Descriptor names	Descriptor count	Reason for removal
MOE3D	E_rele, E_rnb , E_rsol, E_rvdw, pmiX, pmiY, pmiZ, dipoleX, dipoleY, dipoleZ	10	External coordinates
	AM1_*, MNDO_*, PM3_*	21	Quantum descriptors
MOE2D	PC+, PC-, RPC+, RPC-	4	Duplicate descriptors
	lip_druglike, lip_violation, opr_leadlike, opr_violation, rsynth	5	Meta-descriptors
	reactive, nmol	2	Null variance
CDK2D	Lipinski	1	Meta descriptor
	khs.aaSe, khs.dSe, khs.ddssSe, khs.dssSe, khs.sAsH2, khs.sGeH3, khs.sLi, khs.sNH3, khs.sPH2, khs.sPbH3, khs.sSeH, khs.sSnH3, khs.ssAsH, khs.ssBH, khs.ssBe, khs.ssGeH2, khs.ssNH2, khs.ssPH, khs.ssPbH2, khs.ssSiH2, khs.ssSnH2, khs.sssAs, khs.sssGeH, khs.sssNH, khs.sssP, khs.sssPbH, khs.sssSnH, khs.sssdAs, khs.ssssBe, khs.ssssGe, khs.ssssPb, khs.ssssSn, khs.sssssAs, khs.sssssP, nH	34	Null variance

Table S3: detailed list of descriptors removed for each descriptor set. Descriptors having null variance and meta-descriptors which do not correspond to intrinsic chemical or physico-chemical properties (e.g. Lipinski or Oprea flags) were removed. Descriptors based on the absolute coordinates of each molecule were removed from the original set (e.g. the X component of the dipole moment, which depends on the absolute orientation of the molecule), as well as all semi-empirical descriptors which require a prohibitive calculation time for such a large number of molecules.

DETAILED PREPARATION OF CHEMICAL STRUCTURES

The following 11-step Pipeline Pilot² protocol was applied to obtain the final standardized 3D structures. The proportion of compounds eliminated in each step is given in parts per million (ppm).

1. First, compounds with a null molecular weight are removed (442 ppm). These compounds correspond to entries in the SD file for which only properties are defined, without the actual structure.
2. Molecules containing non-natural isotopic atoms are removed (302 ppm). These atoms may be problematic for descriptor calculation; e.g. deuterium atoms (with specific atom type “D”) are read as carbon atoms in the MOE software (2009.10).³ The identification of this problem is very difficult as no error message is displayed, while descriptor values may not seem aberrant at all.
3. Molecules containing aliases are removed (97 ppm). Specific information on structures can be incorporated into the “properties block” of an SD file under the form of a so-called “alias”. These “aliases” are often used to represent commonly used groups such as protecting groups (Boc, Fmoc), and may be problematic because they do not correspond to atoms in the coordinate block of the SD file. A reliable automatic transformation into real atoms with coordinates is difficult. Practically, Pipeline Pilot can detect these aliases as “non standard atom type”.
4. 2D coordinates are generated for all molecules to obtain the same starting point.
5. The largest fragment is kept for each SD file entry. This ensures that each entry of the libraries is composed of a single molecule and that no counter-ions are present. We are aware that there may be some very rare cases where the counter-ion might be of slightly higher molecular weight than the actual compound. Such cases are difficult to identify and handle, however, especially for such a large dataset.

6. Each molecule is then “kekulized”, to avoid aromatic flag sometimes occurring in the SD files of some chemical libraries. According to the standard SD format, aromatized structures are meant to be used for substructure search queries, and should not be used for structure definition. In particular, the IUPAC official InChITM program⁴ will reject aromatized compounds. However, this kekulization process may sometimes fail, e.g. when heteroatoms have to be deprotonated, or when a formal charge has to be added. This type of error represents 1.3 ppm of the database.
7. Hydrogen atoms are added, and protonation states are reassigned; for example, acidic functions are protonated and basic functions are deprotonated. This avoids uncertainties, and thus potential errors, related to the assignment of protonation states at a given pH, which requires in-silico pKa predictions.
8. Several specific chemotypes are normalized: metal atoms (Li, Na, K) covalently bonded to oxygen or sulfur atoms are removed; the bond valences of nitro, diazo, azido, isonitril, sulfone and benzothiadiazol groups are standardized; atoms, covalently bonded to oxygen atoms, with the erroneous “Ac” atom type are expanded to the corresponding acetyl group.
9. Molecules containing the following atoms are retained: H, C, N, O, P, S, F, Cl, Br, I, B, Si and Se. Chemical descriptors, especially those based on molecular forcefield, are rarely defined for excluded atoms. Obviously, this step creates a bias in the definition of the DRCS. Nevertheless, the proportion of compounds removed at this step remains fairly low (218 ppm), and we believe that such a bias is preferable to undetectable errors occurring during the descriptor calculation.
10. Compounds still containing erroneous valences are removed (39 ppm).
11. Lastly, we assign 3D coordinates to each structure using the Corina program⁵. This program is a gold standard in 3D structure generation for small molecules. It allows a fast and robust calculation of 3D coordinates and is well suited for large chemical libraries. Stereochemistries as defined in the original file are conserved during this step. 3D coordinates generation failed

for 446 ppm of the compounds. Errors appear mainly for molecules containing macrocycles, non-aromatics multicyles or sugar moieties.

TOOLS AND COMPUTATIONAL PERFORMANCES

All the programs and algorithms were developed in-house using the JAVA programming language. The PCA was implemented using the JAMA library⁶. Each convex hull was computed using the Graham scan algorithm⁷ (algorithm complexity in $O(n \log(n))$). The programs and source code are freely available and documented⁸, and the methodology will be bundled in the Screening Assistant 2 open-source platform⁹ (paper and release in preparation).

All calculations were performed on recent computers with Intel Core I7 2.93GHz processor and 6GO DDR3 RAM (1,375GHz). The 2D MOE descriptors calculation took nearly 36 hours for all 6.6 million compounds. 2D CDK and 3D MOE descriptors were calculated for the 600 000 compounds used to build each DRCS, and took approximately 3 and 4 days, respectively.

DETERMINATION OF THE NUMBER OF SUBSETS AND MOLECULES PER SUBSET

A global PCA Model was computed on the whole MOE 2D dataset (6.6 million molecules). This model was used as a reference to determine if a similar PCA model could be obtained by averaging multiple PCA models computed on reduced random subsets.

Method

First of all, given N subsets of M molecules, an averaged PCA model is computed as the average of each PCA model computed on the N subsets. For a given subset, the eigenvectors and corresponding eigenvalues are computed on the correlation matrix derived from the scaled descriptors matrix. The eigenvectors of each PCA model are then sorted by decreasing eigenvalues. The average model is

obtained by averaging all eigenvectors and all the corresponding eigenvalues. Sign inversion in PC might randomly occur during the calculation, and may introduce errors in the final averaged model. To detect such an inversion for a given PC, the molecule having the lowest coordinate in this PC is retrieved for the model computed on the first subset, and the sign of this coordinate is defined as the reference. Sign inversion is subsequently detected for a given PC when the sign of the coordinates of this compound is reversed compared to the reference one. In such case, all the signs of the corresponding eigenvectors are inverted.

Parameters and model comparison

To compute the averaged model, one needs to define the number of subsets and the number of molecules per subset. The number of subsets was first empirically defined as 30, which is in our opinion a good compromise between computational cost and statistical significance. The number of molecules was subsequently determined by comparing the global model and averaged models computed using different numbers of molecules per subset, ranging from 1000 to 100000. The idea was to pick a number of molecules leading to a similar model while being small enough to allow the computation of time-consuming descriptors in an acceptable time scale.

To the best of our knowledge, there is no established method to assess the similarity between several PCA models computed using the same set of variables. In the present case, PCA models are basically defined by the % of explained variance on each PC, and by the composition of the eigenvectors which allows the calculation of the reduced space coordinates. These coordinates can subsequently be used to compare two models: a strong linear correlation between the coordinates of a reference set of molecules in both models may suggest a strong similarity between the two models for the PC under consideration. The following two criteria were thus used to assess the similarity between each consensus model and the global PCA model:

1. The % of explained variance. For a consensus model and a given component, it is determined as the average explained variance of all 30 PCAs for that component.

2. The correlation between the reduced coordinates of a set of compounds computed using the global PCA and using the consensus model. For a given component, reduced coordinates are computed as the weighted sum of scaled descriptor values.

Results

Both % of explained variance and correlation between coordinates were monitored for different numbers of molecules used to compute each averaged model. Although one usually uses at most the first three components for visualization purposes, we collected data for the first five PC. Figure S3a shows that the number of molecules per subset has almost no effect on the percentage of explained variance for all PC, which remains practically the same even with few molecules in each subset. A similar observation can be made for the correlation coefficient, as shown in Figure S3b. Although a few variations seem to appear (mainly for the fifth PC with small datasets), the correlations observed remain almost equal to 1 for all PC.

We also note that more variations were observed, especially when considering PC with smaller explained variances (data not shown). After careful analysis, we identified a few dozens of molecules having very extreme values – up to 1000 times the standard deviation – for the weinerPath descriptor. When some of these molecules are present in one or several subsets, correlations decrease to a greater extent than observed here for the first five PC. This is even truer when the number of molecules per subset is small. Care must be taken for such phenomena, especially when the number of molecules and the number of descriptors is small. The averaging procedure used here clearly smooths the influence of these extreme values, allowing very similar models to be obtained for the first components. Data pre-processing is again highlighted as a crucial step in any statistical analysis.

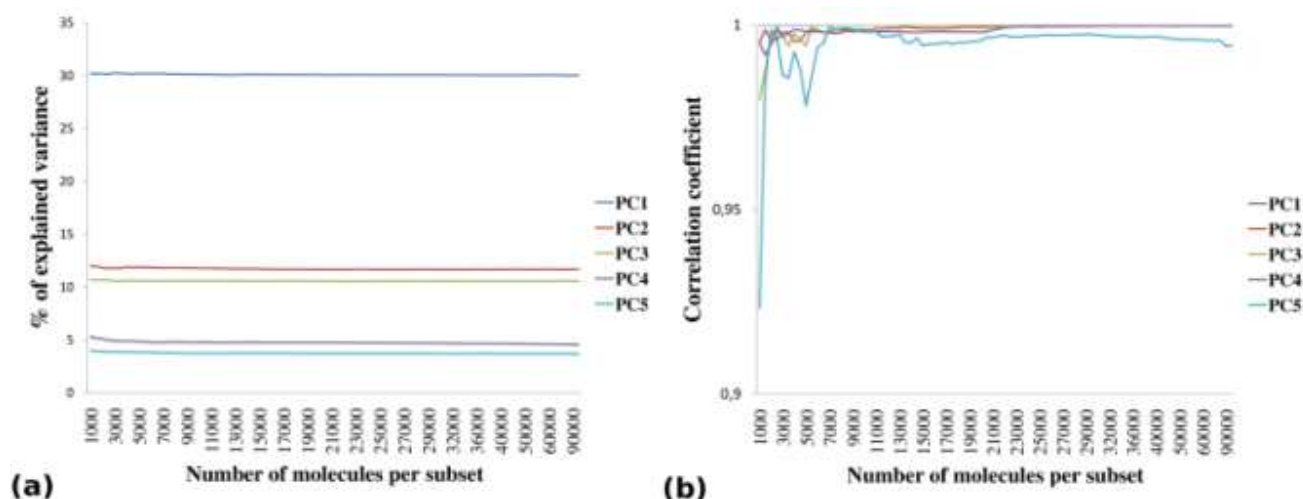


Figure S3: (a) Variation in the proportion of explained variance for the first five PC of average DRCS models computed with different numbers of molecules per subset (MOE 2D dataset). (b) Variation in the correlation coefficient between coordinates of molecules in the global and averaged model computed with different numbers of molecules per subset. For a given number of molecules, the first five PC coordinates of the dataset were computed using (i) the global MOE2D model and (ii) the averaged model. The correlation coefficient between the two sets of coordinates was then calculated for each PC.

Nevertheless, it can be concluded that averaged models very similar to the global model can be obtained, and that the number of molecules has no significant influence for the first PCs considered here, though care must be taken concerning extreme descriptor values. The number of molecules per subset was empirically fixed at 20 000. These 30 subsets of 20 000 molecules finally represent around 11 % of the original dataset. Files MOE2DvsGlobal_PC#.pdf in the supplementary material show a detailed comparison of PCs loadings for the final DRCS MOE2D model obtained and used in this study. It further confirms that they are almost identical.

PARAMETRISATION OF THE CONTOUR CALCULATION

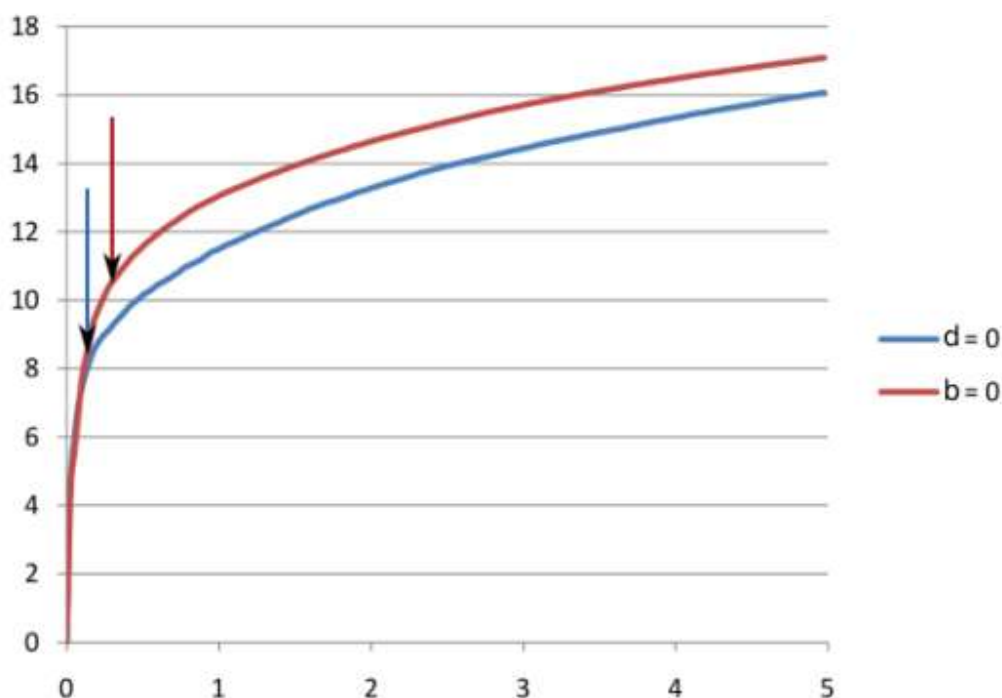


Figure S4: RMSD between the consensus hull computed without any outlier removal ($b = d = 0$) and consensus hulls computed with an increasing % of molecules removed using the barycenter method (blue) and the density method (red). Stability seems to appear slightly sooner with the barycenter-based method (see respective arrows).

This section is dedicated to further investigating the advantage of combining the barycenter and the density outlier removal methods to obtain a stable contour while minimizing the number of molecules removed, and to select appropriate parameter values using a more rational approach. To this end, the RMSD between each MOE2D consensus contour shown in figure 1 of the article, and the initial hull computed without outlier removal (where $d = b = 0$) were computed. Again, the barycenter method alone (figure S4, blue curve) and the density method alone (figure S4, red curve) were successively used. Two distinct behaviors can be identified in both cases: a rapid increase in the RMSD occurs for a small percentage of outliers removed, and a slow and monotonic increase in the RMSD value indicating

the convergence of the shape. From this graphic, one can see that convergence seems to occur for a slightly higher percentage of outliers removed when the density based method is used alone. We suggest that the reason for this is that the density method depends on the distance D defined to calculate the neighborhood density. This distance may or may not be appropriate, depending on the overall neighborhood density and distribution of compounds in the reduced space under consideration. If D is defined too small or too large, a molecule isolated and located in extreme regions of the space may have the same density value as a molecule located in a denser and more centered region of the space. The removal of molecules using the barycenter prior to the density method ensures that the small proportion of molecules located in extreme zones is removed regardless of the D value, thus decreasing the potential influence of an inappropriate value. We therefore decided to calibrate these parameters on the MOE2D dataset, and use the resulting parameters for the other datasets to keep the same ratio of molecules removed. Empirical visual analysis leads to defining b to 0.05 % and D to 3. d was subsequently set to 0.20 %. For a different dataset, the optimal values may be different.

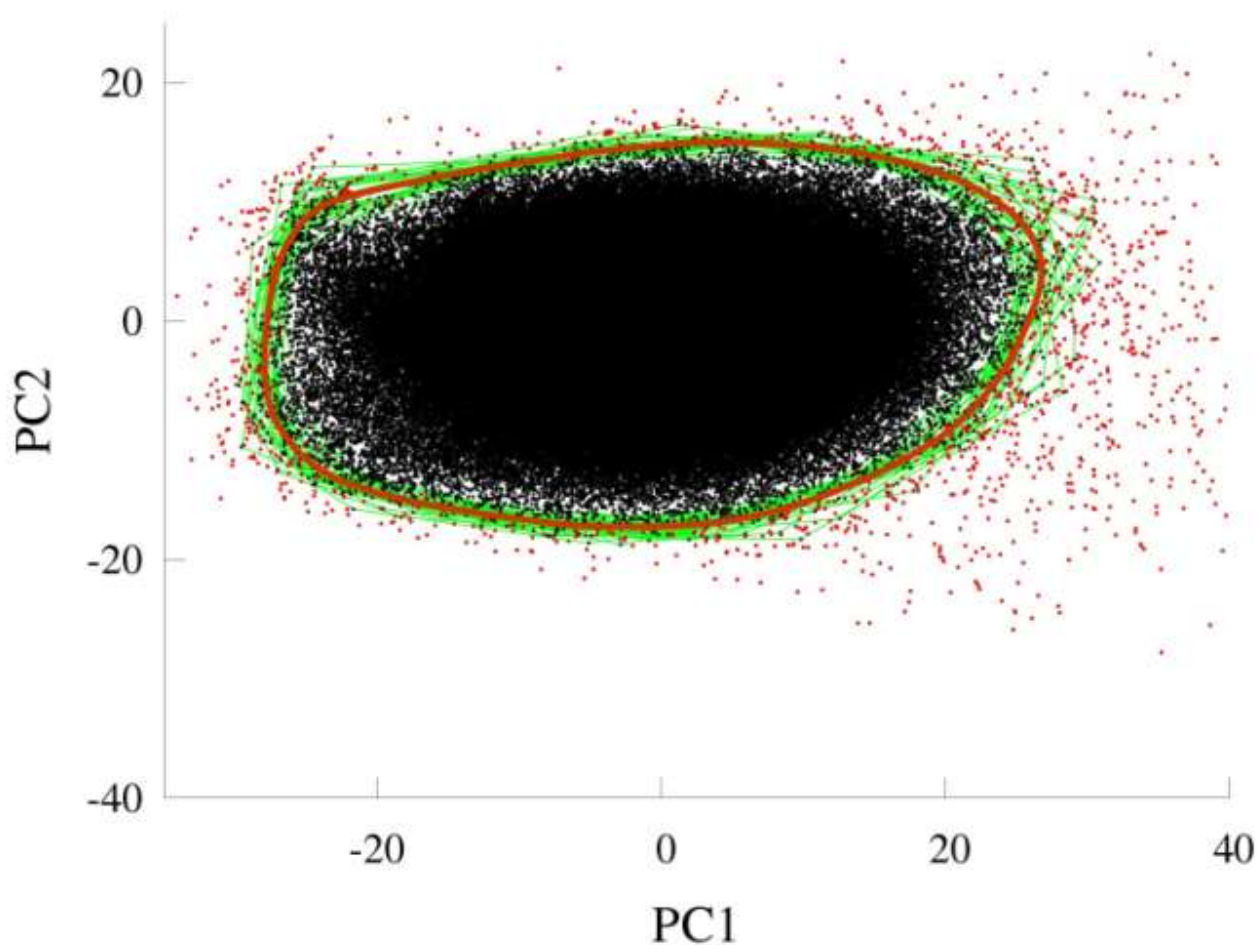


Figure S5: Projection of the 30 subsets (for a total of 600 000 molecules) used to compute the MOE2D space. The 30 roughly-shaped convex hulls obtained for each subset (in green) and the resulting smooth contour (orange) obtained using the presented methodology highlight the interest of the averaging operation.

References of supporting information

1. Originality is a relative concept which depends both on the list of products of the provider but also on the list of compounds of the other providers. If the compounds of a provider are remarketed by another one of the list then the originality connected to these compounds is null for both providers. It is why the comparison with the results obtained in previous work is without object. It is also why this value cannot be used as a commercial argument (either positive or negative)
2. *Pipeline Pilot*, version 6.1.5.0 student edition, Accelrys, San Diego, CA, 2010.
3. *MOE*, version 2009-10; Chemical Computing Group; Montreal, Quebec, Canada
4. *InChI*, 1.03, IUPAC, 2010. INCHI. <http://www.iupac.org/inchi/> (accessed January 15, 2011)
5. Sadowski, J.; Gasteiger, J., From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Reviews* **1993**, *93*, 2567-2581.
6. Javanumerics. <http://math.nist.gov/javanumerics/>. (accessed January 15, 2011)
7. Graham, R. L., An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. *Inform. Process. Lett.* **1972**, *1*, 132-133.
8. DRCS Tools; ICOA-CNRS: Orleans France. <http://www.univ-orleans.fr/icoa/DRCS/>; (accessed January 15, 2011)
9. ScreeningAssistant; <http://www.univ-orleans.fr/icoa/screeningassistant/>; (accessed January 15, 2011)

Chapitre III

Quantification de la diversité à l'aide d'espaces chimiques délimités

1 Introduction

Dans ce chapitre et le chapitre suivant, nous nous intéressons cette fois à l'évaluation numérique de la diversité de chimiothèques. Bien que la projection des molécules dans des espaces DRCS permette une évaluation rapide de la couverture d'une chimiothèque dans l'espace considéré, elle ne permet pas de statuer lorsque ce recouvrement est visuellement similaire pour plusieurs chimiothèques. Les indices de diversité apportent généralement une information complémentaire sur la manière dont l'espace de référence est peuplé et permettent notamment d'évaluer à quel point cette couverture est uniforme.

Nous avons dans le chapitre 1 décrit les différents indices de diversités classiquement trouvés dans la littérature. Pour les indices reposant sur l'évaluation de la couverture d'un espace chimique donné, l'une des problématiques les plus récurrentes est de trouver un moyen de délimiter cet espace. Par définition, les DRCS apportent une solution à cette problématique en délimitant les zones les plus denses couvertes par un sous-ensemble de molécules donné. Nous avons ainsi à nouveau illustré l'utilisation des DRCS afin de générer de nouveaux sous-espaces délimités correspondant à divers types de molécules : les molécules "drug-like" (selon les règles de Lipinski), pharmaceutiques (basées sur une compilation de molécules issues de la DrugBank et de la Comprehensive Medicinal Chemistry), ou adaptées au criblage par fragments.

La méthodologie des DRCS a ensuite été appliquée au calcul d'indices permettant d'évaluer la diversité d'une ou plusieurs chimiothèques par rapport à un sous-espace de référence. Dans un premier temps, les règles de Waldman, qui définissent le comportement idéal d'une métrique

Chapitre III. Quantification de la diversité à l'aide d'espaces chimiques délimités

évaluant la diversité d'une chimiothèque, ont été étendues afin de prendre en compte le cas où des chimiothèques de taille différentes doivent être comparées. Nous avons ainsi défini de nouvelles règles décrivant le comportement idéal d'un indice *relatif* de diversité.

22 indices de diversité ont ensuite été ré-implémentés et adaptés à la méthodologie des DRCS. La délimitation définie par l'enveloppe convexe moyenne est ici utilisée afin de définir les limites des sous-espaces considérés et de permettre ainsi d'apporter une solution à une problématique classique dans l'évaluation de la diversité. Il devient alors possible de discrétiser le sous-espace délimité par l'enveloppe convexe et d'appliquer des méthodes basées sur l'utilisation de grilles. Le comportement de ces indices a alors été analysé dans des situations fictives pour permettre de mettre plus facilement en avant leurs avantages et inconvénients respectifs au regard des règles que nous avons établies sur la base de celles de Waldman.

Il en ressort que certains indices classiquement utilisés ne sont pas adaptés pour l'évaluation de la diversité relative de chimiothèques. D'autres indices ont néanmoins été identifiés comme utiles et surtout complémentaires dans l'information qu'ils apportent. Nous avons ainsi sélectionné cinq indices décrivant différents aspects de la diversité et nous avons proposé un cadre générique pour leur utilisation.

2 Visual Characterization and Diversity Quantification of Chemical Libraries : 2. Analysis and Selection of Size-Independent, Subspace-Specific Diversity Indices

Visual Characterization and Diversity Quantification of Chemical Libraries: 2. Analysis and Selection of Size-independent, Subspace-specific Diversity Indices

Lionel Colliandre,^a Vincent Le Guilloux,^a Stephane Bourg,^b Luc Morin-Allory^{a}*

a) Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans-CNRS, UMR 7311 B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, France

b) Fédération de Recherche, Physique et Chimie du Vivant, Université d'Orléans-CNRS; FR 2708, avenue Charles Sadron, 45071 Orléans Cedex 2, France

AUTHOR EMAIL ADDRESS: luc.morin-allory@univ-orleans.fr

RECEIVED DATE

ABSTRACT

High Throughput Screening (HTS) is a standard technique widely used to find hit compounds in drug discovery projects. The high costs associated with such experiments have highlighted the need to carefully design screening libraries in order to avoid wasting resources. Molecular diversity is an established concept that has been used to this end for many years. In this article, a new approach to quantify the molecular diversity of screening libraries is presented. The approach is based on the Delimited Reference Chemical Subspace (DRCS) methodology, a new method that can be used to delimit the densest subspace spanned by a reference library in a reduced 2D continuous space. A total of twenty-two diversity indices were implemented or adapted to this methodology, which is used here to remove outliers and obtain a relevant cell-based partition of the subspace. The behavior of these indices was assessed and compared in various extreme situations, and with respect to a set of theoretical rules that a diversity function should satisfy when libraries of different sizes have to be compared. Some gold standard indices are found inappropriate in such a context, while none of the tested indices behave perfectly in all cases. Five DRCS-based indices accounting for different aspects of diversity were finally selected, and a simple framework is proposed to use them effectively. Various libraries have been profiled with respect to more specific subspaces, which further illustrate the interest of the method.

INTRODUCTION

Molecular diversity is an established concept that is now routinely used to improve the results of a High Throughput Screening (HTS) campaign. The underlying assumption is that a diverse set of compounds will increase the ratio of hits by maximizing the number of chemically distinct molecules to be tested, thereby avoiding redundant tests¹. It is thus necessary to have tools that allow the description and the comparison of different chemical libraries to be performed, and hence facilitate the selection of the most promising one(s)²⁻⁴.

The concept of chemical space provides a way to represent chemical libraries in a fixed referential, making it possible to assess their relative coverage and absolute diversity⁵. In a previous article⁶, a new graphical representation that can be used to represent a particular library in a reduced chemical space was introduced. Delimited Reference Chemical Subspaces are defined by the combination of a Principal Component Analysis (PCA) model (DRCS model), and a subspace delimitation (DRCS contour) intended to encompass a large proportion of compounds. The delimitation is computed on a reduced (2D) space obtained using the PCA model, and is based on the convex-hull calculation. Isolated compounds (outliers) are excluded prior to the creation of this delimitation, which finally represents the most populated (dense) subspace spanned by the reference library.

Although an intuitive visual inspection is a mandatory step to determine the space coverage of a particular library, a quantitative assessment of molecular diversity can provide complementary information. Automated library design, in particular, makes extensive use of diversity indices as target values that need to be optimized. Diversity indices can also be used as a complementary decision support when visual inspection does not permit to choose between two similar libraries.

Many different types of diversity indices have been developed in past decades⁷. Although intuitive in appearance, molecular diversity is actually difficult to quantify precisely, since what we call

"diversity" is obviously strongly dependent on the method used to describe molecules, the metric used to compare these molecules, and the function used to quantify the diversity itself. In terms of applicability domain, molecular diversity indices are typically used to compare and optimize collections of fixed size extracted from a reference library⁸. However, when one needs to analyze and compare two libraries of different sizes, e.g. after an enrichment process, some of the indices are found inappropriate. Previous research has shown that the size of the library usually has a notable influence on the final value of most diversity indices⁹⁻¹¹.

Diversity functions can be classified in 3 broad categories: (1) distance-based methods, which compute the diversity using compounds similarity / dissimilarity measures, (2) cell-based methods which partition a multi-dimensional descriptor space into a finite number of cells used to assess the space coverage, and (3) structure-based methods, which seek to maximize the number of different substructures in the target library (e.g. scaffolds). A decade ago, Waldman et al^{9, 12} proposed in two excellent papers a framework to better describe the expected behavior of a good diversity function. In particular, they defined a diversity function as being any protocol that quantitatively assesses the coverage of a particular descriptor space⁹. Such a definition implies that a working descriptor space must be defined. Furthermore, as working in an open space is inappropriate in most cases, a way to delimit this space is usually required to focus attention on the populated regions. To this end, cell-based methods are probably the simplest, fastest and most appropriate ways of partitioning a particular descriptor space, and have been extensively used for this purpose^{13, 14} to characterize chemical libraries. It is however well recognized that they usually suffer from the presence of sparse regions, which usually leads to an over-consideration of outlier compounds¹⁵. The choice of the binning scheme used to partition the chemical space usually remains empirical, although valuable improvements have been proposed¹¹.

In this work, we will show that the DRCS methodology can be used to tackle the problem of delimiting a reduced descriptor space and to obtain Relative Diversity Indices (RDI) that are

independent of the cardinality of the library. To do so, we suggest that the densest region of a particular subspace, as defined by the DRCS contour, should be focused on to apply cell-based diversity functions, leaving the outliers to be analyzed separately. As the DRCS contour reflects the overall shape of the subspaces spanned by a chosen library, a more relevant and flexible partition of it can be obtained compared to the classical hypercube partitioning. A total of twenty-two diversity indices have been implemented and analyzed in combination with the DRCS methodology. As no index behaves perfectly, five indices that account for different aspects of diversity were finally selected. Various libraries were profiled using different DRCS subspaces obtained using more specific reference libraries (e.g. Lipinski compliant, fragment or pharmaceutical compounds), which further illustrate the interest of the DRCS methodology.

METHOD

Various real and fictive chemical libraries were used to study the behavior of twenty-two diversity indices. All these libraries and the indices will be further described.

Publicly available data sets. A database of 6.63 million unique, standardized and non-reactive compounds⁶ that were gathered from 73 vendor collections was used to create the following subsets:

- 0.12 M compounds satisfying the rule of three as defined by Congreve et al. (molecular weight < 300, number of hydrogen bond donors ≤ 3 , number of hydrogen bond acceptors ≤ 3 and ClogP ≤ 3 ; no violation of these four rules allowed)¹⁶,
- 6.22 M compounds satisfying the rule of five as defined by Lipinski et al.¹⁷,
- 0.41 M compounds not satisfying the rule of five.

MOE®¹⁸ descriptors were used to create these filtered sets. Several commercial or publicly available compound collections were also used in this study:

1. the Prestwick Chemical Library¹⁹ containing only marketed drugs,
2. the Comprehensive Medicinal Chemistry (CMC)²⁰, a database of pharmaceutical compounds,
3. the Chembridge Kinaset²¹ a target-based library containing compounds similar to kinase inhibitors,
4. the Pyxis Discovery Smart Fragment Library²², a fragment library based on scaffolds found in existing drugs,
5. the EPA Fathead Minnow Acute Toxicity (EPAFHM)^{23, 24} a database containing compounds with known toxicity, often used for the development of predictive quantitative structure-activity relationship (QSAR) models of toxicity,
6. a set of 1420 accepted marketed compounds retrieved from the DrugBank database²⁵,

7. two combinatorial libraries used in the article of Owen et al ²⁶ were used: A3 and B5 libraries. The original denomination of these libraries was kept.

A combined library was created by merging the DrugBank, Prestwick and CMC libraries. This combined library is intended to represent “pharmaceutical compounds” in a broad sense. Moreover, random selections were made to create 60 subsets from the CMC database (8773 compounds). The subsets vary from 100 to 8000 compounds (12 sizes of subsets; 5 subsets per size).

Each library was standardized using the protocol described in the first article and the MOE2D descriptors were subsequently computed.

Fictive data sets. A set of thirty-three fictive libraries was created to study the behavior of diversity indices in various extreme situations. Based on the DRCS, a grid was applied to the 2D subspace, strictly encompassing the DRCS contour. The fictive libraries were created by adding points at the center of selected cells of the grid. Each point represents one compound. Only the cells inside the DRCS can be completed i.e. cells with at least one corner inside the DRCS. Except for libraries H, I, J and K, a grid containing around 2500 cells inside the DRCS was used. The exact number of cells depends on the subspace used, which might lead to an optimal grid that contains a slightly different number of cells (see the section Group 3 below). Schematic views of each library can be found in the results section and in the supporting information.

- Library A was created to perfectly occupy the subspace, i.e. one point was added in each cell.
- Libraries B, C and D were created by duplicating library A two, four or eight times respectively.
- Libraries E, F and G: two points per cell were created for half of the cells. The occupied cells are regularly distributed in the delimited space in library E (two points in the first cell, then zero points in the next cell, etc...). In F only the cells in the left part of the subspace are occupied and in G only the cells in the bottom part of the subspace are occupied.

- Libraries H, I, J and K were created to perfectly occupy the subspace i.e. one point was added in each cell. The four libraries differ in the grids used to create the sets of points (from around 5000 to 50000 cells inside the subspace), and thus in the number of compounds.

- Libraries L to Q: points were added in cells of two independent regions of the delimited space leading to libraries of around 2500 compounds. From L to Q, only the distance between the barycenters of the two groups of points was increased.

- Library R: six points were added on a set of 416 contiguous cells.

- Combined libraries: a series of fifteen libraries was designed corresponding to the association of x % of randomly selected points taken from the fictive library R and $(100 - x)$ % of the fictive library A, with x varying from 1 to 99.

Construction of DRCS and specific subspaces. The MOE2D DRCS, as defined in the previous article⁶, was used for all experiments. Briefly, 174 MOE2D descriptors were computed on 30 subsets of 20 000 molecules randomly selected in the database of 6.63 million molecules. These subsets were subsequently used to build a consensus MOE2D PCA model (43.5 % of the variance is explained by the two first principal components). The HTS subspace (as defined by the DRCS contour) was subsequently built upon the same subsets using their projection coordinates in the first two Principal Components. Briefly, a subspace is defined by a 2D convex delimitation that is obtained by averaging a set of convex hulls computed on each subset of molecules. Prior to the calculation of each individual convex hull, a certain proportion of molecules (referred to as outliers) is removed. This proportion is determined by identifying the smallest percentage that can be used to obtain a stable shape – see ref ⁶ for details. The HTS subspace finally encompassed 99.67 % of the compounds.

Using the specific sets of compounds (see above) and the MOE2D PCA model, four new subspaces were derived:

1. the “Lipinski+” and “Lipinski-” subspaces, based on molecules that satisfy (resp. do not satisfy) the Lipinski rule of five,

2. the “Fragment” subspace, based on molecules that satisfy the rule of three,
3. the “Pharmaceutical” subspace, based on the combined library of pharmaceutical compounds,

The detailed methodology used to obtain all these contours can be found in the supporting information.

Generalities on chemical diversity indices: what is a usable diversity index? Based on their definition of a diversity function, Waldman et al ⁹. proposed a set of theoretical requirements that a “perfect” diversity function should satisfy:

1. *Adding redundant molecules to a system does not change its diversity.*
2. *Adding non-redundant molecules always increases the diversity of the system.*
3. *Space-filling behavior of diversity space should be preferred.*
4. *Perfect (i.e. infinite) filling of a finite descriptor space should result in a finite value for the diversity function.*
5. *If the dissimilarity or distance of one molecule to all others is increased, the diversity of the system should increase.*

We completely agree with these requirements for the definition of an *absolute diversity index* characterizing the overall diversity of a library. But, taking only these criteria into account, the most natural way to increase the diversity is to increase the number of compounds. It must be remembered that the general goal of designing diverse screening libraries is to increase the ratio and the interest of the hits. Hence, from the practical point of view of an experimentalist doing screening, the interest of an index observing rules 1 and 2 is quite debatable. For such an application, a Relative Diversity Index (RDI) describing the mean diversity *per molecule* is certainly more valuable and better adapted. The most natural way to get such a value is to divide the value of the diversity of a library (the absolute diversity index) by the cardinality of this library. It will discriminate two libraries with the same overall diversity but different cardinalities. Moreover, this index will be comparable between two libraries

containing a different number of molecules, which is clearly not possible using an absolute diversity index which is expected to increase with the number of compounds, regardless how similar are these compounds. In their second paper, Waldman et al.¹² went into this problem but did not explicitly adapt the previous rules.

For a RDI, rule one must be modified. The addition of redundant molecules, which does not increase the overall diversity but increases the number of compounds, decreases the index (from the experimentalist point of view, adding redundant molecules decreases the overall interest of a chemical library: the diversity will not increase, but the costs will).

Dealing with the overall diversity of a library, rule two is obvious. Any new (*i.e.* not previously present) compound will increase this diversity, no matter how different it is compared to the existing library. But regarding the relative diversity (*i.e.* the mean contribution to the diversity per product) the addition of a non redundant molecule has no obvious effect. The RDI can either increase or decrease depending on the diversity induced by the new product *vs.* the mean diversity of the library. A new compound very different from the previous ones is likely to increase the RDI, but another one, very similar to some products is likely to decrease this RDI. Thus the second rule has no interest for an index such as RDI.

Rules 3, 4 and 5 concern the space coverage and have the same interest for an absolute diversity index as for an RDI. Their application is sometimes rather difficult. One can note that rule 4, "perfect filling..." implies the use of a finite descriptor space, which can be achieved using the DRCS contour. Furthermore, rule 5 also implies that the RDI value must regularly increase when diverse compounds are added to a library.

The use of an RDI implies one other rule. A representative subset of a library should have the same RDI (or a very similar value) as its parent library. There is no unique definition of representativity in chemoinformatics^{27, 28}. In this work, the statistical definition of representativity is used: one set is representative of another if the two have the same statistical distribution of their properties. The easiest way to obtain such a subset is to perform a simple random selection. Differences can appear between a

random sample and the whole library due to the random process; but they decrease when the cardinality of the sample increases. This criterion, which is quite obvious when using basic descriptors, is not always satisfied with the currently used indices, as shown in the results.

Finally a perfect *relative diversity index* should satisfy the following rules:

1. *The RDI of a representative subset of a library is equal (or very similar) to that of the whole library.*
2. *Adding redundant molecules to a system decreases its RDI.*
3. *Space-filling behavior of diversity space should be favored by the RDI.*
4. *Perfect (i.e. infinite) filling of a finite descriptor space should result in a finite value of the RDI.*
5. *If the dissimilarity or distance of one molecule to all others is increased, the RDI of the system should increase.*

In the following part of this paper we will analyze the various indices using these rules.

Indices implemented. Various indices have been previously developed to describe a chemical library, especially to describe and quantify their chemical diversity^{9-11, 13, 14, 29-34}. Twenty-two of these indices were implemented in our study. Our goal is to select the index(ices) satisfying the rules of an RDI and allowing the best description and characterization of the chemical diversity of libraries through the DRCS methodology.

Six DRCS-free indices were implemented for comparison purposes only. The other sixteen indices were applied in combination with the DRCS methodology (in the 2D representation). The indices were classified in four groups:

- Reference indices: widely used indices that do not depend on the DRCS contour.
- Group 1: index characterizing the number of compounds projected outside the DRCS contour.
- Group 2: non-cell based indices applied on DRCS.

- Group 3: cell based indices applied on DRCS.

For all the four groups, each measurement methodology will be further described and compared.

It is important to note that, for comparison purposes, all the indices were computed using the compounds projected inside the DRCS contour.

Reference indices. For these indices, the DRCS contours only serve as a way to remove outliers prior to calculation.

Molecular scaffolds and frameworks: molecular scaffolds and frameworks are simplified representations of chemical structures³⁵. For each chemical library, the ratio between the number of different scaffolds/ frameworks and the number of compounds was computed:

- *Scaffolds:* the Bemis and Murcko³⁵ definition was used, where only rings and linkers between the rings are kept.
- *Frameworks* are based on the scaffolds, but atom types are removed and bond orders are all set to one. However, unlike the original implementation, we differentiate aromatic from non aromatic bonds for six-member rings³.

For the two indices, an in-house InChI^{36, 37} based script was applied to remove duplicate scaffolds and frameworks.

Fingerprint-based indices: molecular fingerprints are binary strings that encode the presence of a set of chemical features in a compound. They were previously used³⁸ for the characterization of chemical libraries. In this study, two fingerprints implemented in Pipeline Pilot³⁹ called ECFP_4#S and EPFP_4#S that take into account the stereochemistry of the compounds were computed. Briefly, ECFP_4#S generates extended-connectivity fingerprints^{38, 40} whereas EPFP_4#S generates Daylight-style path-based fingerprints⁴¹. In these methodologies, for a given chemical library, each bit of the fingerprint represents a chemical feature found in a chemical structure. Finally, the length of the

fingerprints corresponds to the number of different chemical features found in all the compounds of the library.

The following indices were derived for both fingerprints:

- *NumFPFeatures*: number of chemical features present in a database scaled by the number of compounds.
- *AvgDistance*: the average Tanimoto distance for all pairs of molecules.

Group 1: external compounds. The DRCS contour encompasses 99.67% of the HTS compounds used to compute the DRCS model⁶. However, using the same DRCS (model + contour), the percentage of compounds projected outside the contour can vary significantly depending on the library under consideration. These compounds represent “exotic” molecules with respect to the library used to delimit the subspace, i.e. molecules having a combination of molecular properties not found in the reference library. Thus, for each chemical library, the percentage of compounds projected outside the DRCS contour will be computed. It will be further referred to as the *Out(%)* index. These external compounds will not be considered for the calculation of all the other indices.

Group 2: non cell based indices applied on DRCS. For these indices compounds will be characterized by their coordinates in the DRCS model.

Euclidean distance-based functions: The Euclidian distances in the reduced 2D space were used to derive the following indices:

1. *MeanMinEucDist*: the average Euclidean distance between each molecule and its closest neighbor.
2. *MeanEucDist*: the average Euclidean distance for all pairs of molecules in the library.

These indices require the calculation of all intermolecular distances, which is computationally very expensive (complexity in $O(N^2)$).

Diversity integral based indices: other distance-based functions exist which are less time-consuming. In particular the Diversity Integral methodology of Cerius2 C²-Lib⁴², used by Pascual et al. and others^{11, 31} is comparable to the calculation of the *MeanMinEucDist* index. It is based on the average Euclidean distance between random points and their closest compound in the library under consideration.

Similarly, in this work, a set of fixed reference points was used instead of random points to define the Diversity Integral index (*DivInt*). This implies that the complexity of the methodology drops to $O(N)$. For each subspace, a set of 1000 reference points was equally dispersed to cover the entire subspace (see the supporting information) and, for each reference point, the minimum distance to a product of the considered library is computed. The average of these distances is calculated to obtain the *DivInt* index. Hence, the lower the *DivInt* is, the better the distribution of the projected compounds inside the DRCS is.

Group 3: cell based indices applied on DRCS. As mentioned previously, cell-based partitioning is an intuitive way of assessing the space coverage of a particular library. Because chemical space is almost infinite, the space is usually partitioned into a hypercube binned on each descriptor, and the diversity is usually expressed using the proportion of occupied cells. Pascual *et al*^{11, 31} divided the ranges of principal components in such a way that the number of occupied cells is always less than or equal to the number of molecules to select. This leads to a large number of empty cells at the extreme part of the space, which are irrelevant to consider. To avoid the use of parts of the chemical space that are not informative, it is necessary to focus on what Agrafiotis¹⁵ called the “Accessible space” determined by focusing on a representative subspace which minimizes the influence of outer regions. Cummins *et al.*²⁹ proposed to remove the compounds present in the low density space (e.g. cells that are

occupied by few compounds) in order to optimize the size of the considered space and of the grid. This allows one to focus on the most representative subspace.

The DRCS contour provides an accurate delimitation of the densest part of a particular delimited subspace. Consequently, a partitioning of this subspace is expected to be more relevant than using a traditional hyper cubic or hyper spherical delimitation. A specific partitioning was therefore applied to each subspace delimited by its corresponding DRCS contour, as described in the following section.

DRCS-based partitioning. To assess the coverage of a library for a particular partitioned subspace, the following assumption was made: “if the chemical library covers the subspace defined by the DRCS contour in an optimal manner, the proportion of occupied cells must be 100 %”. In other words, the partitioning should be defined in such a way that each cell will be filled by one and only one compound in the case of an ideal library. Using a unique grid to compare libraries of different sizes is thus impossible. This leads to one important practical consequence in the present case: the number of cells falling inside the DRCS must be equal to the number of compounds, leading to a different grid for each library of different cardinality.

The first step is therefore to create a 2D grid that strictly encompasses the entire subspace. In the case of a typical squared-shaped subspace, the number of bins N_{bins} for each dimension of the grid (which is assumed to be the same) would be determined as:

$$N_{bins} = \sqrt{N_{mol}}$$

where N_{mol} is the number of molecules in the library. In our case however, the shape of each contour is not straight and regular. Thus, encompassing the entire subspace using such a $N_{bins} \times N_{bins}$ grid will inevitably lead to cells falling outside it (Figure 1), and the final number of cells located inside the contour will be clearly different from the number of molecules. This suggests that a relationship might exist between the number of bins N_{bins} and the final number of cells located inside a particular DRCS contour $N_{cells-in}$. For the HTS contour, a set of grids was computed with N_{bins} varying from 10 to 2500.

For each grid, the actual number of cells falling inside the contour was computed (a cell is considered as being inside the contour if at least one of its corners is inside the contour). Various regression-based relationships between N_{bins} and $N_{cells-in}$ were tested. Interestingly, we found that a power regression is able to establish a strong correlation for all the cases (i.e. for all the various subspaces considered in the previous and in the present work). Following this analysis, a power relationship was established and validated on all subspaces to determine the optimal number of bins based on the number of molecules. Finally, given a subspace S and a library containing N_{mol} molecules, the number of bins in each dimension of the grid is determined as:

$$N_{bins} = \alpha_S \times N_{mol}^{\beta_S}$$

where α_S and β_S are the two parameters determined by the power regression obtained on the subspace S . The coefficient values for each subspace presented herein can be found in the supporting information. Consequently, each grid is specific both to the subspace under consideration and to the library to be analyzed. Once N_{bins} has been determined, the grid is positioned to strictly encompass subspace S , and cells located outside the subspace (i.e. cells that have all their corners outside the contour) are subsequently removed. All the indices described below were applied using this optimized grid.

Filling(%). The simplest index that can be defined based on a cell-based partitioning is the percentage of occupied cells. Given N_{occ} occupied cells among N_{total} cells in an optimized grid, the filling percentage (*Filling(%)*) is defined as:

$$Filling(\%) = \frac{N_{occ}}{N_{total}} * 100$$

It quantifies the overall coverage of a particular subspace, but does not take into account the evenness of this distribution.

Shannon entropy: Shannon entropy (SE) was originally developed for application in digital communication theory⁴³. It was transferred to the chemical domain and applied to measure the information of molecular descriptors distribution in compound database⁴⁴⁻⁴⁶.

Shannon entropy was also successfully applied to quantify the distribution uniformity of compounds in cells when chemical spaces are partitioned by a grid^{13, 31, 33}. It is defined as:

$$SE = - \sum_{i=1}^{N_{full\ cells}} p_i \log_2 p_i$$

where $p_i = N_i / N_{cpds}$ with N_i being the number of compounds in cell i and N_{cpds} the number of compounds in the DRCS.

To compare this SE value between libraries of different sizes (hence different numbers of cells inside the DRCS contour), it has to be scaled. The meaning of the scaled value depends on the reference used for the scaling. If one uses, as Godden & Bajorath⁴⁶ did, the total number of cells as reference (total number of bins), one obtains:

$$sSE_{all\ cells} = SE / \log_2(N_{cells})$$

with N_{cells} being the number of cells inside the DRCS contour. Thus, $sSE_{all\ cells}$ captures the uniformity of the distribution of the compounds through the entire grid (i.e. all the chemical space defined by the DRCS).

The latter index depends on the percentage of occupied cells (i.e. *the Filling(%)*), thus Shannon entropy can be scaled using the number of occupied cells instead of the total number of cells:

$$sSE_{occ\ cells} = SE / \log_2(N_{occ\ cells})$$

with $N_{occ\ cells}$ being the number of occupied cells inside the DRCS contour. The $sSE_{occ\ cells}$ captures the uniformity of the distribution of the compounds only through the occupied cells, i.e. the occupied

chemical space defined by the DRCS. These two sSE values vary from zero to one (maximum uniformity of the distribution).

Shanmugasundaram and Maggiora³³ pointed out that the cell-based Shannon entropy treats cells as “positionally independent”. To avoid the loss of this information they thus introduced a new Shannon-like index to measure the uniformity of the distribution of occupied cells along each dimension of the grid. The application of this index to our DRCS gives:

$$SE_{PCx} = - \sum_{i=0}^{Nbins} p_i \log_2 p_i$$

with PCx being the principal component X of the DRCS space and $p_i = N_{occ\ cells\ i} / N_{occ\ cells}$ where $N_{occ\ cells}$ is the number of occupied cells inside the DRCS and $N_{occ\ cells\ i}$ is the number of occupied cells in the i^{th} bin. Section 6 of supporting information illustrates the overall process.

Shanmugasundaram and Maggiora do not scale this entropy value using the number of intervals in each dimension, but average the values obtained for all the dimensions. For the application of this index to our DRCS, scaling is mandatory due to the use of grids of variable sizes. But in our grid scaling is not trivial because the bins are not all equivalent, i.e. the bins do not all contain the same number of cells. To avoid this problem, the proportion of occupied cells for each bin was used instead of the number of occupied cells for each bin. We thus obtain a new SE_{PCx} value where $p_i = p_{celli} / \sum p_{celli}$ and $p_{celli} = N_{occ\ cells\ i} / N_{cells\ i}$. This value can be scaled:

$$sSE_{PCx} = SE_{PCx} / \log_2(\text{total number of intervals of the grid})$$

The sSE_{PCx} values were kept for the first two principal components of the DRCS. An additional index $sSEPCmean$ was also defined as the average value of sSE_{PC1} and sSE_{PC2} .

Cluster diversity index: data clustering is broadly used to assess the quality/diversity of data sets⁴⁷. The CLIQUE⁴⁸ (CLustering In QUEst) algorithm was used to find subspaces of a partitioned space with high

density clusters. A simplified version of this algorithm was used to evaluate the number of clusters in our DRCS, where the clusters are based on all the occupied cells and not only on the most occupied ones (the “densest cells”). A cluster is defined as a set of contiguous occupied cells (two cells are contiguous if they have a common face, see Figure 1). Only the number of clusters is kept.

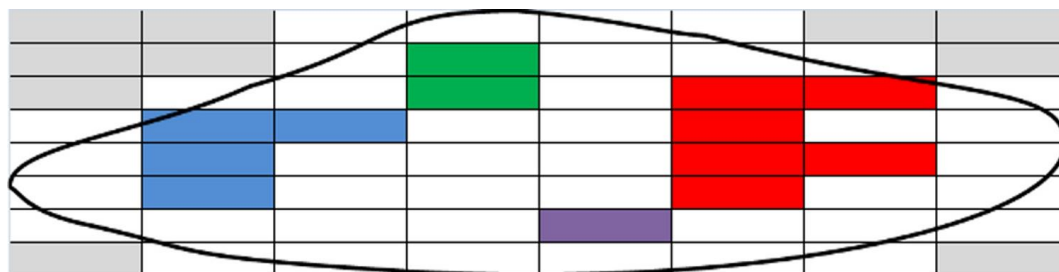


Figure 1: Simulation of the projection in a DRCS of a library having four clusters of occupied cells. The cells considered to be outside the DRCS are in gray, the DRCS contour is in black, the four clusters of occupied cells are colored differently.

The homogeneity of the projection can be characterized by the number of clusters scaled by the number of occupied cells, leading to the Cluster Diversity Index (*ClusterDiv*):

$$ClusterDiv = \frac{\text{number of clusters}}{\text{number of occupied cells}}$$

Kolmogorov-Smirnov index: the Kolmogorov-Smirnov (*KS*) criterion was first applied by Rassokhin & Agrafiotis^{10, 49} to quantify molecular diversity. The *KS* criterion measures how well an experimental distribution is approximated by a particular distribution function⁵⁰. It is defined as the maximum value of the absolute difference between two cumulative functions:

$$KS = \max_{-\infty < x < +\infty} |P(x) - P^*(x)|$$

where $P(x)$ is a known cumulative distribution of a uniform sample and $P^*(x)$ is the experimental cumulative distribution. Thus, the *KS* criterion measures the extent to which the real distribution deviates from the theoretical one. In our study, the optimal (known) distribution is a uniform distribution, where all molecules are evenly distributed over all the cells of the grid. Since the *KS* criterion is a measure of dissimilarity, $KS = 0$ is obtained when the experimental distribution is optimal. As we want to obtain a maximum value when the distribution is optimal, a *KS*-based diversity index called *Dks* was defined as:

$$Dks = 1 - \max_{-\infty < x < +\infty} |P(x) - P^*(x)|$$

The Manhattan distances between cells of the optimal grid were used to compute the $Dks(ManhDist)$ index. For each possible value of Manhattan distance (between zero and twice the number of bins), the ratio between the number of distances and the total number of distances is considered. Then, the theoretical cumulative distribution is based on the Manhattan distances for all the cells ($P^*(R_{ManhDist})$) and the experimental cumulative distribution on the Manhattan distances for the occupied cells ($P(R_{ManhDist})$). The comparison of the two cumulative distributions leads to the $Dks(ManhDist)$ index:

$$Dks(ManhDist) = 1 - \max_{0 < i < 2N_{bins}} |P(R_i(ManhDist)) - P^*(R_i(ManhDist))|$$

Since the $Dks(ManhDist)$ implies the calculation of the Manhattan distances for all the occupied cells inside the DRCS contours, it is computationally expensive. Two new Dks indices were created for each principal component of the DRCS model ($Dks(PC1)$ and $Dks(PC2)$). Based on the optimized grid, the ratio between the number of occupied cells and the total number of occupied cells is considered from the first to the last bin of the grid for each PC. Then the experimental ($P(R_{occ\ cells})$) and theoretical ($P^*(R_{occ\ cells})$) cumulative distributions are based on the real and expected distributions of the occupied cells through all the bins (the shape of the DRCS contour is taken into account). The comparison of these two cumulative distributions thus leads to the $Dks(PCx)$ indices:

$$Dks(PCx) = 1 - \max_{0 < i < N_{bins}} |P(R_i(occ\ cells)) - P^*(R_i(occ\ cells))|$$

These $Dks(PCx)$ indices computed on one dimension are faster to compute than the $Dks(ManhDist)$. Section 7 of supporting information illustrates the overall process.

Manhattan distance-based indices: similarly to non cell-based indices, distance-based functions can be applied on the occupied cells. The following indices were thus computed:

1. *MeanMinManhDist*: the average Manhattan distance between each occupied cell and its closest neighbor.

2. *MeanManhDist*: the average Manhattan distance for all pairs of occupied cells.

Final procedure to compute DRCS based indices. From a pre-computed DRCS (model + contour) based on a reference library, the following final procedure was used to compute DRCS based indices of a new library:

- 1- standardize the molecules of the studied library,
- 2- compute the descriptors corresponding to the DRCS,
- 3- compute the compounds' coordinates in the DRCS model,
- 4- test whether the molecules are inside or outside the DRCS contour,
- 5- compute the optimal number of grid bins,
- 6- test whether the grid cells are inside or outside the DRCS contour,
- 7- for compounds and grid cells inside the DRCS contour, attribute each compound to a grid cell,
- 8- compute the DRCS based diversity indices.

RESULTS & DISCUSSIONS

Twenty-two indices were implemented to study the diversity of chemical libraries (see Method section). The behavior of these indices was further explored with respect to the rules that a relative diversity index (RDI) should satisfy (see Method section). The objective of this work is to select a few indices based on the DRCS methodology that respect the maximum number of these rules and provide non-redundant information on diversity.

Representativity of diversity indices. Rule 1 implies that the RDI values of representative subsets of a library must be equal to the RDI value of the entire library. This property was verified by computing and studying the variation of the indices for the 60 random subsets extracted from the CMC library. The CMC database was used in order to study the implemented indices on a real library. Furthermore, the CMC database has one of the highest coverages of chemical space that we have found ⁶.

In order to obtain references, two simple indices were studied: the average molecular weight and the *MeanMinEucDist*. According to the statistical definition of representativity, the average molecular weight must be constant regardless of the cardinality of the subset. On the contrary, the average of the distances of each compound from its closest neighbor (i.e. the *MeanMinEucDist* index) should obviously decrease when the number of compounds increases. The observations derived from Figure 2 are consistent with these predictions. For each size of subset, the average value of the five subsets was plotted with the error bars limited to one standard deviation. Obviously the error bars for the small sizes of subset are the largest. The average molecular weight remains quite constant (within the error bars) whereas the average *MeanMinEucDist* shows strong variations. The first index thus satisfies the first rule of an RDI but the second one is unusable.

Globally all the indices show either of the two variations observed previously. For some of them it is difficult to analyze these variations for the small subsets (under 1000 compounds) since the random variations are too great, but for the subsets between 1000 and 8000 the results are quite clear. All the

graphs are presented in the supporting information Table S2 and the results are summarized in the column "Stability" in Table 1.

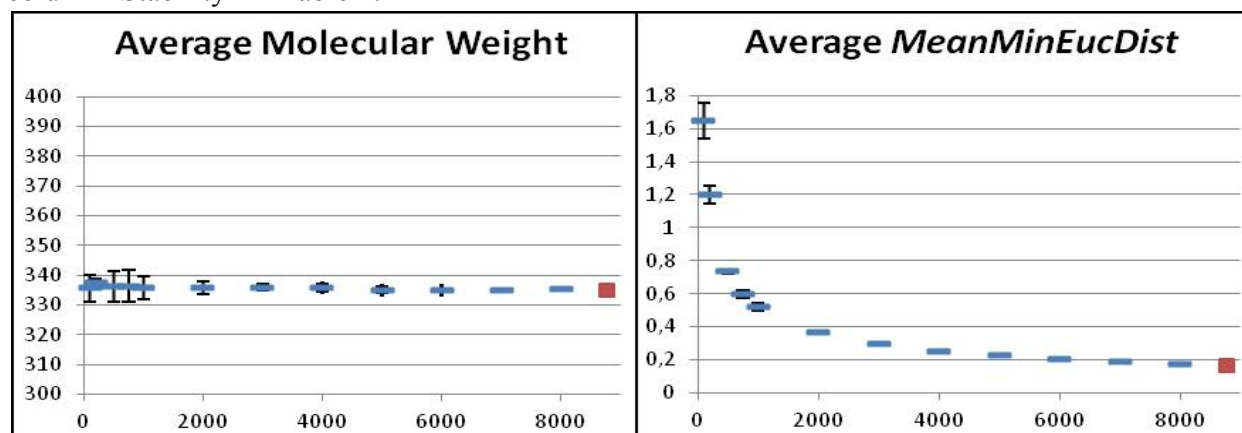


Figure 2: average molecular weight and average *MeanMinEucDist* for five subsets randomly selected from the CMC database. The size of the subsets varies from 100 to 8000 compounds. The error bars correspond to one standard deviation. The red point indicates the indices values for the whole CMC database.

Table 1: summary of the results for rule 1 of an RDI. +: stable index; -: size-dependent index.

Group	Description of the indices	Name of the indices	Rule 1 of an RDI : stability
REF	Percentage of scaffolds	<i>Scaffolds</i>	-
	Percentage of Frameworks	<i>Frameworks</i>	-
	Percentage of chemical features in molecular fingerprints ECFP_4#S	<i>NumFPFeatures (ECFP_4#S)</i>	-
	Percentage of chemical features in molecular fingerprints EPFP_4#S	<i>NumFPFeatures (EPFP_4#S)</i>	-
	Average Tanimoto distance for ECFP_4#S molecular fingerprints	<i>AvgDistance (ECFP_4#S)</i>	+
	Average Tanimoto distance for EPFP_4#S molecular fingerprints	<i>AvgDistance (EPFP_4#S)</i>	+
1	Percentage of compounds outside the DRCS contour	<i>Out(%)</i>	+
2	Mean of minimal Euclidean distances	<i>MeanMinEucDist</i>	-
	Mean of Euclidean distances	<i>MeanEucDist</i>	+
	Diversity integral	<i>DivInt</i>	-
3	Filling percentage for optimized grid	<i>Filling(%)</i>	+
	Scaled Shannon entropy with reference to the total of cells	<i>sSE_{all cells}</i>	-
	Scaled Shannon entropy with reference to the occupied cells	<i>sSE_{occ cells}</i>	-
	Scaled Shannon entropy for PC1	<i>sSE_{PC1}</i>	-
	Scaled Shannon entropy for PC2	<i>sSE_{PC2}</i>	-
	Mean of Shannon entropy for all PC	<i>sSE_{PCmean}</i>	-
	Scaled number of clusters for the occupied cells	<i>ClusterDiv</i>	+
	Kolmogorov-Smirnov diversity based on Manhattan interdistances of occupied cells	<i>Dks(ManhDist)</i>	+
	Kolmogorov-Smirnov diversity based on the occupied cells distribution on the PC1	<i>Dks(PC1)</i>	+
	Kolmogorov-Smirnov diversity based on the occupied cells distribution on the PC2	<i>Dks(PC2)</i>	+
	Mean of minimal Manhattan distances	<i>MeanMinManhDist</i>	+
	Mean of Manhattan distances	<i>MeanManhDist</i>	-

Representativity of non DRCS-based indices. Among the indices of the reference group, *Scaffolds* and *Frameworks* depend on the cardinality of the subsets. For the fingerprint-based indices, *NumFPFeature* also depends on the size of the subsets, but not *AvgDistance*. The same holds for both types of fingerprint. *AvgDistance* is hence the only representative index in the reference group.

Representativity of DRCS-based indices. The percentage of compounds projected outside the contour (*Out(%)*) is obviously stable for all the subsets. For the non cell-based indices applied to the DRCS (group 2), *MeanMinEucDist* and *DivInt* vary with the cardinality of the subsets whereas *MeanEucDist* remains constant. For cell-based indices (group 3), the Shannon entropy based indices and *MeanManhDist* are found to be unstable, while all the other cell-based indices are stable (see supporting information Table S2).

Finally, among the twenty-two indices implemented, only eleven (50%) are independent of the cardinality of the subsets and hence satisfy the first rule. The same conclusions were obtained for larger datasets (N between 20 000 and 100 000) extracted from our in-house database (data not shown).

These results show that many diversity indices, including some DRCS-based indices, are not appropriate for comparing libraries of different cardinalities. In the remainder of this section, only the DRCS-based indices that satisfy rule 1 of an RDI will be considered.

Behavior of DRCS-based indices. To examine rules 2 to 5, N fictive library projections were used to represent various extreme cases (see Method section). As these libraries were created inside the DRCS contour, the *Out(%)* index is systematically 0, and only the 7 remaining stable indices were analyzed. The results are summarized in Table 2.

Table 2: summary of the results for the stable DRCS based indices according to rules 2 to 5 of an RDI. +: index satisfying the rule; -: index not satisfying the rule; ±: index satisfying the rule with some applicability limits.

		Rules of an RDI
--	--	-----------------

Group	Name of the indices	2:	3:	4:	5:	
		redundancy	space-filling	perfect filling	dissimilarity	monotony
1	<i>Out</i> (%)	Not appropriate for describing the compounds inside the DRCS				
2	<i>MeanEucDist</i>	-	+	+	+	+
3	<i>Filling</i> (%)	+	-	+	-	+
	<i>ClusterDiv</i>	-	+	+	-	-
	<i>Dks</i> (<i>ManhDist</i>)	-	+	+	-	±
	<i>Dks</i> (<i>PC1</i>)	-	+	+	-	±
	<i>Dks</i> (<i>PC2</i>)	-	+	+	-	±
	<i>MeanMinManhDist</i>	+	+	+	-	-

Redundancy: fictive library A has an optimum coverage of the DRCS. Based on it, libraries B, C and D were assembled by duplicating library A two, four or eight times. Thus in libraries B, C and D each product is present in two, four or eight copies.

A good index should discriminate between these four fictive libraries, ranking them in the ABCD order. As shown in Table 3, the *Filling*(%) value decreases with the redundancy (the small differences obtained compared to the expected values [100, 50, 25 and 12.5 for A, B, C and D respectively] stem from the rounding of N_{bins} induced by the power regression). The *MeanEucDist* and *Dks* indices do not seem to change. *ClusterDiv* and *MeanMinManhDist* values increase, while *ClusterDiv* seems to reach a plateau for libraries C and D. This shows that *MeanEucDist*, *ClusterDiv*, and the *Dks* indices do not satisfy rule 2.

Table 3: values of eight stable indices for the fictive libraries A, B and C.

Fictive library	<i>MeanEucDist</i>	<i>Filling</i> (%)	<i>ClusterDiv</i>	<i>Dks</i> (<i>ManhDist</i>)	<i>Dks</i> (<i>PC1</i>)	<i>Dks</i> (<i>PC2</i>)	<i>MeanMinManhDist</i>
A	20.84	100.00	0.04	1.00	1.00	1.00	1.00
B	20.84	49.73	18.03	1.00	0.99	0.99	1.01
C	20.84	24.89	100.00	0.99	0.99	0.99	2.00
D	20.83	12.51	100.00	0.99	0.99	0.99	2.49

Space-filling: in fictive libraries E, F and G 50% of the cells are occupied. In library E the occupied cells are regularly distributed on the grid. In library F and G, the cells are filled in the left and bottom part of the subspace, respectively (Figure 3) .

With the same number of occupied cells, fictive library E is obviously better than F or G. The indices that differentiate these situations will thus satisfy rule 3 of an RDI. Results are given in Table 4.

For all three fictive libraries, the *Filling(%)* is constant (the small difference for library G comes from an artifact of the construction of the fictive library as explained above). *ClusterDiv*, *Dks(ManhDist)* and *MeanMinManhDist* have equal values for F and G but a different one for E. The combination of the two *Dks(PC)* indices makes it possible to differentiate the three libraries. Indeed, for these indices we have the maximum value ($Dks(PC) = 1$) when the points are equally projected along each axis and the middle value ($Dks(PC) = 0.5$) when the points are projected only on half of the axis. Finally, only *MeanEucDist* gives three different values for the three libraries.

Rule 3 of an RDI implies that the size of the empty regions of a chemical space must be minimized. With the same coverage, libraries F and G have one empty region representing 50% of the space whereas library E has the same proportion of empty regions but spread out through all the delimited space. Thus because they differentiate library E from the other two, all the indices except *Filling(%)* satisfy rule 3.

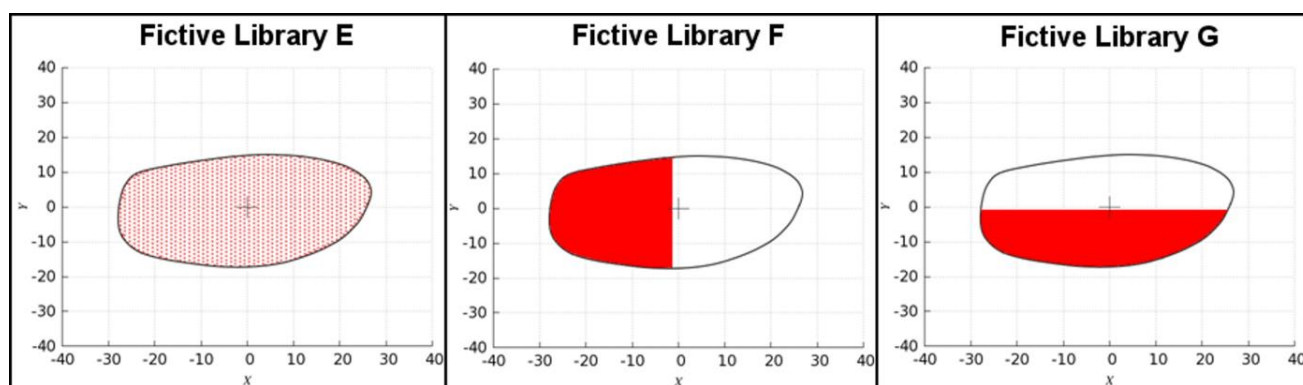


Figure 3: projections of the fictive libraries E, F and G inside the DRCS.

Table 4: values of seven stable indices for the fictive libraries E, F and G.

Fictive library	Mean EucDist	Filling (%)	ClusterDiv	Dks (ManhDist)	Dks (PC1)	Dks (PC2)	MeanMin ManhDist
E	20.86	50.16	100.00	0.99	1.00	1.00	2.00
F	14.22	50.16	0.08	0.79	0.50	0.97	1.00
G	17.52	49.09	0.08	0.78	0.96	0.49	1.00

Perfect-filling: based on a given number of compounds, fictive libraries A, H and I were created to simulate perfect filling of the subspace. This perfect filling is defined as the completion of all the cells

of the optimal grid in the finite and delimited space. All the indices have finite values for these libraries (Table 5), and hence satisfy rule 4 of an RDI. It can be noted that some of the indices are constant (*Filling(%)*, *Dks* and *MeanMinManhDist*). Among the others, *ClusterDiv* tends to zero because in all the cases there is only one cluster but each time the number of compounds increases. *MeanEucDist* also decreases but this does not question the stability (rule 1) of this index.

Table 5: values of seven stable indices for the fictive libraries A, H, I, J and K. The number of fictive compounds in the library is indicated.

Fictive library	Number of fictive compounds	Mean EucDist	Filling (%)	ClusterDiv	Dks (ManhDist)	Dks (PC1)	Dks (PC2)	MeanMin ManhDist
A	2540	20.84	100.00	0.04	1.00	1.00	1.00	1.00
H	4930	20.75	100.00	0.02	1.00	1.00	1.00	1.00
I	9982	20.68	100.00	0.01	1.00	1.00	1.00	1.00
J	19956	20.62	100.00	0.00	1.00	1.00	1.00	1.00
K	49779	20.57	100.00	0.00	1.00	1.00	1.00	1.00

Dissimilarity: Rule 5 of an RDI specifies that when the distance between the compounds (their dissimilarity) increases, the diversity must increase. Six fictive libraries with the same coverage of the chemical space were created to study this rule (Figure 4). For each library, the compounds were separated into two groups covering two small regions of the delimited space. From L to Q, the distance between the barycenters of the two groups increases, thus according to rule 5 the diversity of the libraries increases.

Filling(%), *ClusterDiv*, *Dks(PC2)* and *MeanMinManhDist* are constant for the six libraries (Table 6). This means that they are not able to characterize alone the difference of diversity. Note that *Dks(PC2)* is constant because the projection of the compounds on the y-axis of the DRCS does not change from L to Q. *Dks(PC1)* and *Dks(ManhDist)* increase and rapidly reach a plateau. This limit is clearly due to the very specific examples chosen, but this demonstrates that the ability of the *Dks* indices to measure the homogeneity of the compounds projection inside the DRCS is limited. Finally *MeanEucDist* increases from L to Q. It is the only index that strictly satisfies rule 5 of an RDI.

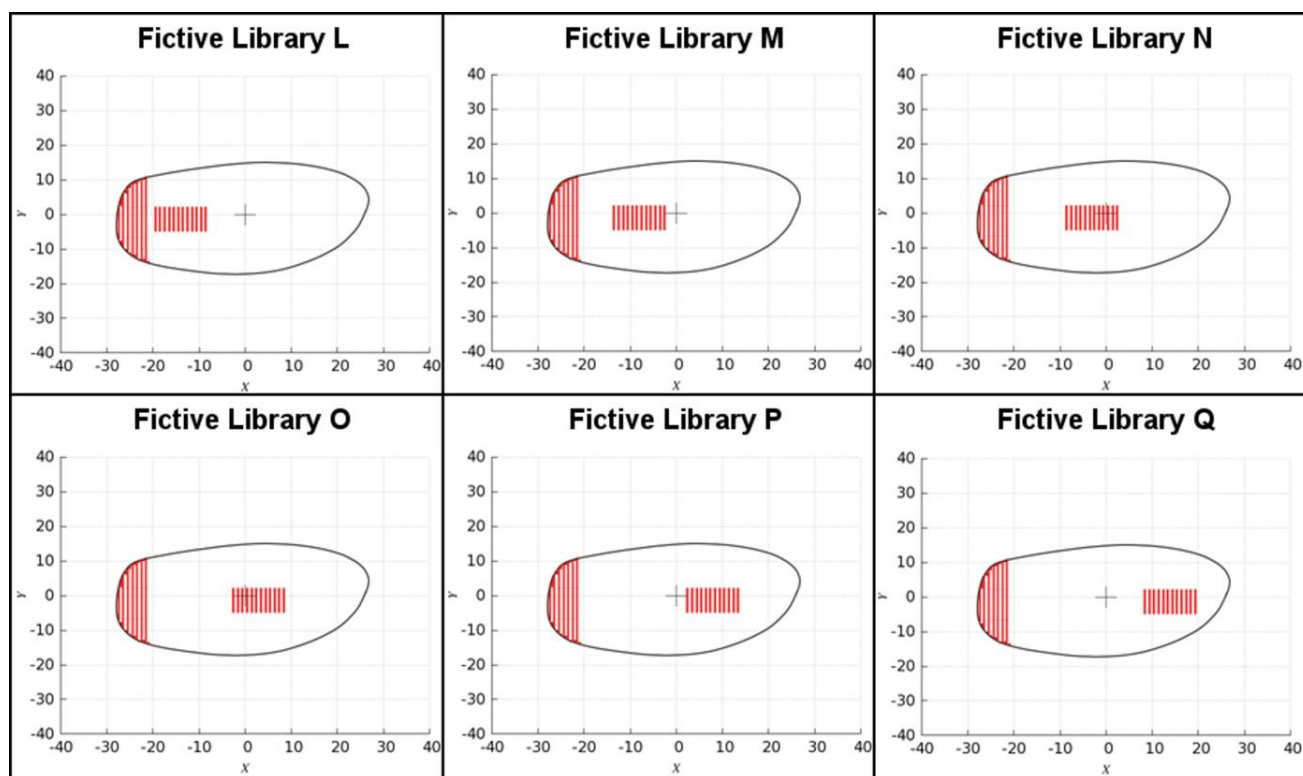


Figure 4: projections of the fictive libraries L to Q inside the DRCS.

Table 6: values of seven stable indices for the fictive libraries L to Q.

Fictive library	Mean EucDist	Filling (%)	ClusterDiv	Dks (ManhDist)	Dks (PC1)	Dks (PC2)	MeanMin ManhDist
L	9.37	15.71	0.50	0.51	0.35	0.81	1.00
M	11.99	15.71	0.50	0.61	0.46	0.81	1.00
N	14.32	15.71	0.50	0.71	0.46	0.81	1.00
O	17.19	15.71	0.50	0.79	0.46	0.81	1.00
P	19.62	15.71	0.50	0.79	0.46	0.81	1.00
Q	22.55	15.71	0.50	0.79	0.46	0.81	1.00

The previous fictive libraries provide some indication about the behavior of the computed indices for such extreme cases, but no indication is available on their evolution between these cases. One consequence of rule 5 of an RDI is that the evolution of the indices must be monotone i.e. the indices values must increase regularly from the least to the most diverse case.

A series of seventeen fictive libraries was created. It corresponds to the evolution from the least diverse library R to the most diverse case A (Figure 5). Results of the indices are given in Table 7. Figure 6 shows the evolution of the seven stable indices values versus the percentage of compounds of the fictive library A in the combined libraries.

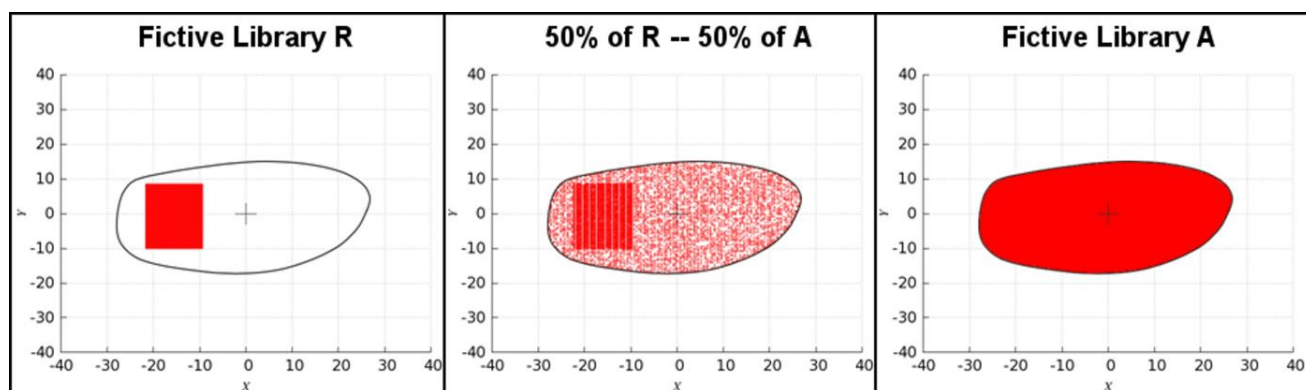


Figure 5: projections of the fictive libraries R and A and of one combined library comprising 50% of compounds from the fictive libraries R and A.

Table 7: values of seven stable indices for the fictive libraries R, A and the combined libraries from R to A.

Fictive library	Mean EucDist	Filling (%)	ClusterDiv	Dks (ManhDist)	Dks (PC1)	Dks (PC2)	MeanMin ManhDist
R (100-0)	8.34	16.38	0.24	0.48	0.33	0.83	1.00
99-1	8.56	17.50	4.58	0.53	0.36	0.84	1.22
95-5	9.53	20.87	17.67	0.70	0.48	0.88	1.37
90-10	10.55	24.95	26.67	0.80	0.59	0.91	1.36
80-20	12.70	33.31	30.20	0.91	0.73	0.93	1.25
70-30	14.37	41.40	26.00	0.95	0.81	0.95	1.16
60-40	16.10	50.15	17.44	0.97	0.87	0.96	1.09
55-45	16.73	54.13	13.62	0.97	0.89	0.97	1.06
50-50	17.32	58.01	10.91	0.98	0.90	0.98	1.05
45-55	17.98	62.38	7.60	0.99	0.92	0.98	1.04
40-60	18.55	66.21	4.37	0.99	0.93	0.99	1.02
30-70	19.39	74.50	1.29	0.99	0.96	0.99	1.01
20-80	20.06	82.36	0.23	1.00	0.98	1.00	1.00
10-90	20.44	90.86	0.04	1.00	0.99	1.00	1.00
5-95	20.63	95.21	0.01	1.00	1.00	1.00	1.00
1-99	20.67	99.03	0.01	1.00	1.00	1.00	1.00
A (0-100)	20.84	100.00	0.04	1.00	1.00	1.00	1.00

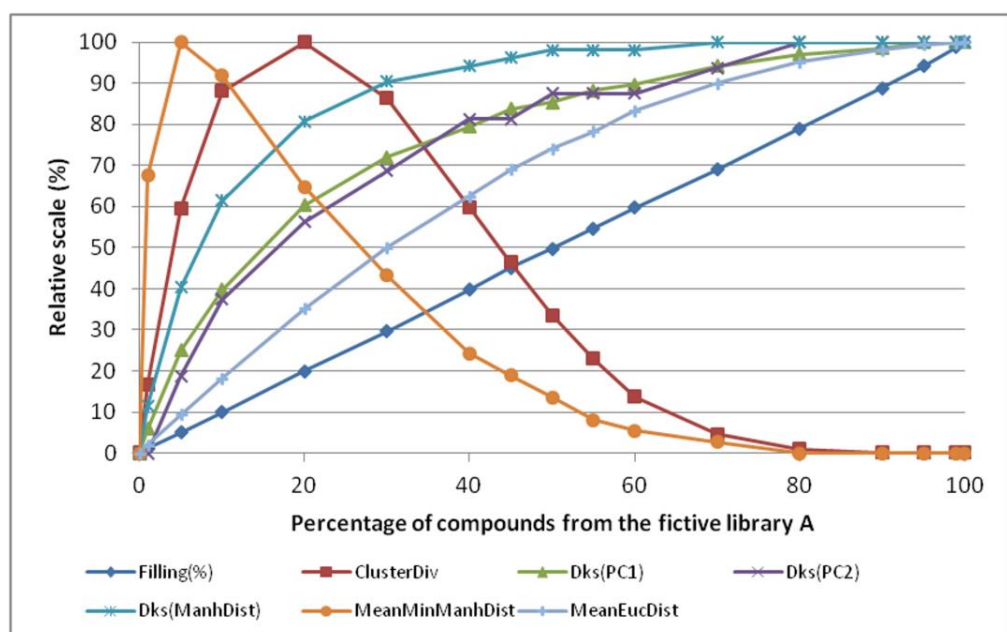


Figure 6 : evolution of seven stable indices computed on combined fictive libraries between R and A (the indices values are indicated on a relative scale between 0 and 100%).

Clearly, *Filling(%)* has a linear evolution from the fictive libraries R to A. It well characterizes the linear increase in the chemical space coverage by the fictive combined libraries.

MeanEucDist and all the *Dks* indices have a non linear but monotone evolution. The combined libraries containing more than 40-60% of the library A (*Filling(%)* > 50-60%) have similar values of the *Dks* indices. These indices are thus difficult to use for comparing chemical libraries having *Filling(%)* value higher than 50%. This is not a problem, however, because this is not the case for the majority of real chemical libraries.

Finally, *ClusterDiv* and *MeanMinManhDist* indices form bell curves. Their values increase up to a maximum and then decrease with the diversity. These indices thus do not have a monotone behavior. Furthermore, the possibility of having the same value for two differently diverse projections is clearly not acceptable.

So, *Filling(%)*, *MeanEucDist* and *Dks* indices satisfy the consequence of rule 5 concerning their evolution (see column “Monotony” Table 2). This result shows the importance of testing the monotony of the evolution of the indices values, and whether this evolution is similar to that of the diversity.

Selection of the DRCS based diversity indices. Sixteen DRCS based indices were implemented and tested against the rules that an RDI should satisfy. Table 1 and Table 2 summarize the results. None of the implemented indices satisfy all the rules. Some indices do not satisfy some rules in a critical way e.g. the non stable and non monotone indices. Other indices are characteristic of some rules and do not satisfy other rules without altering their interpretability e.g. the *Filling(%)* characterizes the coverage of the delimited subspace. Thus, depending on their characteristics, a selection of indices can be made to produce a simple description of the chemical diversity of a library with regard to the DRCS used. In our opinion, these indices must be used in the following order if they are to be interpreted clearly:

1. *Out(%)*: calculation of the proportion of external compounds with regard to the DRCS used.

A chemical library has to be diverse in a given subspace and it should also contain external compounds to explore other subspaces. The calculation of this proportion of external

compounds is mandatory as it enables one to know the proportion of compounds on which the other indices are computed. That is why the more compounds projected outside the DRCS contour there are, the less usable the computed indices are. No strict rule can be implemented, but we consider that the indices are representative and interpretable if the library contains less than 10% of compounds outside the DRCS contour used. If *Out(%)* is higher than 10%, the DRCS used is maybe ill-adapted. Further, the other selected indices will not be computed in this situation.

2. *Filling(%)*: coverage of the studied delimited subspace. This second index gives the proportion of the subspace that the library explores. The higher the *Filling(%)*, the higher the diversity is.
3. *Dks*: homogeneity of the distribution of the occupied cells inside the delimited subspace. For similar *Filling(%)* values, two libraries can have very different distributions of the occupied cells. The *Dks* indices enable one to know whether the occupied cells are grouped together or not. *Dks(ManhDist)* index gives information about the homogeneity of the distribution of the occupied cells in all the DRCS. *Dks(PCI and 2)* indices are faster to compute and provide the same information but through each axis of the DRCS model. They thus offer a greater description of the diversity. That is why they will be further used to characterize libraries. Nevertheless, *Dks(ManhDist)* could be preferred for an automatic application to optimize the chemical diversity of libraries.
4. *MeanEucDist*: homogeneity of the distribution of the compounds inside the delimited subspace. Similar *Dks* values can mask a non-homogeneous distribution of the compounds inside the DRCS. This information can be obtained with the *MeanEucDist* index. However it depends on the calculation of the Euclidean distances for all the pairs of compounds, which implies a high computational cost. Thus this index will be used if the previous indices are unable to differentiate the compared libraries.

The previous selected diversity indices were implemented in radar graphs, with which various chemical libraries can be easily compared with regard to the computed indices.

Specific DRCS contours. DRCS-based diversity indices characterize the coverage and the homogeneity of the projection of a chemical library inside a DRCS. They are representative of the diversity with respect to the contour used and thus to the chemical subset used to construct the contour. It is therefore important to know which DRCS contour has to be used e.g. using the HTS contour for characterizing fragment subsets is totally inappropriate.

Based on the DRCS model and on the databases prepared (see Method section), four specific contours were computed (see the supporting information). The initial DRCS contour (representative of the HTS compounds) and the other four specific contours are shown in Figure 7.

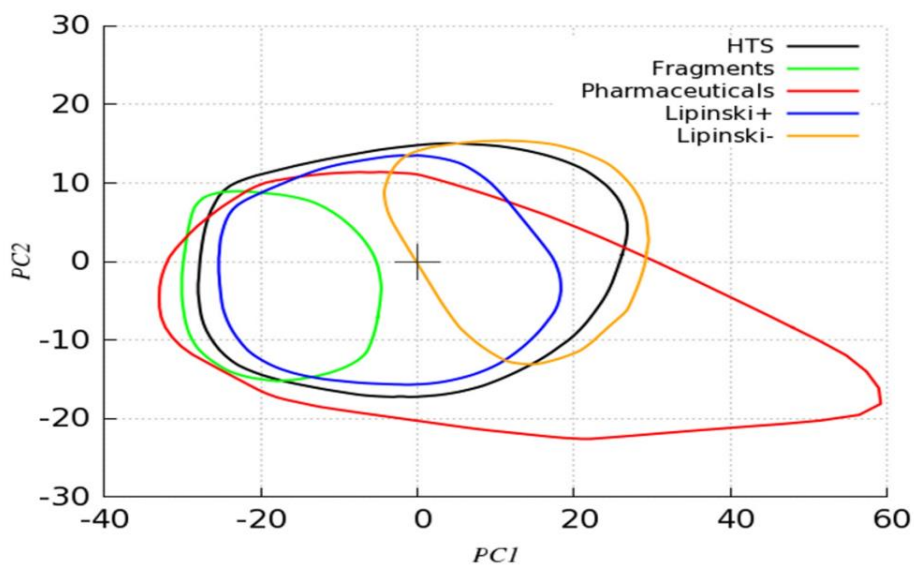


Figure 7: multiple contours representation in the DRCS model.

In the same way as for the initial DRCS contour, the use of representative contours instead of compound projections facilitates the visualization and the characterization of chemical libraries. The projection of a library means that the coverage of different chemical subspaces represented by specific contours can be rapidly estimated, and unexplored zone(s) in these subspaces identified.

Moreover, because the contours represent specific chemical subspaces, the spaces can be compared without the need to project all the compounds. Figure 7 shows that the subspace of

pharmaceutical compounds (red contour) is much broader than the subspace of HTS commercial compounds (black contour). Likewise, the subspace of fragment compounds (green contour) clearly occupies only a small part of the pharmaceutical or HTS commercial compounds spaces. The blue and orange contours represent the subspaces of the compounds that respectively satisfy or do not satisfy the “rule of five”. One region of the space is common to these two contours. This fuzzy limit between the two subspaces is the consequence of the exception accepted in the “rule of five”.

The distribution of a chemical library in different contours reflects different occupations and chemical diversities with reference to each contour. Consequently, the DRCS based indices will not be equivalent but depend on the DRCS contour used. The contour will modify the interpretation of the diversity indices.

Application to the characterization of libraries. The DRCS based methodology (graphical tool and diversity indices) was applied to six publicly available chemical libraries: Prestwick Chemical Library, Chembridge Kinaset, two combinatorial libraries (CL-A3 and CL-B5), Pyxis Discovery Smart Fragment Library and EPAFHM (see Method section for description). This application will illustrate how this methodology should be used, and what kind of information it provides.

Compounds projection: the six libraries were projected in the DRCS. Figure 8 shows the compounds projections in the five pre-computed contours, and allows their visual comparison. The Prestwick collection has a substantial coverage of the HTS contour. Moreover, the majority of the compounds projected outside the DRCS contour appear to be in the pharmaceuticals subspace. The compounds of the Chembridge Kinaset collection are concentrated in a reduced part of the HTS or pharmaceuticals contours. This is what is expected for target-specific libraries. The two combinatorial libraries have the same behavior as the Chembridge Kinaset library. They focus on a small part of the HTS contour. The two libraries contain the same number of compounds (1000), but visually the A3 combinatorial library seems to occupy a higher subspace than the B5 library. The Pyxis collection of fragment compounds

homogeneously occupies a small part of the HTS or pharmaceuticals subspaces, but seems to cover a high part of the fragments contour. The EPAFHM collection also seems cover a high part of the fragments subspace with a high proportion of compounds outside this subspace.

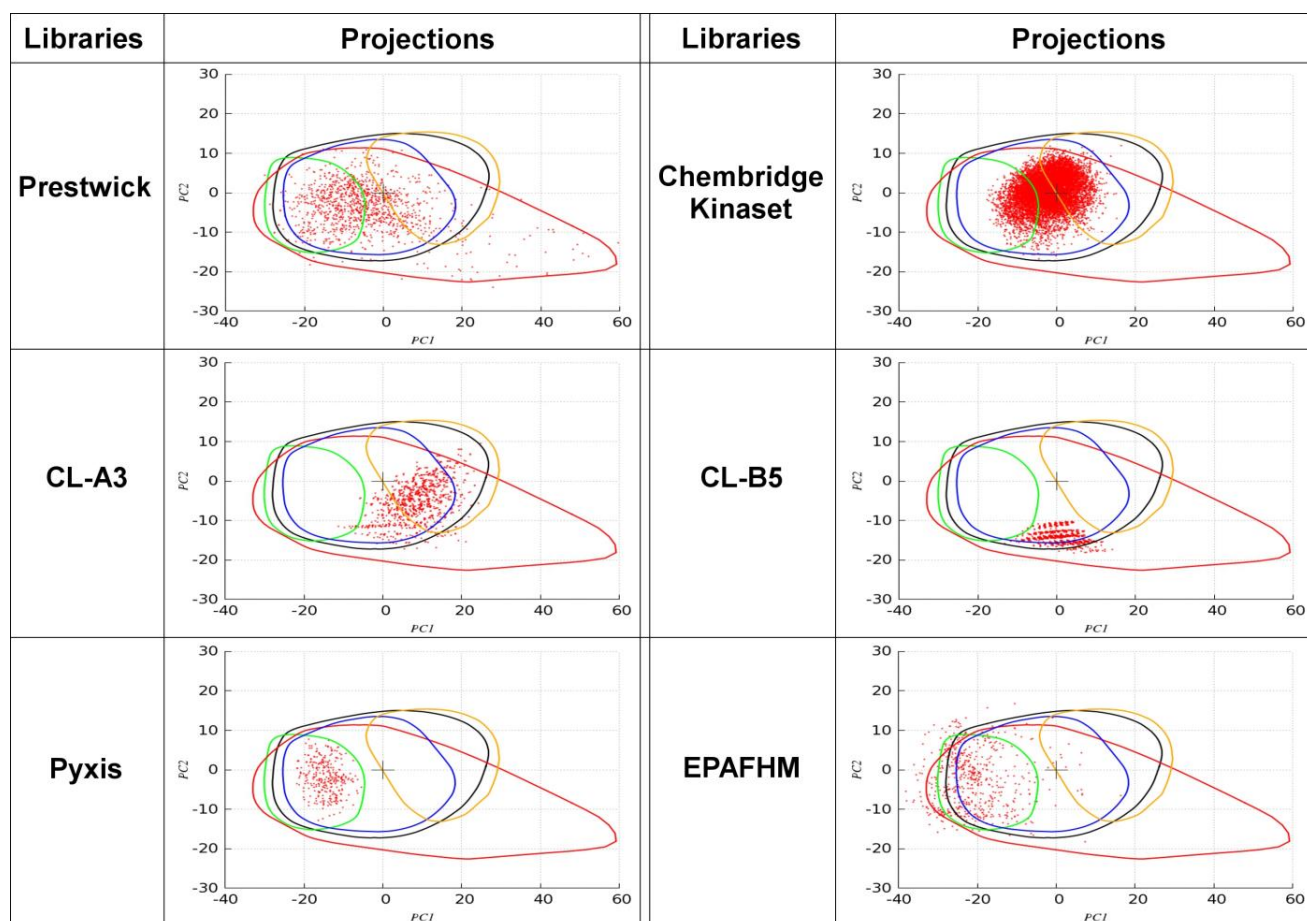


Figure 8: projections of chemical libraries on the DRCS. The compounds are shown by red dots. The contours are presented in the same way as in Figure 7.

DRCS based diversity indices calculation: the five selected DRCS based diversity indices were computed for the same six chemical libraries. The DRCS model and four representative contours (HTS, Pharmaceuticals, Lipinski+ and Fragments; see Methods section for description) were used. Figure 9 to Figure 12 show the four resulting radar graphs (one for each contour).

It is important to note that the EPAFHM library is not represented in the radar graphs. This is because of the large proportion of compounds projected outside the contour (>10%) for all subspaces. In the light of its projection shown in Figure 8, the applicability domains of QSPR models built using this

reference library could obviously be questioned. This library indeed contains a high proportion of very small compounds compared to the other projected libraries, or to the libraries used to build the specific contours. As this seems to demonstrate that none of the subspaces are appropriate to quantify the diversity of this library, it will not be further studied.

- HTS contour: the five libraries have less than 10% of compounds projected outside the HTS contour with the Pyxis and Chembridge Kinaset collections having 100% of the compounds projected inside the contour. Thus these two libraries do not have distinct compounds with respect to the HTS commercial compounds.

The *Filling*(%) shows that the Prestwick library has the highest coverage of the HTS subspace. With 5.81% of external compounds and 40.85% of coverage, this library is the most diverse and covers the subspace of HTS commercial compounds quite well. The Pyxis and CL-B5 libraries, on the contrary, have a low coverage of the subspace (18.13% and 7.57% respectively). They cover the two axes differently, as shown by the significant differences found for the *Dks* indices. The Pyxis library has a better coverage of the PC2 axis ($Dks(PC1) = 0.79$) compared to the PC1 axis ($Dks(PC2) = 0.42$), while the opposite is true for CL B5, which has a better coverage on the PC1 axis ($Dks(PC1) = 0.72$) than on the PC2 axis ($Dks(PC2) = 0.19$). These differences are a good illustration of the complementarity of the selected indices, which account for different information that would not be captured by a single index. Chembridge Kinaset and CL-A3 libraries have similar coverage of the subspace but the *Dks(PC1 and PC2)* indices indicate that Chembridge Kinaset library has a better distribution on each axis. Nevertheless, CL-A3 has a higher value for the *MeanEucDist* index compared to the Chembridge Kinaset library, indicating a better homogeneity of the coverage in the occupied regions. This confirms the visual interpretation of their projection in the DRCS (Figure 8).

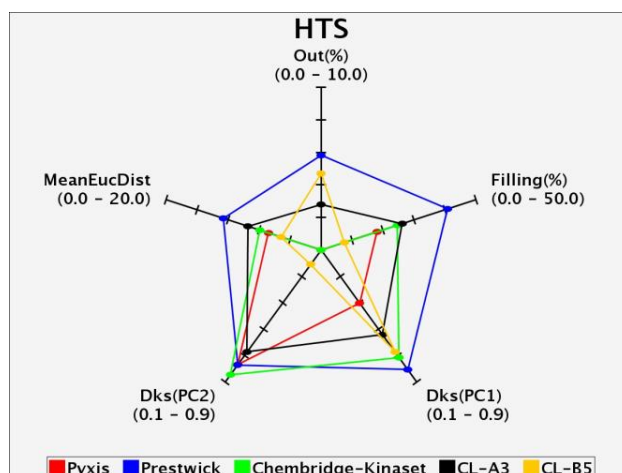


Figure 9: radar graph of the DRCS based diversity indices computed for five chemical libraries. The DRCS model and the HTS commercial compounds contour were used for all the calculations.

- Pharmaceuticals contour: the libraries have a similar profile to that in the HTS contour (Figure 10). However, differences can be observed in the *Dks(PC1 and PC2)* indices. The distribution of the occupied cells in the Prestwick and CL-A3 libraries is the same through each axis of the DRCS model, whereas the homogeneity of Chembridge Kinaset is lower. It can also be noted that in comparison with the HTS contour, the values of the indices are lower in the pharmaceuticals contour for all the libraries. All these differences reflect the incapacity of the libraries to cover the region of pharmaceutical compounds with high values on the PC1 axis (Figure 8).

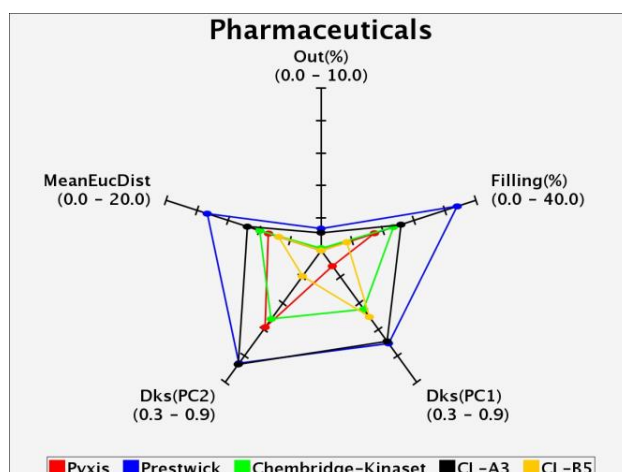


Figure 10: radar graph of the DRCS based diversity indices computed for five chemical libraries. The DRCS model and the pharmaceuticals contour were used for all the calculations.

- Lipinski+ contour: it does not appear to be suitable for studying libraries CL-A3 and CL-B5 (14.00% and 24.80% of external compounds respectively; Figure 11). With a high proportion of external compounds (8.76%), the Prestwick collection has the best coverage of this subspace (46.33%), while Chembridge Kinaset and Pyxis libraries have a lower diversity. It can be noted that differences in the coverage of the subspace by the three libraries impacts only on their distribution through PC1 (similar values for $Dks(PC2)$).

In this case, *MeanEucDist* calculation is not mandatory as the previous indices clearly show the differences between the libraries.

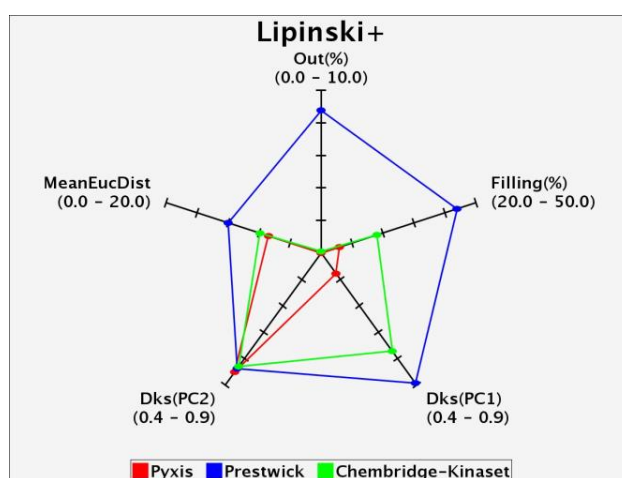


Figure 11: radar graph of the DRCS based diversity indices computed for three chemical libraries. The DRCS model and the Lipinski+ contour were used for all the calculations.

- Fragments contour: only the Pyxis library has less than 10% of external compounds with regard to the fragments contour (Figure 12). This reflects the high specificity of this subspace for the fragments compounds. Here, the Pyxis library has a quite high coverage of this subspace (34.50%), comparable to the coverage of the HTS subspace by the Prestwick collection (Figure 9). Finally, in contrast to the other subspaces, the Pyxis library has a similar distribution through the two axes of the DRCS model. This shows that this library is clearly more diverse and better suited to this subspace.

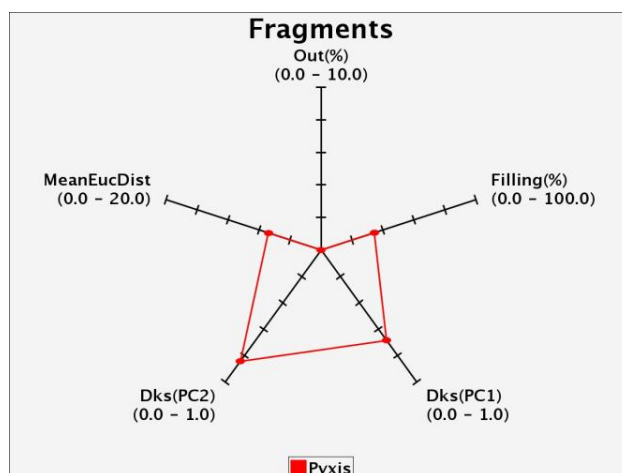


Figure 12: radar graph of the DRCS based diversity indices computed for the Pyxis library. The DRCS model and the fragments contour were used for all the calculations.

It should be noted that the Lipinski- contour was not used. All the chemical libraries studied have more than 10% of external compounds with respect to this contour (in the best case CL-A3 has 26.30% of compounds outside this contour). This is a typical example of a subspace that, in most cases, will not be useful to estimate the diversity of chemical libraries, as most screening libraries are designed to contain drug-like molecules. Such a subspace may rather be used to estimate how a particular library covers an undesired subspace, and if this coverage is acceptable for the problem at hand.

CONCLUSION AND PERSPECTIVES

In a previous paper, the DRCS methodology was introduced as the combination of a PCA model and a subspace delimitation. A subspace, as defined by the DRCS contour, provides a relevant delimitation of the densest zone of the chemical space spanned by the reference library. In this work, the DRCS methodology was used to create a new set of subspaces representing specific types of compounds, and to derive new diversity indices that are independent of the cardinality of the library. As it obviously does not make much sense to e.g. assess the space coverage of a fragment library in general-purpose HTS subspace, we showed that the DRCS methodology can be used to focus on the relevant part of the chemical space that corresponds to the problem at hand. The general-purpose Lipinski, Pharmaceutical and Fragment subspaces were presented, and we showed that despite some significant overlaps, they all span a different zone of the chemical space.

The DRCS contour was also used to obtain a relevant and flexible partition of the corresponding subspace, making it possible to derive diversity indices. The behavior of reference as well as DRCS-based diversity indices was assessed in the light of a new set of rules that a diversity index should satisfy when comparing libraries of different cardinalities. Following this analysis, it was found that, even in simplified fictive situations, none of these indices can account for all the important aspects that characterize the diversity of screening libraries (space coverage, redundancy, uniformity of the distribution...). Hence, five complementary indices were finally selected that account for these different aspects of diversity, and a simple framework is proposed to use them effectively. These indices can be applied to compare libraries containing different numbers of molecules (e.g. after an enrichment operation or to decide between two libraries proposed by external vendors), as well as to other classical diversity problems (e.g. automatic library design).

In conclusion, the methodology proposed in these two papers provides complementary visual and numerical tools that can be used to analyze the diversity of chemical libraries. The overall idea can be summarized as the following: intuitively, the real chemical space is defined by both very dense and

very sparse regions. Dense regions typically contain the large majority of available compounds, corresponding to widely studied chemotypes. In contrast, sparse regions generally contain specific and somewhat exotic molecules, which might nevertheless be included in a screening campaign as well, depending on the context. It becomes evident that the sampling of the two regions should not be performed the same way, and the delimitation of the densest subspace represents a simple way of tackling this issue. The diversity indices as well as the visual characterization provides a way to assess the coverage of a library with respect to the most explored region of a reference library, leaving the sparse regions to a separate analysis. The use of a 2D representation obtained by PCA projection is certainly the most limiting factor of the methodology. The easiest way to overcome this is either to extend the analysis to a 3rd dimension, or to perform the same analysis using different Principal Components, although at the expense of simplicity and ease of interpretation. Another possibility would be the use of non-linear multidimensional scaling techniques, such as Generative Topographic Maps ²⁶. These non-linear methods would be especially useful to define biologically relevant subspaces, e.g. target-specific subspaces, without losing the advantage of visual analysis.

ACKNOWLEDGMENT

The authors thank Accelrys for providing, free of charge, the software "Pipeline Pilot Student edition", and the "Conseil Régional du Centre" for supporting this research. The authors also thank Peter Schmidtke for helpful comments on the manuscript. VLG thanks the "Conseil Général du Loiret" for funding his PhD.

Supporting Information Available

S1: Schematic view of the creation of the fictive libraries A to K; S2: Creation of specific subspaces; S3: Determination of the mathematical relation between N_{bins} of the optimal grid and N_{mol} of a library for a given DRCS; S4: *DivInt* index: definition of the reference points; S5: Behavior of DRCS-based indices: rule 1, influence of the size of the library; S6: Illustration of the overall process of Shannon entropy calculation on PC1; S7: Illustration of the overall process of the Kolmogorov-Smirnov index on PC1 ;Description of tools and computational performances. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

1. Sukuru, S. C.; Jenkins, J. L.; Beckwith, R. E.; Scheiber, J.; Bender, A.; Mikhailov, D.; Davies, J. W.; Glick, M. Plate-based diversity selection based on empirical HTS data to enhance the number of hits and their chemical diversity. *J. Biomol. Screen.* **2009**, *14* (6), 690-699.
2. Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29* (1), 55-67.
3. Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Diversity* **2006**, *10* (3), 389-403.
4. Dubois, J.; Bourg, S.; Vrain, C.; Morin-Allory, L. Collections of Compounds - How to Deal with them? *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 156-168.
5. Medina-Franco, J. L.; Martinez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (4), 322-333.
6. Le Guilloux, V.; Colliandre, L.; Bourg, S.; Guenegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. Visual Characterization and Diversity Quantification of Chemical Libraries: 1. Creation of Delimited Reference Chemical Subspaces. *J. Chem. Inf. Model.* **2011**, *51* (8), 1762-1774.
7. Gillet, V.; Dean, P.; Lewis, R., Background Theory of Molecular Diversity. In *Molecular Diversity in Drug Design*, Springer Netherlands: 2002; pp 43-66.
8. Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14* (6), 501-506.
9. Waldman, M.; Li, H.; Hassan, M. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graphics Modell.* **2000**, *18* (4-5), 412-426, 533-536.
10. Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **2001**, *22* (5), 488-500.
11. Rabal, O.; Pascual, R.; Borrell, J.; Teixido, J. Cell-Integral-Diversity Criterion: A Proposal for Minimizing Cluster Artifact in Cell-Based Selections. *J. Chem. Inf. Model.* **2007**, *47* (5), 1886-1896.
12. Brown, R. D.; Hassan, M.; Waldman, M. Combinatorial library design for diversity, cost efficiency, and drug-like character. *J. Mol. Graphics Modell.* **2000**, *18* (4-5), 427-437.
13. Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *40* (1), 63-70.
14. Bayley, M. J.; Willett, P. Binning schemes for partition-based compound selection. *J. Mol. Graphics Modell.* **1999**, *17* (1), 10-18.
15. Agrafiotis, D. K. A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (1), 159-167.
16. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876-877.
17. Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Del. Rev.* **1997**, *23*, 3-25.
18. MOE version 2009-10; Chemical Computing Group; Montreal, Quebec, Canada
19. Prestwick. <http://www.prestwickchemical.com/> (accessed January 15, 2011)
20. CMC. <http://www.akosgmbh.de/Symyx/software/databases/cmc-3d.htm> (accessed January 15, 2011)
21. Chembridge. <http://www.chembridge.com> (accessed January 15, 2011)
22. Pyxis. <https://www.chemonaut.com> (accessed January 15, 2011)
23. EPAFHM; US EPA Computational Toxicology Program: NC; http://www.epa.gov/ncct/dsstox/sdf_epafhm.html. (accessed November 2, 2010).

24. Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16* (5), 948-967.
25. DrugBank; <http://www.drugbank.ca/>. Accessed January 15, 2011
26. Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; Lopez-Vallejo, F. Visualization of Molecular Fingerprints. *J. Chem. Inf. Model.* **2011**, *51* (7), 1552-1563.
27. Clark, R. D.; Langton, W. J. Balancing Representativeness Against Diversity using Optimizable K-Dissimilarity and Hierarchical Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 1079-1086.
28. Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 1-10.
29. Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases :Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds *J. Chem. Inf. Comput. Sci.* **1996**, *36* (4), 750-763.
30. Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 305-312.
31. Pascual, R.; Mateu, M.; Gasteiger, J.; Borrell, J. I.; Teixido, J. Design and Analysis of a Combinatorial Library of HEPT Analogues: Comparison of Selection Methodologies and Inspection of the Actually Covered Chemical Space. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (1), 199-207.
32. Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2011**, *14* (3), 325-330.
33. Shanmugasundaram, V.; Maggiora, G. Application of Shannon-like diversity measures to cell-based chemistry spaces. *J. Math. Chem.* **2011**, *49* (2), 342-355.
34. Viswanadhan, V. N.; Rajesh, H.; Balaji, V. N. Atom Type Preferences, Structural Diversity, and Property Profiles of Known Drugs, Leads, and Nondrugs: A Comparative Assessment. *ACS Comb. Sci.* **2011**, *13* (3), 327-336.
35. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887-2893.
36. Stein, S. E.; Heller, S. R.; Tchekhovskoi, D., An Open Standard for Chemical Structure Representation - The IUPAC Chemical Identifier. In *2003 International Chemical Information Conference*, Colliers, H., Ed. Infonortics Ltd., Tetbury, UK: Nimes, 2003; pp 131-143.
37. InChI, 1.03, IUPAC, 2010. InChI. <http://www.iupac.org/inchi/> (accessed January 15, 2011)
38. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742-754.
39. *Pipeline Pilot*, student edition, Accelrys, San Diego, CA, 2010.
40. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1177-1185.
41. Daylight Chemical Information Systems, Inc. PO Box 7737, Laguna Niguel, CA 92677; USA
42. Accelrys, Inc. 10188 Telesis Court, Suite 100 San Diego, CA 92121 USA
43. Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **1948**, *27*, 379-423.
44. Dash, M.; Liu, H.; Terano, T.; Chen, A., Feature Selection for Clustering; Knowledge Discovery and Data Mining. Current Issues and New Applications. In Springer Berlin / Heidelberg: 2000; Vol. 1805, pp 110-121.
45. Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 796-800.
46. Godden, J. W.; Bajorath, J. An Information-Theoretic Approach to Descriptor Selection for Database Profiling and QSAR Modeling. *QSAR & Combin. Sci.* **2003**, *22* (5), 487-497.

47. MacQueen, J. B., Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press: Berkeley, 1967; Vol. 1, pp 281-297.
48. Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P. Automatic Subspace Clustering of High Dimensional Data. *Data Min. Knowl. Disc.* **2005**, *11* (1), 5-33.
49. Rassokhin, D. N.; Agrafiotis, D. K. Kolmogorov-Smirnov statistic and its application in library design. *J. Mol. Graphics Modell.* **2000**, *18* (4-5), 368-382.
50. von Mises, R., *Mathematical theory of probability and statistics*. Academic Press: New York, 1997.

Visual Characterization and Diversity Quantification of Chemical Libraries: 2. Analysis and Selection of Size-independent, Subspace-specific Diversity Indices

Lionel Colliandre, Vincent Le Guilloux, Stephane Bourg, Luc Morin-Allory

Supporting Information

S1: Schematic view of the creation of the fictive libraries A to K

S2: Creation of specific subspaces

S3: Determination of the mathematical relation between N_{bins} of the optimal grid and N_{mol} of a library for a given DRCS

S4: *DivInt* index: definition of the reference points

S5: Behavior of DRCS-based indices: rule 1, influence of the size of the library

S6: Calculation of scaled Shannon entropy

S7: Calculation of Kolmogorov-Smirnov index

S1 - Schematic view of the creation of the fictive libraries A to K

A set of fictive libraries was created to study the behavior of diversity indices in various extreme situations. Based on the DRCS, a grid was applied to the 2D subspace, strictly encompassing the DRCS contour. The fictive libraries were created by adding points at the center of selected cells of the grid. Each point represents one compound. Only the cells inside the DRCS can be completed i.e. cells with at least one corner inside the DRCS.

A grid of 2540 cells was used to create library A. It was created to perfectly occupy the subspace i.e. one point was added in each of the cells. Based on the same grid, libraries B, C and D were created by duplicating library A two, four or eight times respectively i.e. two, four or eight points were added in each of the 2540 cells. A schematic view of the creation of these libraries is presented in Figure S1.

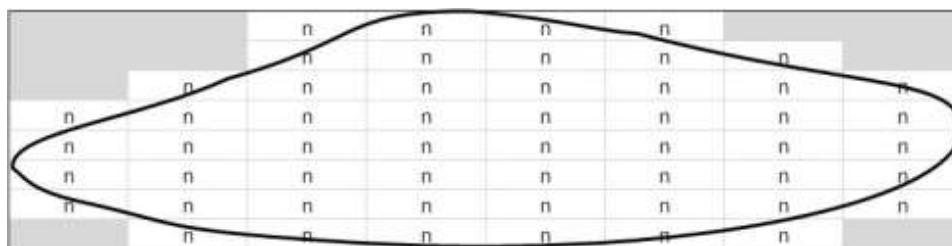


Figure S1: schematic view of the creation of the fictive libraries A, B, C and D. The cells in gray are outside the DRCS. Only the cells in white can be completed. n is the number of points created in each cell (n=1 for library A, 2 for library B, 4 for library C and 8 for library D).

In the same grid, libraries E, F and G were created adding two points per cell for half of the cells. In library E the occupied cells are regularly spread out (Figure S2). In F only the cells in the left part of the subspace are occupied (Figure S3), and in G only the cells at the bottom part of the subspace are occupied (Figure S4).

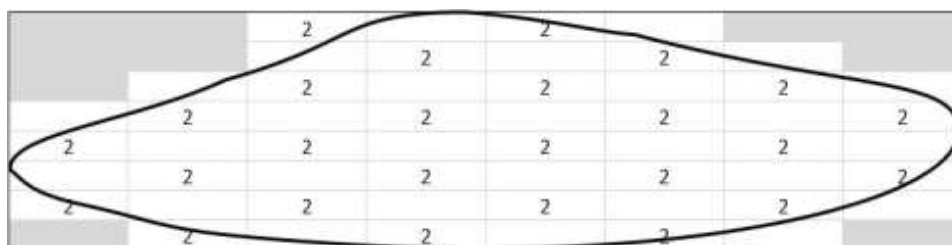


Figure S2: schematic view of the creation of the fictive library E. The cells in gray are outside the DRCS. Only the cells in white can be completed. The number corresponds to the number of points created in the corresponding cell.

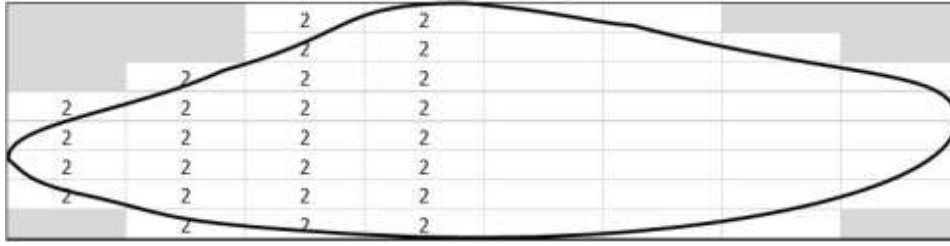


Figure S3: schematic view of the creation of the fictive library F. The cells in gray are outside the DRCS. Only the cells in white can be completed. The number corresponds to the number of points created in the corresponding cell.

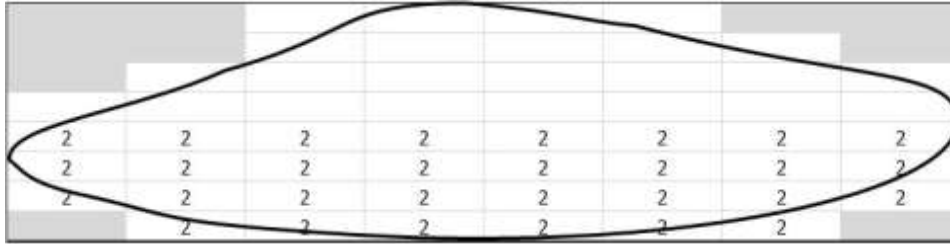


Figure S4: schematic view of the creation of the fictive library G. The cells in gray are outside the DRCS. Only the cells in white can be completed. The number corresponds to the number of points created in the corresponding cell.

As for library A, libraries H, I, J and K were created to perfectly occupy the subspace but grids from around 5000 to 50000 cells inside the DRCS were used (Figure S5).

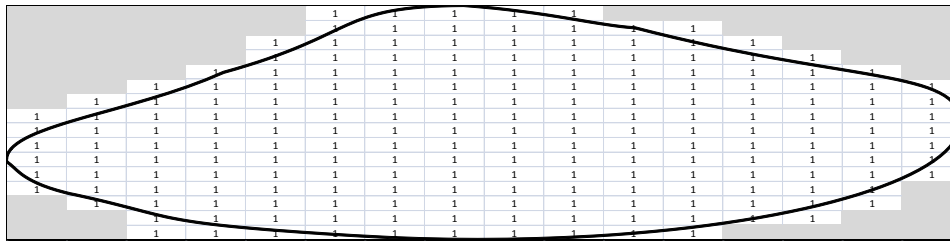


Figure S5: schematic view of the creation of the fictive libraries H, I, J and K. The four libraries differ in the grids used to create the sets of points (here one example presented). The cells in gray are outside the DRCS. Only the cells in white can be completed. The number corresponds to the number of points created in the corresponding cell.

S2 - Creation of specific subspaces

The specific subspaces were computed using nearly the same methodology as that described in a previous article¹. Each contour is computed as the average of 30 convex hulls computed on 30 random subsets extracted from each reference library. For each convex hull, a certain proportion d of outliers is removed using a density-based methodology prior to the calculation of the averaged convex hull. The exact value of d is determined by monitoring the variation in the surface of the final contour for an increasing value of d (ranging from 0% to 10%, increased by 0.1%). The variation in the surface of the final contour converges to a constant value once a sufficient proportion of outliers has been removed. The value of d is selected at the very beginning of this convergence, i.e. once the removal of $d + 0.1\%$ of compounds does not significantly affect the shape of the averaged contour.

In the first article, the size of each subset used to compute the HTS subspace corresponded to around 0.3% of the entire HTS library. To ensure a good correspondence between the two articles, this same percentage was used for the other libraries. For small libraries however, a minimum of 2 000 compounds was fixed: as the convergence of the final contour is determined by adding 0.1 to d at each iteration, this number ensures that at least 2 compounds will be removed at each iteration. Table S1 summarizes the results obtained for each specific library.

Table S1: size of each library, and parameters used to compute the specific subspaces.

Subspace	Number of molecules	Number of molecules	Percentage of outliers
		per subset	removed (d)
HTS	6 630 000	20 000	0.25%
Lipinski+	6 200 000	18 000	0.4%
Lipinski-	413 098	2 000	2.2%
Fragment	122 951	2 000	1.1%
Pharmaceutical	8 932	2 000	4.2%

The larger value of d for the Pharmaceutical subspace can be explained by the higher diversity of this library. As there are many more compounds located in sparse regions of the space, the densest subspace is more difficult to locate, and a larger proportion of outliers have to be removed to obtain a stable shape. This observation holds true for the Lipinski- non-druglike subspace, where only compounds that do not satisfy the rule of 5 (typically very large or highly hydrophobic compounds) criteria are considered. For all the other subspaces, the final percentage of removed molecules remains relatively low.

S3 - Determination of the mathematical relation between N_{bins} of the optimal grid and N_{mol} of a library for a given DRCS

For the implementation of cell-based indices, we used grids that strictly encompass the DRCS contour. In these grids, that are adapted to the studied library, the number of cells that are inside the DRCS (N_{cells}) must be equal to the number of compounds in the library (N_{mol}). However, because the contour is not squared, the number of bins (N_{bins}) that need to be used to create the grid is not proportional to N_{cells} (and thus not proportional to N_{mol}). The relation between N_{bins} and N_{mol} (or N_{cells}) has been found to be a power relation:

$$N_{bins} = \alpha_S \times N_{mol}^{\beta_S}$$

The two coefficients are dependent on the shape of the contour. To obtain these coefficients, 250 grids were created with N_{bins} from 10 to 2500 (one grid every 10 N_{bins}). Then for each grid the number of cells falling inside the DRCS was computed (a cell is considered as being inside the contour if at least one of its corners is inside the contour). Thus a power regression between N_{bins} and N_{mol} (equal to N_{cells}) can be made to obtain the two coefficients. The coefficients for the seven DRCS presented in this series of two papers are given in Table S.

Table S2: coefficients of the power relation between N_{bins} and N_{mols} for seven DRCS.

DRCS			Coefficients	
Descriptors used for the DRCS calculation	Compounds used for the PCA model calculation	Compounds used for the Contour calculation	α	β
MOE2D	HTS	HTS	1.0725	0.5025
MOE3D	HTS	HTS	1.0859	0.5027
CDK2D	HTS	HTS	1.1307	0.5030
MOE2D	HTS	Pharmaceutical	1.1753	0.5032
MOE2D	HTS	Fragment	1.1476	0.5028
MOE2D	HTS	Lip+	1.0816	0.5025
MOE2D	HTS	Lip-	1.1145	0.5029

S4 - *DivInt* index: definition of the reference points

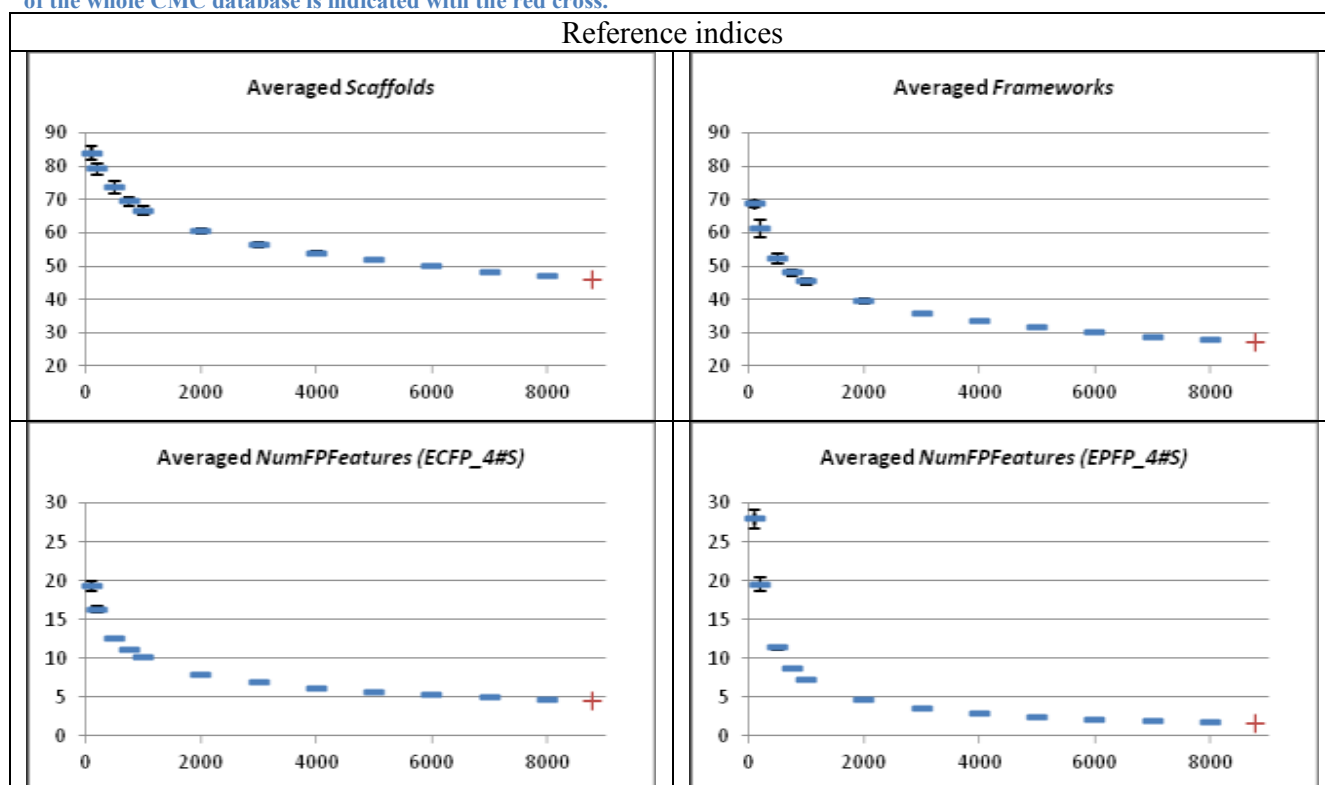
For each calculation of the *DivInt* index, one set of around 1000 points is used as a reference. In order to obtain a homogeneous distribution of the points, their positions are deduced from an optimal grid.

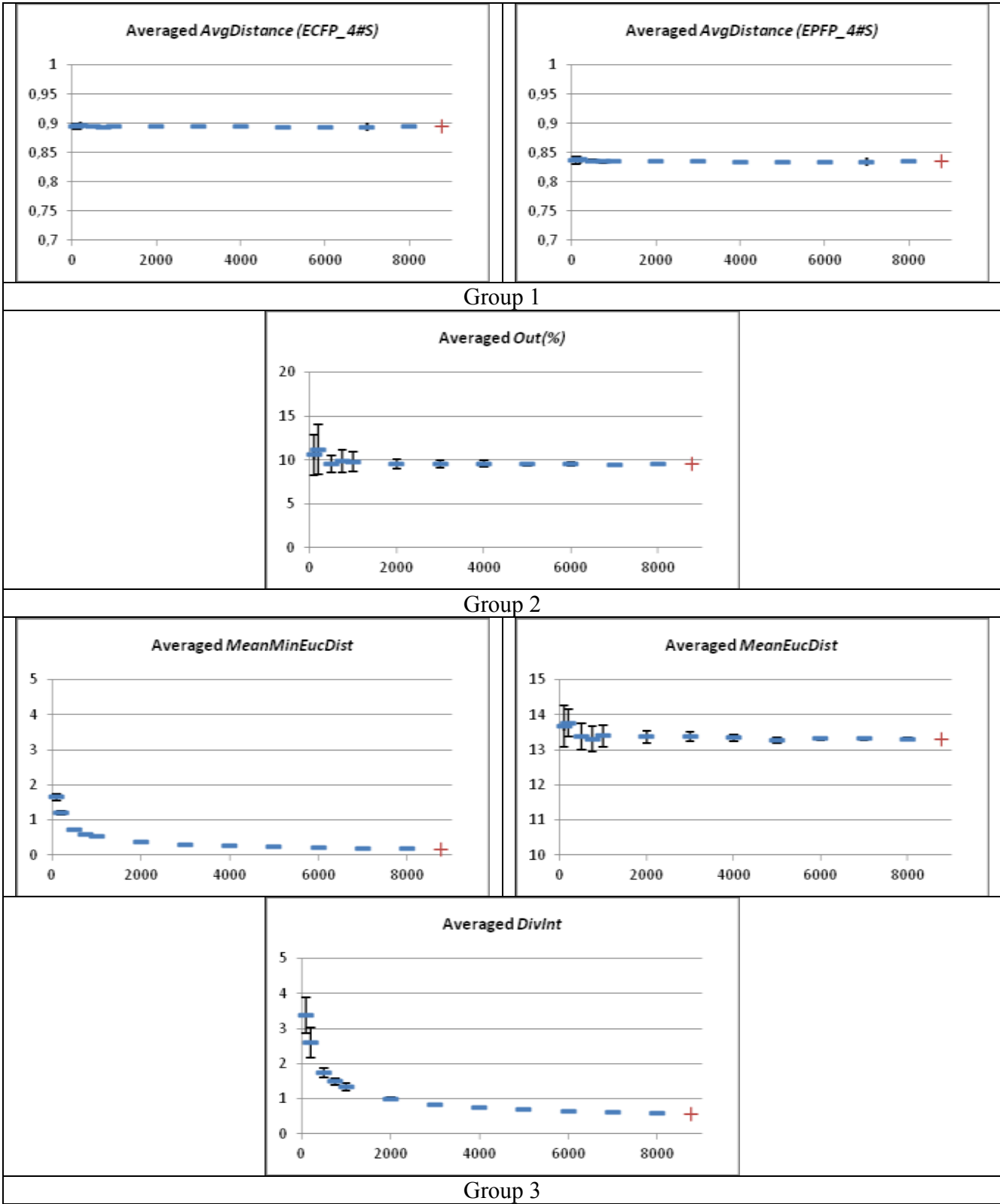
For each DRCS contour, a mathematical relation was computed to obtain, from the number of cells required to be inside the DRCS contour (N_{cells}), the optimal number of bins of the corresponding grid (N_{bins}). For the *DivInt* index, we used this mathematical relation to create a grid containing around 1000 cells inside the DRCS contour. N_{bins} is thus deduced from $N_{\text{mol}} = 1000$. Because the value of N_{bins} that is obtained is not an integer, the N_{bins} used corresponds to the rounded value. In the created grid, one point was created at the center of all the cells inside the DRCS. When the computed point is not inside the DRCS contour i.e. for the cells on the edges of the contour, the point is deduced from the coordinate of one of the corners of the cell that is inside the DRCS contour. We thus obtain as many points as cells in the optimal grid. Because the grid construction depends only on the shape of the DRCS contour used, for one DRCS contour the same set of points will be computed and used for all the calculation of the *DivInt* index. The *DivInt* values will therefore be strictly comparable for all the libraries projected in the same DRCS contour.

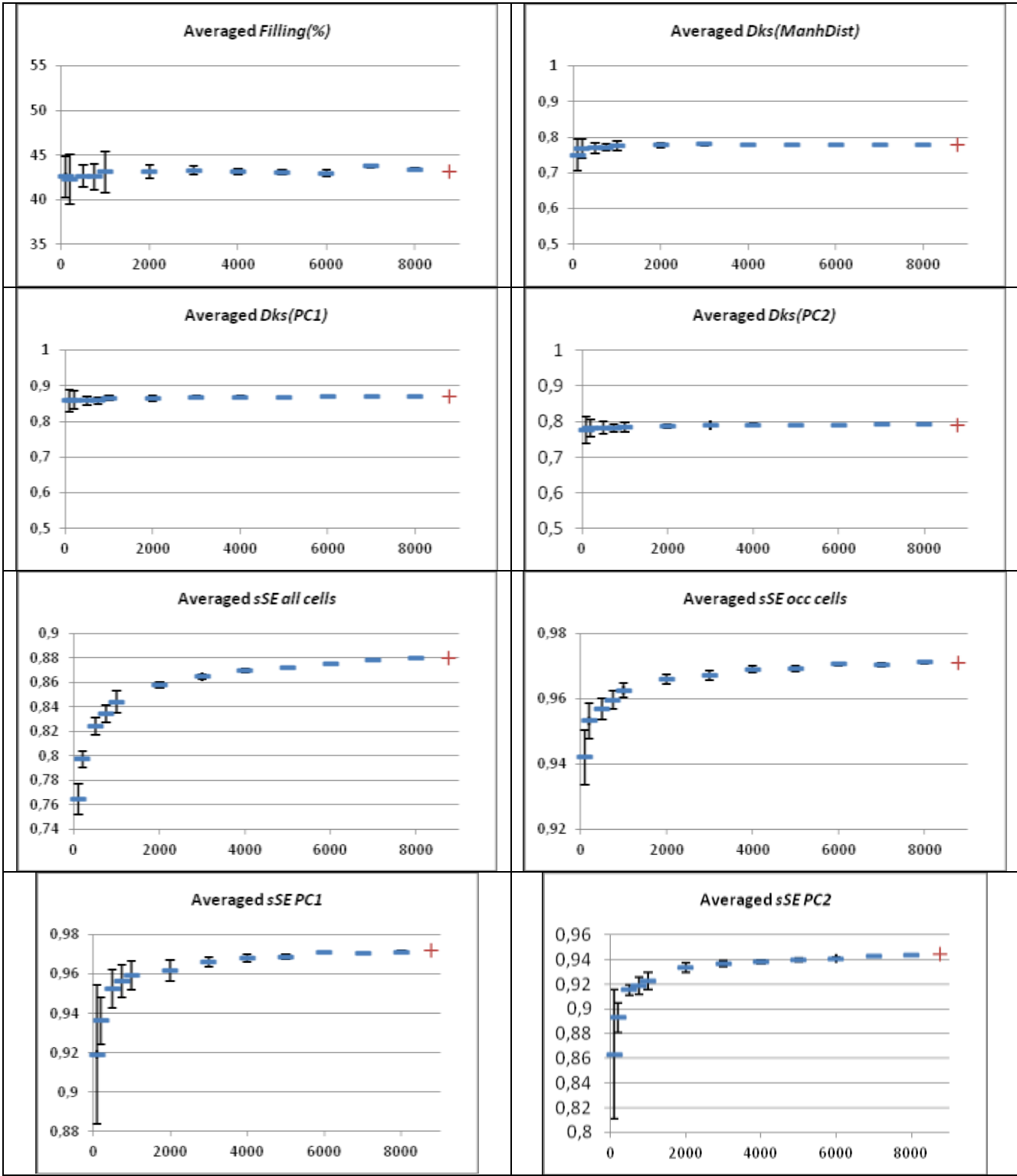
S5 - Behavior of DRCS based indices: rule 1, influence of the size of the library

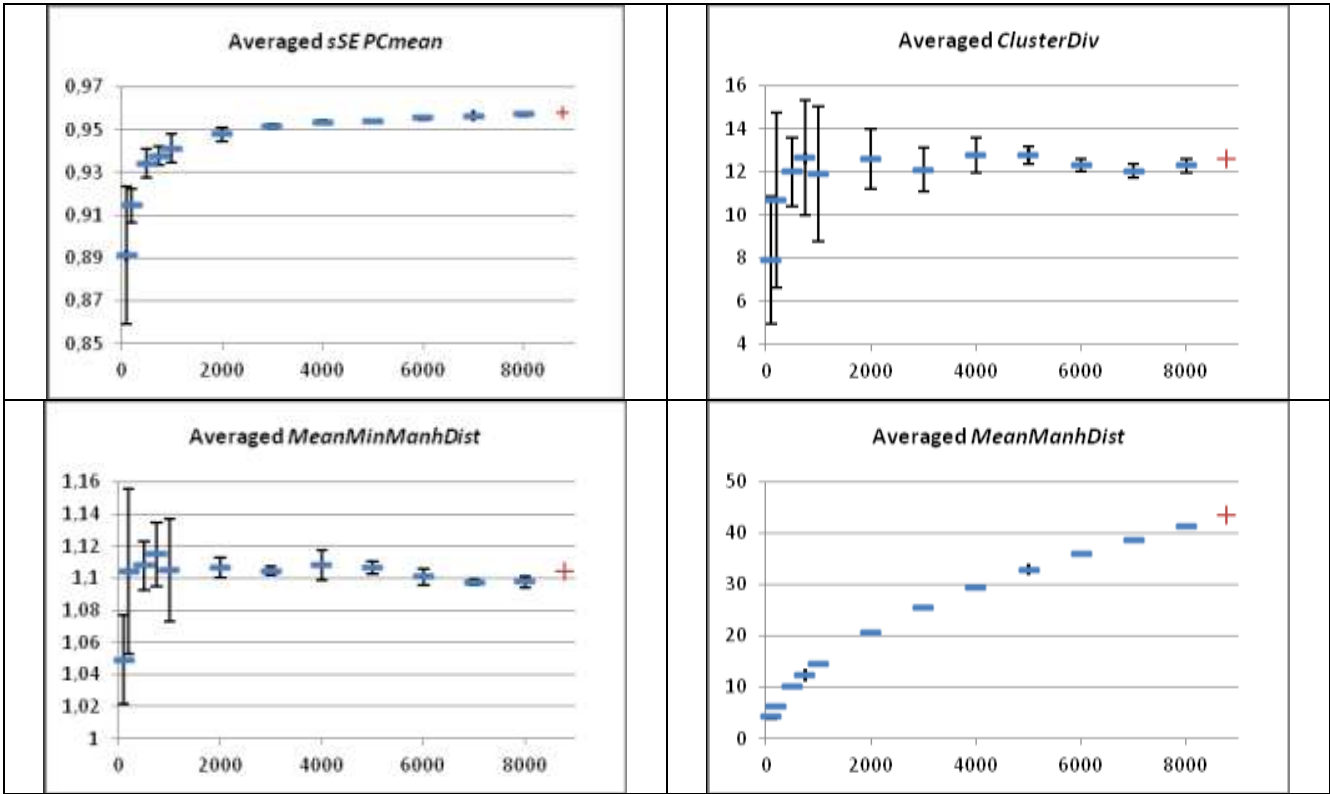
The representativity of the DRCS based diversity indices was verified by computing and studying the variation in the indices with respect to the size of the CMC database subsets. All the indices implemented were computed for all the 60 random subsets of the CMC database (see Method section of the paper). The subsets vary from 100 to 8000 compounds (12 sizes of subsets; 5 subsets per size). For each possible size of the subsets, the 5 computed values were averaged. Graphs presenting the evolution of the values of the DRCS based indices are shown in Table S2.

Table S2: graphs of the twenty-two DRCS based indices implemented versus the size of the CMC subsets. The indices are presented according to their group (see Method section of the paper). Each point in blue represents the average of the index value of five subsets of the same size. The error bars are limited to one standard deviation calculated on the five subsets. The index value of the whole CMC database is indicated with the red cross.









S6 - Scaled Shannon entropy: supplementary information

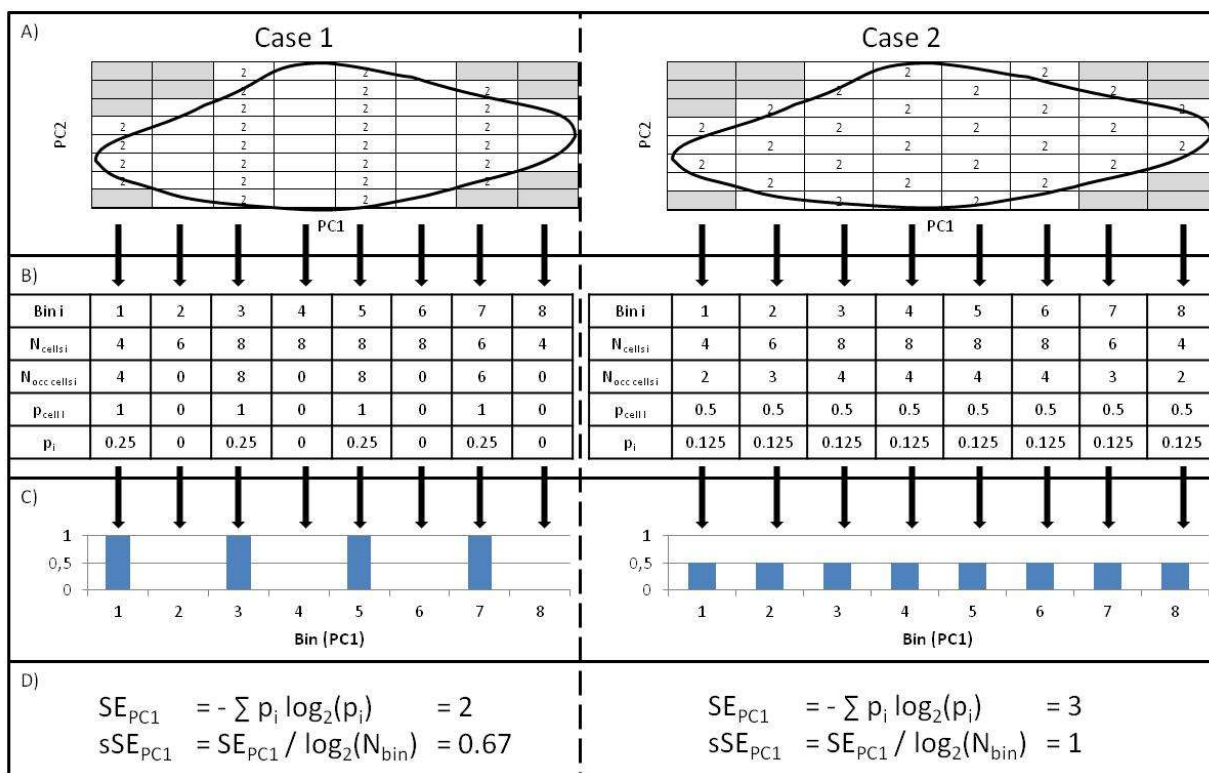


Figure S6: For two fictive libraries made of 52 products (then 52 cells inside the DRCS), description of the successive steps for the sSE_{PC1} indice calculation. A) compounds are projected inside the DRCS and associated to a cell of the optimized grid. B) for PC1, each column of the grid is considered as a bin. For each bin i the probability p_i is computed. C) the probability of occupancy of each bin is obtained ($\sum p_i = 1$). D) Shannon entropy and scaled Shannon entropy are obtained.

S7 - Kolmogorov-Smirnov index: supplementary information

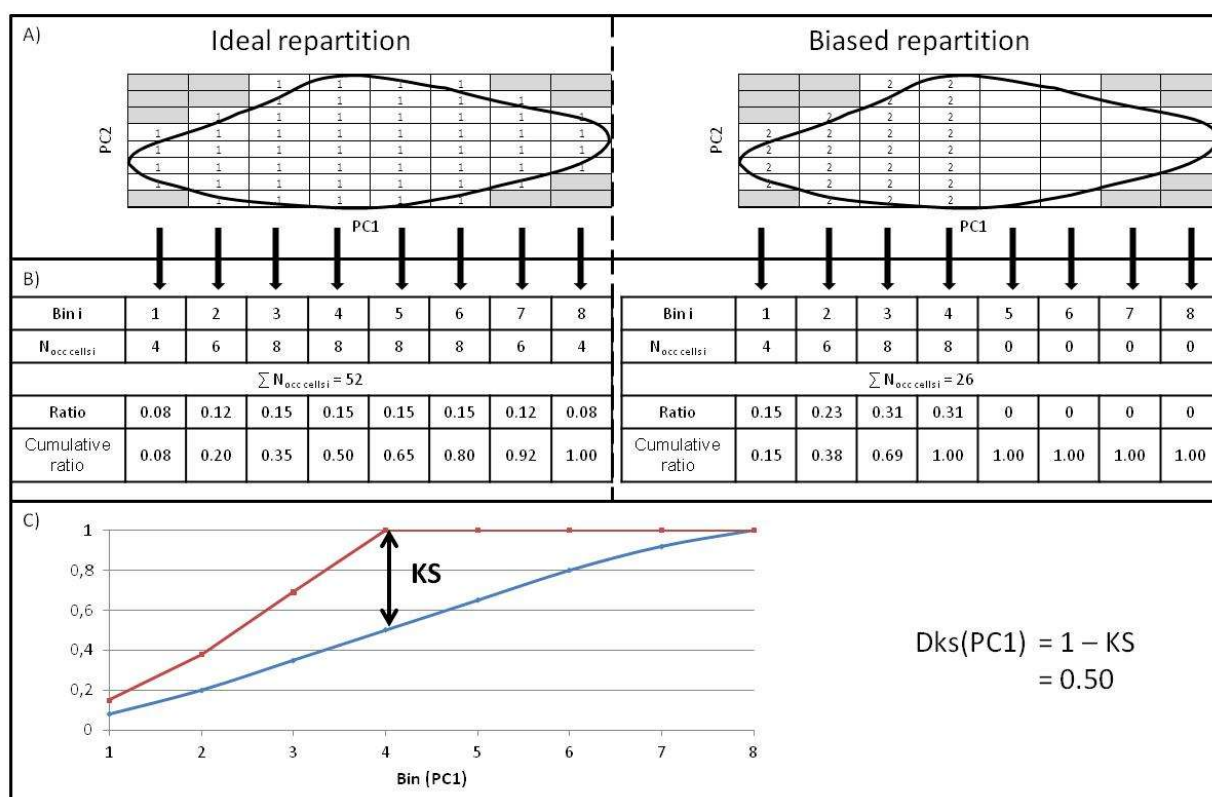


Figure S7: For two fictive libraries made of 52 products (then 52 cells inside the DRCS), description of the successive steps for the $D_{ks}(PC1)$ indice calculation. A) compounds are projected inside the DRCS and associated to a cell of the optimized grid. B) for PC1, each column of the grid is considered as a bin. For each bin i the ratio between the number of occupied cells in the bin and the total number of occupied cells in the DRCS is computed. Then the cumulative ratio from the first to the last bin is obtained. C) the experimental (here biased) cumulative ratio is compared to the cumulative ratio for the ideal repartition to obtain the KS. The $D_{ks}(PC1)$ index is deduced from the KS value.

Tools and computational performance

All the programs and algorithms were developed in-house using the JAVA programming language. The power regression was implemented using the Statistics class of JFreeChart library². The radar graphs were also generated using the JFreeChart library.

All calculations were performed on recent computers with Intel Core 2 Duo P9400 2.40GHz processor and 4GO DDR3 RAM. The creation of 250 grids of N_{bins} from 10 to 2500, the calculation of the number of cells inside the DRCS for each grid, and the power regression with the 250 points took nearly 3 hours. The calculation of the 16 DRCS based indices (see groups 1, 2 and 3 indices in the paper) took nearly 3 minutes for a library of 20000 compounds. Here the Euclidean distances based indices led to the highest time cost calculation (more than 2 minutes).

References of Supporting Information

1. Le Guilloux, V.; Colliandre, L.; Bourg, S.; Guenegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. Visual Characterization and Diversity Quantification of Chemical Libraries: 1. Creation of Delimited Reference Chemical Subspaces. *J. Chem. Inf. Model.* **2011**, *51* (8), 1762-1774.
2. JfreeChart library; <http://math.nist.gov/javanumerics/> (accessed January 15, 2011).

Chapitre IV

Conception et développement de Screening Assistant 2

Dans ce chapitre nous présenterons le développement de Screening Assistant 2 (SA2), un logiciel open-source dédié à la gestion et à l'analyse de chimiothèques à l'aide de méthodes chémoinformatiques. Nous commencerons tout d'abord par présenter l'historique de ce projet et nous ferons ensuite un état de l'art des outils chémoinformatiques disponibles en libre. Nous décrirons ensuite les objectifs et le cahier des charges que nous avons mis en place. Nous présenterons enfin les principales fonctionnalités du logiciel en discutant les différents choix qui ont été faits, aussi bien en termes d'architecture logicielle que de sélection des outils chémoinformatiques sur lesquels repose SA2. Nous conclurons en discutant des points forts et des points faibles de l'outil et des développements futurs qui pourraient être envisagés. Le logiciel a par ailleurs été décrit et son utilisation illustrée à travers l'analyse de chimiothèques dans une publication qui sera présentée dans le chapitre suivant.

1 Bref historique du projet

Screening Assistant (SA) est un projet s'inscrivant dans la continuité des travaux de thèse entrepris par Aurélien Monge entre 2002 et 2006, et présentés en 2006 [25]. Dans le cadre de cette thèse, il a développé Screening Assistant 1 (SA1), un logiciel dédié à la gestion et l'analyse d'ensembles de molécules utilisées pour le criblage. L'origine du projet tient au constat qu'il n'existe aucun logiciel open-source répondant à l'ensemble des problématiques liées à la gestion de chimiothèques : gestion des doublons, mesure de diversité, sélection de sous-ensembles divers, création et visualisation d'espaces chimiques, gestion de gros volumes de données, etc. L'ICOA dispose pourtant de nombreux outils (MOE, Schrödinger, etc.) permettant des traitements avancés en modélisation moléculaire, mais aucun d'entre eux ne permettait (à la date

du début du projet) une gestion complète et efficace de grands ensembles de molécules dédiées au criblage. Ces logiciels contiennent pourtant une grande partie des 'briques' nécessaires au développement d'un tel outil : API pour le traitement informatique de molécules, calcul de descripteurs moléculaires, standardisation des molécules. Ces briques sont néanmoins souvent proposées sous la forme de modules et n'ont pas été reliées de manière à former un ensemble cohérent permettant de gérer efficacement les problématiques liées à la gestion de chimiothèques. Bien qu'il soit tout à fait possible de développer des extensions de ces outils, le choix a été fait de ne pas se rendre dépendant de logiciels commerciaux. Le groupe de modélisation de l'ICOA a par ailleurs souhaité contribuer à la croissance de l'écosystème des logiciels open-source en chémoinformatique en faisant le choix de développer un outil libre.

Une première version (SA1) a donc été mise à disposition de la communauté en 2006 sous licence GPL, une licence open-source très utilisée et qui autorise la redistribution de l'outil à la condition que cette même licence soit conservée. Les objectifs initiaux de SA étaient les suivants [25] :

- Le logiciel doit être gratuit ou très abordable.
- Le logiciel doit être utilisable par un chimiste médicinal (sans connaissances particulières en chémoinformatique).
- Un seul exemplaire de chaque structure doit être stocké (cela implique une gestion efficace des doublons).
- La visualisation des propriétés (diversité, doublons, pourcentage de composés « drug-like »...) des bases des fournisseurs doit être possible.
- Le logiciel doit permettre de générer des ensembles de molécules de manière pertinente (cela implique le calcul de propriétés physicochimiques, la prédiction des caractères « drug-like » et « lead-like » des composés, la génération d'ensembles divers...).

Tout en conservant l'idée initiale du projet, à savoir proposer un logiciel libre de gestion de chimiothèques, nous avons souhaité apporter de nombreuses améliorations à l'outil afin de le rendre plus polyvalent, plus facile d'utilisation, et plus flexible dans la gestion des ensembles de molécules.

Dans la mesure où le logiciel a été développé sous licence open-source, nous allons tout d'abord réaliser un bref état de l'art des outils gratuits et redistribuables dans le domaine de la chémoinformatique.

2 L'open-source en chimoinformatique

Un outil permettant de gérer et d'analyser des chimiothèques devra avant toute chose être capable de comprendre ce qu'est une molécule et offrir la possibilité de les manipuler. Dans le contexte du développement d'un outil open-source, il devient donc nécessaire de n'utiliser que des librairies gratuites et pouvant être redistribuées. Plutôt que de réinventer la roue, nous avons choisi de nous tourner vers l'utilisation de librairies chimoinformatiques open-source.

Contrairement à la bioinformatique, les logiciels libres sont un peu moins répandus dans le domaine de la chimoinformatique et sont plus souvent accessibles sous forme commerciale. Nous présenterons ici brièvement les différentes librairies disponibles en open-source dont la Figure IV.1 en retrace l'historique. Nous ne discuterons donc pas les librairies commerciales ou ne pouvant pas être redistribuées, telles que CACTVS, Daylight Toolkit, OELib ou encore Marvin Beans. Bien que certaines de ces librairies soient disponibles gratuitement pour les académiques, elles ne peuvent être utilisées qu'en interne et non dans le cadre de redistribution d'un outil open-source comme c'est le cas pour Screening Assistant. Nous avons par ailleurs souhaité ne pas nous lier à des outils commerciaux et donc propriétaires, pouvant être sujet à des modifications de politique de licences (comme ce fut le cas par exemple pour Pipeline Pilot il y a peu !).

2.1 Avantages et inconvénients

L'utilisation d'outils open-source présente probablement autant d'avantages que d'inconvénients qu'il convient de garder à l'esprit. Du point de vue de la disponibilité, l'utilisation d'un outil libre garanti sa pérennité puisque l'on n'est pas restreint à l'utilisation d'une API propriétaire. La mise à disposition du code source laisse par ailleurs aux développeurs et aux contributeurs une grande flexibilité, que ce soit pour analyser ou vérifier un algorithme, corriger un bug, ou encore améliorer une méthode existante. Elle offre en outre une plus grande transparence sur la compréhension des traitements exécutés, certains logiciels commerciaux ayant parfois tendance à passer sous silence certains événements non gérés.

L'open-source est également un moyen efficace de partage entre les chercheurs du monde entier. Une nouvelle méthode diffusée par ce moyen pourra ainsi être réutilisée par tout un chacun. Cela permet entre autre une plus large diffusion du savoir et offre à tous la possibilité de comprendre, réutiliser et améliorer une méthode existante. De ce point de vue, l'un des

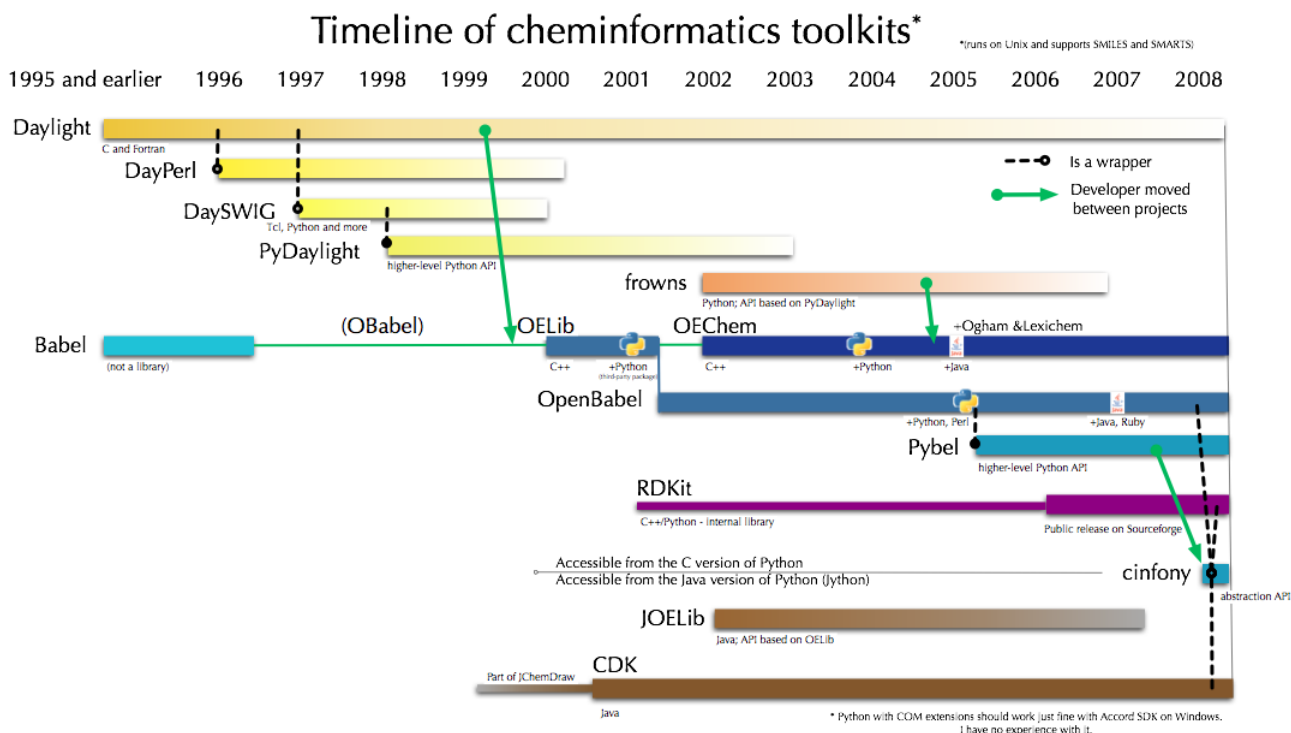


Figure IV.1 – Historique d'apparition des librairies en chémoinformatique. Figure réalisée par Andrew Dalke et présentée lors de l'EuroQSAR 2008

avantages clés du modèle open-source est certainement de permettre la reproduction de résultats publiés. De nombreuses études pointent en effet du doigt l'absence ou l'extrême difficulté à reproduire des résultats issus d'études scientifiques, la modélisation ne faisant pas exception à la règle [18, 22, 35]. Cet état de fait est d'autant plus surprenant pour un domaine où l'informatique joue un rôle central; reproduire un résultat issu de l'utilisation d'un programme informatique est à priori bien plus facile que pour des résultats expérimentaux. L'utilisation d'un outil commercial restreint dès le départ la possibilité de reproduire les résultats aux utilisateurs possédant l'outil, ce qui finalement réduit la portée des nouvelles méthodes développées de cette manière.

Bien que de nombreux avantages aient été mis en évidence, il existe aussi certains inconvénients non négligeables à l'utilisation d'outils libres. Développer, tester, maintenir et documenter un logiciel sont des activités chronophages, et à quelques exceptions près, le maintien de logiciels open-source se fait sur la base du volontariat. L'évolution de ces outils dépend donc fortement du degré d'implication des contributeurs principaux et du temps dont chacun dispose pour participer aux développements. On trouve ainsi de nombreux exemples de projets open-

source ayant été plus ou moins abandonnés. En chimoinformatique, on pourra par exemple citer JOELib [4] ou encore MX [5].

L'un des principaux reproches fait aux logiciels open-source est le manque de documentation. Il en va de même pour la qualité du code, que ce soit en termes de conception, de performance ou encore de couverture des tests unitaires. Ce dernier point reste néanmoins discutable. Il semble logique qu'un éditeur commercial mette en avant à ses clients que son logiciel est parfaitement bien conçu et testé. Il n'existe néanmoins aucun moyen de vérifier de telles affirmations (sauf pour les logiciels open-source!), et chacun sait qu'au final, un logiciel sans bug n'existe malheureusement pas.

Si l'on se réfère au fonctionnement des communautés open-source en informatique "généraliste", le cas de la chimoinformatique (et de tous les domaines situés à l'intersection de plusieurs autres domaines scientifiques) s'avère néanmoins assez spécifique. Tout comme en bioinformatique, le développement de nouvelles méthodes et outils se fait la plupart du temps par des scientifiques n'étant pas formés à l'origine pour le développement informatique. De plus, les développements se focalisent généralement sur les aspects scientifiques, ce qui reste parfaitement logique dans le cadre d'un projet de recherche. C'est une des raisons qui peuvent expliquer le manque de documentation, de tests, et parfois de performances, le but n'étant généralement pas systématiquement de produire un logiciel "professionnel" et redistribuable, mais de développer et tester une méthode en vue de la publier. C'est également l'une des raisons pointées par Walters [35] et qui expliquerait en partie le manque de reproductibilité dans notre domaine.

Étonnamment, il n'existe pratiquement aucune organisation officielle fédérant les projets libres en chimoinformatique. A titre de comparaison, l'informatique généraliste regorge de consortiums et autres fondations à but non-lucratif, comme par exemple la fondation Apache [1]. Ces organisations présentent l'avantage de fournir aux équipes de développement un encadrement "professionnel" à travers la définition de règles de bonnes pratiques qui détermineront l'acceptation ou le refus d'un projet au sein de l'organisation. Les logiciels sont alors généralement beaucoup plus stables, maintenus et documentés. La fondation Apache est un très bon exemple de ce point de vue puisque certains de ses logiciels sont aujourd'hui utilisés partout dans le monde (les serveurs Apache étant certainement l'exemple le plus connu).

De telles organisations n'existent pas (encore) dans notre domaine. Il convient néanmoins de noter l'existence de l'initiative Blue Obelisk [16, 26], qui tente de fédérer les logiciels open-

source en chimoinformatique, et de promouvoir des formats standards et libres pour faciliter la diffusion de données chimiques et biologiques. Bien que prometteuse, cette initiative se focalise actuellement sur la centralisation des outils existant, sans proposer de règles concernant le développement et la qualité des outils y étant associés.

2.2 Librairies chimoinformatiques

On définira ici une librairie chimoinformatique comme étant une API de programmation permettant de faire des opérations avancées sur les molécules : lecture et écriture aux formats standards, recherche de sous-structure, manipulation (ajout / suppression d’atomes ou de liaisons, modifications, fragmentation, etc.), ou encore calcul de descripteurs moléculaires. Ces librairies sont généralement utilisées par des programmeurs, au sens où elles ne peuvent être utilisées qu’à travers des appels à des méthodes exposées par l’API. Elle forment bien souvent le cœur d’applications destinées aux utilisateurs et disposant d’une interface graphique, comme c’est le cas pour Screening Assistant. Les librairies principales de chimoinformatique sont listées dans le tableau IV.1, ainsi que la présence ou l’absence de fonctionnalités essentielles dans la perspective du développement d’un logiciel de gestion de chimiothèque (tableau IV.2).

Librairie	Langage	Licence	Actif	Ref
CDK	JAVA	LGPL	Oui	[32, 33]
Indigo	C++, JAVA	GPL	Oui	[3]
JOELib	JAVA	GPL	Non	[4]
Openbabel	C++, Python	GPL	Oui	[26]
RDKit	C++, Python	BSD	Oui	[9]
InChI	C	BSD	Oui	[7, 17]

Tableau IV.1 – Liste des librairies principales de chimoinformatique open-source. La colonne ‘Actif’ indique si le projet est toujours maintenu. Trois licences peuvent être trouvées ici. La GPL est la plus restrictive, puisque que tout outil basé sur une librairie GPL doit lui même être redistribué sous cette même licence. La LGPL est assez similaire, mais permet la redistribution de la librairie sous sa forme binaire sans imposer la licence du produit final. La BSD est la plus permissive et permet globalement au développeur de faire ce qu’il souhaite avec les outils distribués sous cette licence.

Comme on peut le constater, la plupart des librairies possèdent presque toutes les fonctionnalités de base requises pour travailler sur des objets moléculaires. Étant donné que SA2 est écrit en JAVA, nous avons tout d’abord conservé les librairies pouvant être utilisées avec ce langage (JOELib, CDK et Indigo à la date de début du projet), afin de ne pas avoir à écrire

Librairie	CDK	Indigo	RDKit	OpenBabel	JOELib	InChI
Lecture de SDF	Oui	Oui	Oui	Oui	Oui	Oui
Ecriture de SDF	Oui	Oui	Oui	Oui	Oui	Oui
Code canonique	Oui	Oui	Oui	Oui	Oui	Oui
Recherche de sous-structures	Oui	Oui	Oui	Oui	Oui	Non
Recherche de motifs SMARTS	Oui	Oui	Oui	Oui	Oui	Non
Fingerprints	Oui	Oui	Oui	Non	Non	Non
Descripteurs	Oui	Non	Oui	Oui	Oui	Non
SMIRKS	Non	Oui	Oui	Non	Non	Non
Génération d'images	Oui	Oui	Oui	Oui	Non	Non
Outil de tracé de molécules	Oui	Non	Non	Non	Non	Non

Tableau IV.2 – Présence de fonctionnalités essentielles à la manipulation de molécules dans les librairies principales de chimoinformatique. Note : l'outil de tracé de molécule du CDK fait partie du projet JChemPaint, mais celui-ci est intimement lié au CDK et les auteurs de ces deux outils proposent une librairie combinant les fonctionnalités du CDK et de JChemPaint. Par ailleurs, GGA software (l'entreprise derrière indigo) propose un outil de tracé web (utilisable uniquement dans un navigateur internet) qui ne peut donc pas être utilisé en client lourd.

de code intermédiaire permettant de faire le lien entre deux langages différents, et ayant par ailleurs un impact sur la performance. Nous verrons par la suite que nous avons finalement utilisé ces trois librairies pour le développement de SA2, toutes ayant leurs avantages et leurs inconvénients.

A noter également l'existence de Cinfony [27], une librairie permettant d'utiliser la plupart des libraires chimoinformatiques à travers une API unique écrite en Python, facilitant ainsi la comparaison entre les différentes implémentations existantes.

2.3 Cartouches moléculaires : la chimie en base de données

Nativement, les systèmes de gestion de bases de données (SGBD) ne connaissent pas la chimie. Le développement d'outils permettant d'enregistrer des molécules et d'effectuer des traitements avancés sur celles-ci est donc un axe de recherche important en chimoinformatique. Ces outils sont généralement désignés sous le terme de cartouche (cartridge) moléculaire. Quelques outils open-source ont été développés à cet effet ces dernières années. On peut citer notamment MyChem [6] pour MySQL, PgChem [8], Bingo [2] et RDKit [9] pour PostgreSQL,

ou encore OrChem [29] pour Oracle. Ces outils exposent des fonctionnalités permettant l'enregistrement et la recherche de molécules dans une base de données classique : détection de doublons, recherche par sous-structure, recherche par similarité, etc. Leur originalité réside dans l'exposition de ces fonctions via une API SQL, permettant ainsi l'exécution des traitements directement sur le serveur hébergeant la base de données. Elles sont généralement optimisées pour accélérer des traitements lourds tels que les recherches par similarité et par sous-structure.

2.4 Outils graphiques

A la différence des librairies, les outils graphiques proposent des interfaces utilisateur facilitant l'utilisation de librairies chimioinformatiques. La plupart de ces outils repose sur une ou plusieurs librairies citées précédemment. Bioclipse par exemple [31] utilise le CDK ainsi que JOELib dans l'optique de proposer une plateforme combinant des méthodes chimioinformatiques et bioinformatiques. Il est ainsi possible de charger des fichiers SDF et visualiser les molécules, de calculer des descripteurs, mais également de charger et visualiser des protéines au sein de cette même plateforme.

KNIME est un outil de gestion de flux de traitements permettant de combiner et d'automatiser des tâches récurrentes. Bien qu'il ne soit pas orienté chimioinformatique, de nombreux modules ont été développés et mis à disposition de tous. Pratiquement toutes les librairies citées précédemment sont ainsi encapsulées dans KNIME. Taverna [24] est un logiciel similaire à KNIME, mais moins riche en termes de fonctionnalités chimioinformatiques. A ce jour, seul le CDK a été intégré à ce logiciel [21].

Ambit [19] est un outil basé sur le CDK. Il facilite l'enregistrement de molécules dans le cadre des directives européennes REACH. Il est distribué sous la forme d'un outil bureautique, mais également sous la forme d'une application web exposant des services permettant d'interagir à distance avec la base de données. Il permet notamment de calculer des descripteurs, d'effectuer des recherches, ou encore de construire des modèles QSAR. Ce logiciel est assez proche de Screening Assistant sur son architecture, dans le sens où il permet de stocker les données dans une base MySQL. Il n'est en revanche pas dédié à l'analyse de chimiothèques dans le contexte du criblage haut débit.

Enfin, bien que relativement éloigné de SA2, il convient de mentionner Scaffold Hunter [36].

Ce programme permet de générer une hiérarchie de scaffolds en se basant sur la méthodologie décrite par Ertl et al. [15], et de stocker ces informations dans une base de données MySQL. Une interface graphique permet ensuite de naviguer de manière interactive dans ces données. Scaffold Hunter est également basé sur le CDK.

3 Objectifs

Les développements de Screening Assistant ont pour objectif de rendre la gestion et l'analyse d'ensembles de molécules plus simple. L'idée est de regrouper au sein d'une même plateforme un ensemble de méthodes chémoinformatiques afin de répondre au mieux aux problématiques liées à l'analyse et à la comparaison de chimiothèques. L'outil se positionne en amont du processus de criblage (figure IV.2). Bien qu'il soit possible, comme nous le verrons plus loin, d'associer aux molécules un nombre quelconque de propriétés, et donc potentiellement des résultats de criblage expérimental (ou virtuel), le logiciel n'est pas spécifiquement destiné à gérer l'inventaire physique des molécules utilisées. Les plaques, les formes de stockage des produits (poudre, liquide, etc.) ou encore leur localisation physique n'ont pas vocation à être gérés dans cette nouvelle version. L'objectif est donc de s'en tenir aux molécules, à leurs propriétés (unitaires ou globales), et à leur comparaison. Du point de vue du criblage, il s'agit de permettre le chargement de molécules dans l'outil, d'y associer leur provenance, et de permettre d'effectuer un certain nombre d'analyses facilitant la décision quant à la sélection d'une ou plusieurs librairies à cribler.

Nous donnerons ici la liste des principaux points du cahier des charges que nous avons établi initialement. Comme nous le verrons, de nouvelles problématiques ont nécessité la modification de celui-ci, et nous le une fois ces problématiques présentées :

- Conserver les principales fonctionnalités de SA1 (notion de *Providers*, calcul de descripteurs par défaut, visualisation sous forme de tables, génération de rapports de diversité, sélection par diversité).
- Permettre l'ajout unitaire de molécules à l'aide d'un outil de tracé de molécules.
- Permettre la suppression de molécules dans la base.
- Permettre l'ajout de descripteurs supplémentaires (calculés ou importés).
- Autoriser la modification, la suppression ou l'ajout de SMARTS pour marquer les molécules indésirables.

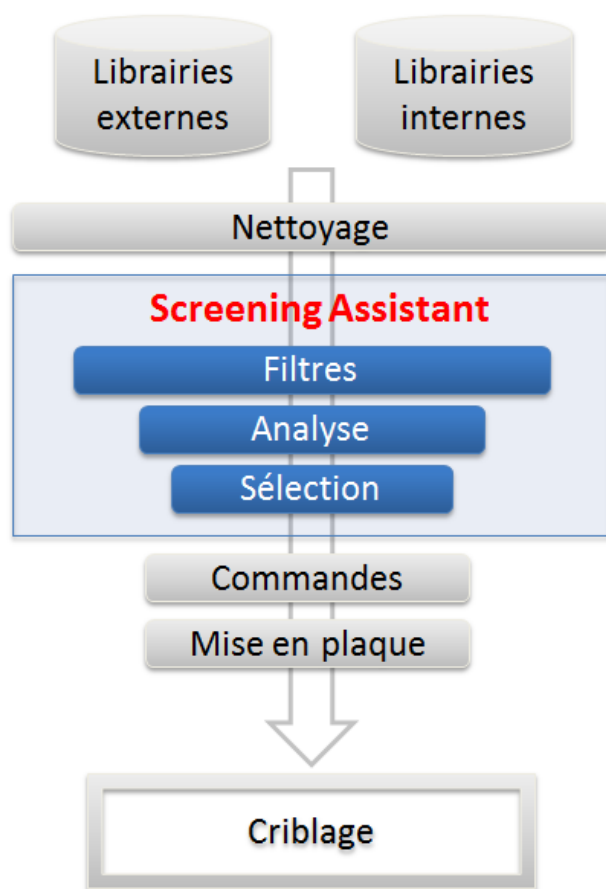


Figure IV.2 – Périmètre d'utilisation de SA2.

- Permettre la visualisation 2D de la structure des molécules et des scaffolds.
- Proposer des fonctions de recherche avancées : par sous-structure, par similarité, etc.
- Proposer une gestion flexible des sous-ensembles de molécules.
- Proposer différentes méthodes de création de sous-ensembles, et notamment des sous-ensembles divers.
- Permettre de calculer de nouveaux espaces chimiques à l'aide de modèles ACP calculés sur les descripteurs choisis par l'utilisateur.
- Permettre une visualisation interactive des données projetées dans des espaces chimiques 2D.
- Permettre une visualisation interactive des données projetées dans des espaces chimiques 3D.
- Introduire la méthodologie DRCS décrite dans le second chapitre.

Après une première analyse de l'architecture de Screening Assistant 1, il s'est avéré que les changements nécessaires à la structure initiale de la base de données (ainsi qu'au programme JAVA responsable des traitements) étaient trop importants pour pouvoir se baser sur SA1. Partant de ce constat, nous avons fait le choix de recoder l'ensemble du logiciel afin de repartir sur des bases permettant l'ajout des nouvelles fonctionnalités définies dans le cahier des charges. Cela nous a également permis d'identifier un certain nombre de points techniques pouvant être améliorés à travers l'utilisation d'outils plus récents.

4 Screening Assistant 2

Nous ne présenterons pas ici l'ancienne version du logiciel, qui est décrite dans la thèse d'Aurélien Monge [25]. En revanche, nous tenterons de lier à celle-ci les choix qui ont été faits durant le développement de SA2. Mis à part la visualisation 3D des espaces chimiques, l'ensemble du cahier des charges a pu être respecté. Nous présenterons dans cette section l'architecture du logiciel, et nous décrirons ensuite comment les molécules sont traitées et stockées dans la base de données. Nous présenterons enfin les principales fonctionnalités du logiciel.

4.1 Architecture

4.1.1 Langage

Nous avons conservé le langage JAVA utilisé pour le développement de SA1. Plusieurs raisons expliquent ce choix, les plus importantes étant la portabilité de celui-ci ainsi que l'existence d'un très grand nombre de bibliothèques tierces, et notamment de bibliothèques chémoinformatiques comme nous avons pu le voir.

4.1.2 Plateforme NetBeans

SA2 se présente sous la forme d'un logiciel dit bureautique (ou client lourd), par opposition à une application web utilisant le navigateur comme source d'interaction avec l'utilisateur et les protocoles de transferts standards type HTTP pour la transmission de données. A cet effet, on retrouve dans ce type de logiciel la plupart des composants graphiques formant ce type d'application : une barre de menu, des boutons, des fenêtres et sous-fenêtres, etc. Bien qu'essentielle au confort d'utilisation, l'architecture de base permettant la gestion de ces composants n'apporte en elle-même aucune valeur ajoutée à un travail de thèse n'y étant pas dédié. Malheureusement,

JAVA ne propose aucun outil natif permettant de gérer simplement une interface graphique. Pour la petite histoire, un premier prototype de SA2 avait été développé sans aucune aide extérieure pour la construction et la gestion du fenêtrage, des événements, des processus longs, etc. Constatant alors à quel point le développement de ces outils était long et fastidieux (et sans intérêt), nous avons décidé de dédier quelques jours à la veille technologique afin d'identifier des outils nous permettant de nous concentrer sur les aspects métiers de SA2, et non au développement d'interface graphique.

Il existe plusieurs outils de développement fournissant un environnement de base sur lequel les développeurs peuvent se reposer pour le développement d'outils bureautiques. A l'heure actuelle, les plus robustes et donc les plus utilisés dans la communauté JAVA sont les plateformes NetBeans (NBP) et Eclipse RCP (Rich Client Platform). Ces deux plateformes sont en fait liées aux environnements de développement NetBeans et Eclipse, respectivement, en ce sens qu'ils en forment le coeur. Ils sont connus pour être équivalents en termes de fonctionnalités, et la préférence pour l'un ou l'autre se résume souvent à des détails cosmétiques. De notre point de vue, étant donné que nous utilisons NetBeans pour les développements JAVA, le choix fut vite fait. Par ailleurs, deux constats supplémentaires ont fait pencher la balance vers NetBeans : (1) La plateforme Eclipse est basée sur SWT (Standard Widget Toolkit), une librairie graphique non-standard nécessitant donc du travail supplémentaire pour se former à son utilisation, et (2) les expériences utilisateurs tendent à affirmer que Eclipse RCP est beaucoup plus complexe d'utilisation. Avec le recul et en tenant compte de la difficulté de prise en main de la plateforme NetBeans, ce choix nous semble aujourd'hui d'autant plus justifié.

La plateforme NetBeans est une application formée par un ensemble de modules génériques fournissant la majorité des ressources nécessaires au développement des éléments structurant une application bureautique : gestion des menus, des fenêtres, configuration, gestion des mises à jour, etc. Le développement d'applications sur la base de la plateforme consiste alors au développement de modules qui apporteront des fonctionnalités supplémentaires à l'application en se basant sur les modules existants formant le coeur d'une application NetBeans. En quelques clics, il est donc possible dans NetBeans, de créer un projet RCP et d'obtenir ainsi un prototype standard d'application offrant une gestion native du fenêtrage, de la soumission et du monitoring de traitements longs, etc.

Du point de vue du programmeur, l'une des fonctionnalités intéressante de la plateforme NetBeans est la notion de *services*. Un service est une interface JAVA définissant le squelette

d'un ensemble de traitements cohérents. Par exemple, un service *FingerprintMetric* définirait une méthode *getSimilarity(Fingerprint1, Fingerprint2)* permettant de récupérer la similarité entre deux fingerprints, et une classe *TanimotoMetric* implémentant ce service fournirait une implémentation concrète de celui-ci. L'avantage procuré par cette notion est qu'il est ensuite possible de demander à NetBeans de retrouver automatiquement toutes les implémentations existantes pour un service donné. Il devient donc aisé d'ajouter des fonctionnalités à SA2 (une nouvelle métrique de comparaison de fingerprints par exemple), sans avoir à modifier l'interface graphique. Une fois le service développé, celui-ci sera automatiquement enregistré par la plateforme et sera donc disponible partout où il est utilisé.

4.1.3 Base de données

L'un des objectifs principaux du projet est le stockage de chimiothèques et *a posteriori* la consultation des données. A cet effet, les systèmes de gestion de base de données (SGBD) sont à l'heure actuelle les solutions les plus appropriées (et les plus robustes) pour gérer des gros volumes de données. Screening Assistant 2, dans la droite lignée de la version 1, repose sur le SGBD MySQL (version 5.3 ou supérieure). Contrairement à SA1, nous utilisons le moteur InnoDB, qui permet, entre autre, d'assurer l'intégrité référentielle des données et de gérer les transactions. Ces deux caractéristiques ont un impact négatif sur la performance (comparé au moteur MyISAM utilisé par SA1), mais le choix a été fait de privilégier l'intégrité et la cohérence des données. Concrètement, lorsqu'une molécule est supprimée, l'ensemble des données y étant associées sont également supprimées (ce qui n'est pas le cas avec le moteur MyISAM). De même, lors de l'insertion de données, les transactions permettent de s'assurer que l'ensemble des traitements ont été réalisés avec succès. Dans le cas contraire (plantage en plein milieu des traitements par exemple), elles garantissent que l'état de la base de données puisse être rétabli tel qu'il était avant le début du traitement.

La figure IV.3 donne un aperçu global de la base. Un total de 41 tables sont actuellement présentes dans SA2. Les détails seront fournis plus loin dans les sections appropriées et dans les annexes.

Notons d'ores et déjà que SA2 ne fait usage d'aucune cartouche moléculaire. A la date de début du projet, MyChem était la seule solution de gestion de base de données chimiques pour MySQL et était encore en phase de développement (la première version 0.7 étant sortie en septembre 2009). C'est d'ailleurs toujours le cas aujourd'hui pour ce SGBD, et l'outil ne

4.2 Interface graphique

La plateforme NetBeans offre une grande flexibilité dans la gestion du fenêtrage. Il est ainsi possible de déplacer les fenêtres simplement en les faisant glisser à l'endroit voulu, ou de les réduire sur l'un des bords de la fenêtre principale. L'application se souviendra également de la dernière disposition sauvegardée, qui sera donc rétablie lors du prochain lancement de SA2.

4.3 Gestion des molécules

La figure [IV.4](#) présente les tables principales permettant le stockage des informations de base associées aux molécules. Chaque molécule est stockée dans la table *Mol*, qui contient la table de connectivité de celle-ci, son SMILES, son InChiKey, ainsi que certains descripteurs calculés automatiquement lors de l'import de nouvelles molécules (la liste détaillée est fournie dans la table [A.1](#) en annexe. Directement reliées à cette table, on trouvera les informations concernant l'association des molécules aux fournisseurs et aux librairies, aux scaffolds et aux frameworks, ou encore aux coordonnées des molécules dans les différents espaces chimiques disponibles dans SA2 (toutes ces notions seront décrites plus bas).

4.3.1 Identifiants

Chaque molécule doit posséder un identifiant. Il n'est pas nécessaire de le rendre unique : une molécule pourra ainsi être associée à plusieurs identifiants si elle a été chargée plusieurs fois (par exemple si elle est retrouvée dans plusieurs bases de données fournisseur).

4.3.2 Format d'entrée

SA2 n'a pas pour vocation de gérer les nombreux formats de fichier existant en chémoinformatique. Afin de nous focaliser sur les aspects liés aux comparaisons de chimiothèques, le choix a été fait de ne supporter que deux formats standards : le format MDL MOLfile pour les molécules et le format texte CSV pour les propriétés, ces deux formats étant des incontournables dans le domaine.

Lors de l'import d'un fichier SDF (de même que pour un CSV), les propriétés associées à chaque molécule peuvent également être importées dans la base. La manière dont ces propriétés sont gérées et stockées sera décrite plus bas.

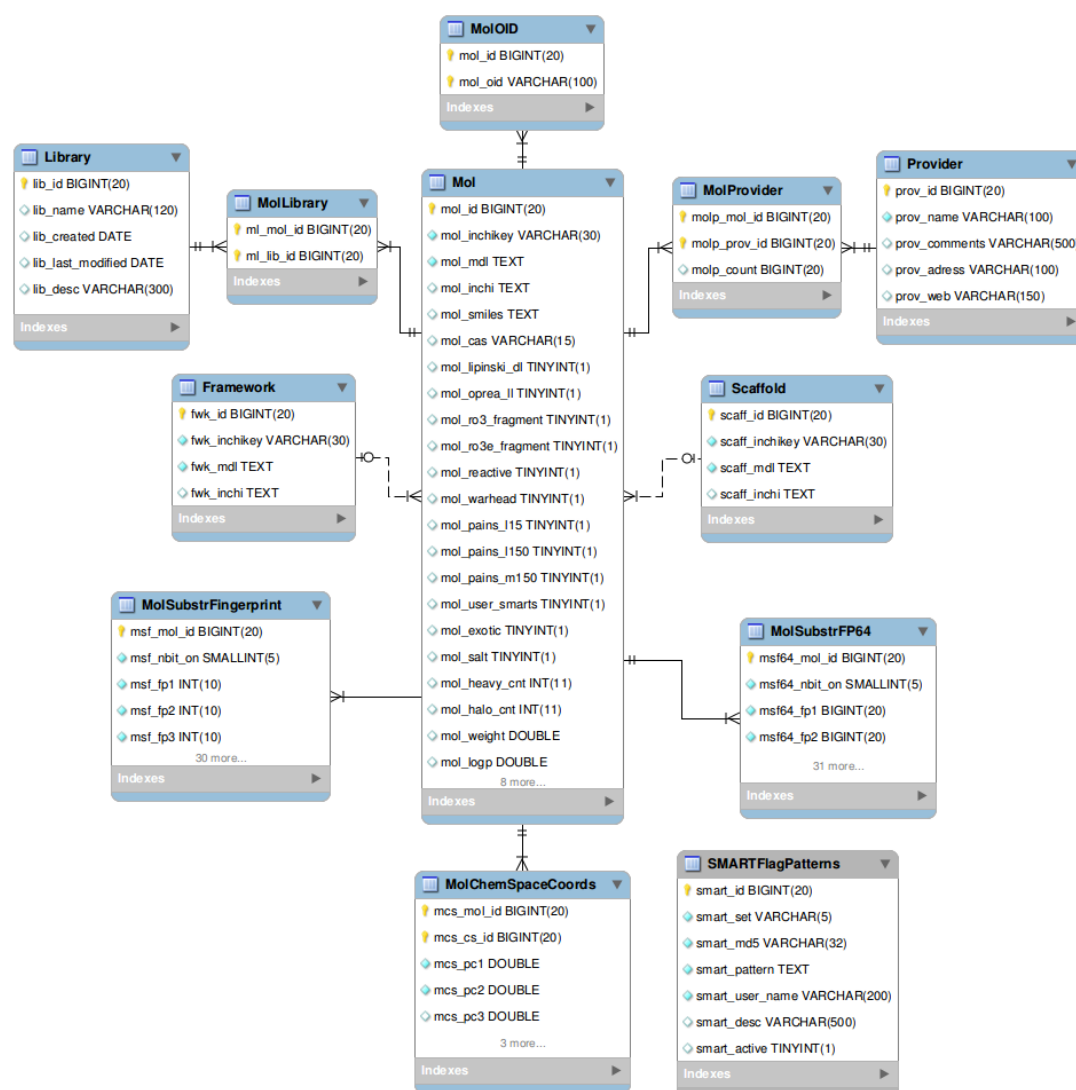


Figure IV.4 – Tables principales pour la gestion des molécules.

4.3.3 Standardisation et unicité

SA2 stocke les molécules en vérifiant au préalable leur existence dans la base de données. La notion de doublon, en apparence évidente, est en réalité bien plus complexe qu'il n'y paraît. D'un point de vue structural, il reste possible de se limiter à une comparaison stricte de deux molécules, en les considérant comme doublons si elles possèdent les mêmes atomes, les mêmes liaisons entre ces atomes et les même propriétés (charges, types de liaisons, stéréochimie). En revanche, du point de vue d'un chimiste ou d'un biologiste, d'autres facteurs entrent en compte, tels que la présence de contre ions, l'état de protonation des atomes ou les différentes formes

tautomères.

Afin de détecter les doublons, deux possibilités peuvent être envisagées : les SMILES canoniques générés par le CDK, Indigo ou JOELib et le code InChI. Le choix a été fait de conserver le code InChI (version 1.0.4 de l'outil), pour les raisons suivantes :

- Le SMILES canonique représentera l'ensemble de la structure telle qu'elle a été définie et ne pourra donc pas détecter les différentes formes tautomères des composés.
- A contrario, le code InChI est capable de gérer certaines formes de tautomérie.
- L'étude comparative effectuée par Aurélien Monge au cours de sa thèse a démontré que l'InChI était le plus performant pour détecter les doublons comparé à MOE, OEChem et Marvin.

Lorsqu'une molécule est importée, et contrairement à la version précédente de SA, aucune modification n'est apportée à celle-ci. Ce choix, tout a fait discutable, a été fait pour deux raisons principales :

1. Il n'existe aucune procédure standard et reconnue pour la standardisation de molécules. Ce processus est de plus généralement spécifique du problème à résoudre : certains logiciels de calcul attendent en entrée des molécules représentées sous leur forme aromatisées, tandis que d'autres ne tolèrent pas ce type de représentation (qui n'est d'ailleurs en théorie pas autorisée par le format SDF pour la définition de la structure d'une molécule). Nous avons ainsi souhaité n'introduire aucune modification (éventuellement indésirable du point de vue de l'utilisateur) aux molécules. De même nous ne modifions pas les sels, afin de ne pas introduire d'erreur lors de leur suppression. Pour ces derniers, les molécules ayant été détectées comme contenant deux sous-structures déconnectées seront marquées comme telles afin de conserver au moins partiellement cette information.
2. Du point de vue de l'unicité, et comme nous l'avons vu précédemment, le programme InChi effectue lui même ses propres normalisations, en ce sens qu'il est notamment capable de détecter certaines formes tautomères.

Un service a néanmoins été défini (*MolTransformer*) à cet effet. Il est donc possible pour un programmeur d'introduire un processus de normalisation en implémentant ce service, qui sera ensuite automatiquement appliqué avant l'insertion de toute molécule. Ce module s'adresse

néanmoins aux développeurs capables d'une part de programmer en JAVA, et d'autre part d'utiliser les bibliothèques chimioinformatiques fournies par SA2. D'autres approches sont aujourd'hui possibles et seront discutées dans la conclusion de ce chapitre.

4.4 Insertion des molécules

Une fois l'InChI calculé, et si la molécule n'est pas un doublon, un certain nombre d'opérations sera effectué sur chacune d'entre elles afin de stocker des informations de base. La figure IV.5 résume les traitements effectués pour chaque molécule lors de l'insertion d'un fichier SDF.

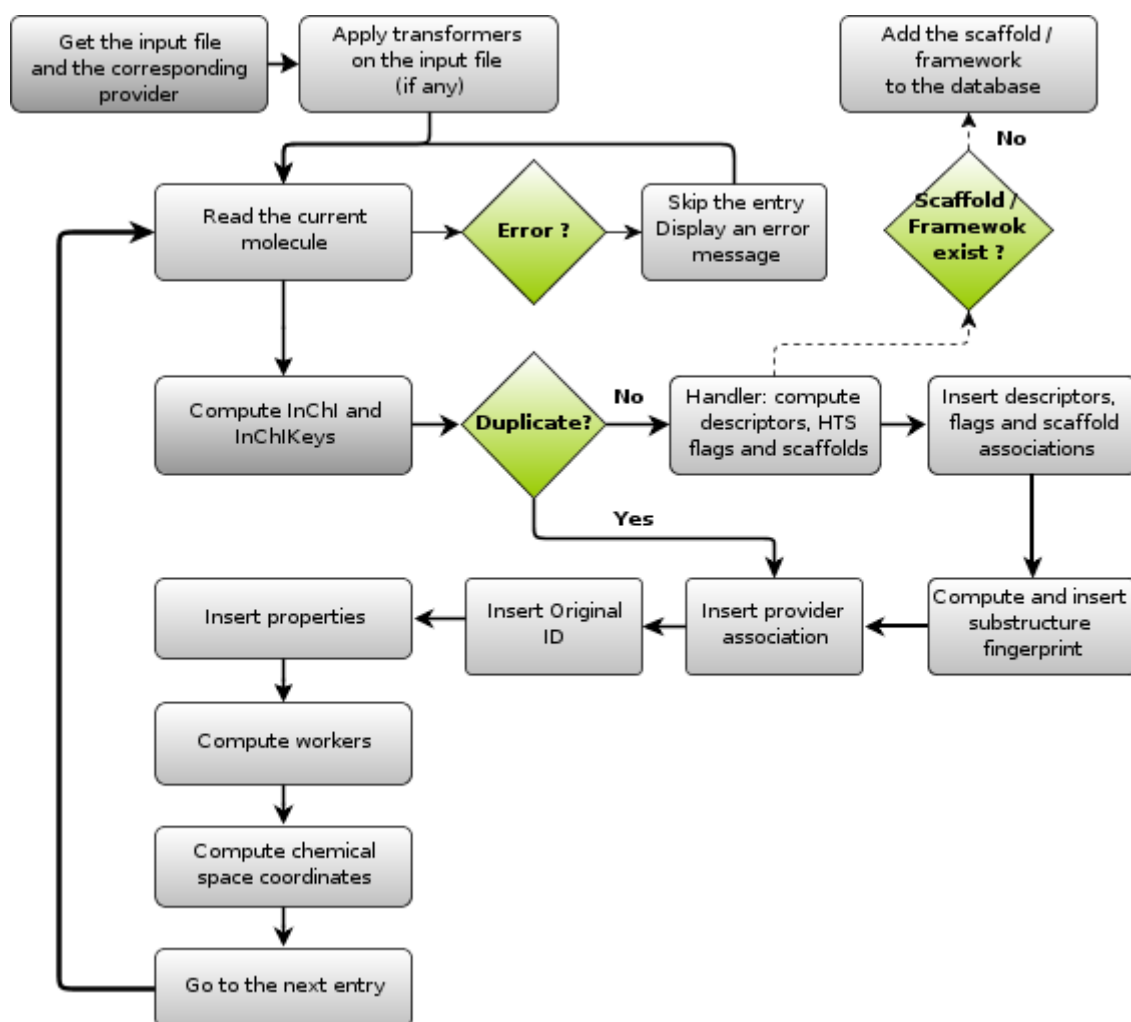


Figure IV.5 – Processus d'insertion de molécules dans SA2.

4.4.1 Perception des molécules

Afin de réaliser des traitements avancés sur chaque molécule (calculs de descripteurs, scaffolds, etc...), il devient nécessaire d'utiliser une librairie chémoinformatique pour obtenir une représentation manipulable des molécules. Dans un premier temps, nous avons sélectionné le CDK, pour la simple raison qu'il était le seul projet actif et supposé stable parmi les librairies JAVA, JOELib ayant été abandonné, et Indigo étant tout juste en version 0.9 beta. Nous avons néanmoins rencontré au fil des développements de nombreux problèmes avec cette librairie (bugs, types atomiques non supportés, valeurs de descripteurs abérantes, etc.).

Pour cette raison, nous avons introduit la notion de *Molecular Handler* sous la forme d'un service définissant les traitements à réaliser lors de l'insertion des molécules : chargement des atomes et des liaisons, calcul des descripteurs, calcul des squelettes moléculaires, marquage des molécules indésirables. Ainsi, on découple la définition des traitements à effectuer de leur implémentation. L'introduction de cette notion permet de ne pas se contraindre à l'utilisation d'une seule librairie et facilite le passage de l'une à l'autre en cas de besoin, ce qui a été le cas pour SA2 (nous avons tout d'abord utilisé le CDK, pour ensuite revenir à JOELib).

Deux implémentations sont donc disponibles pour la dernière version de SA2 : JOELib et CDK. JOELib est utilisé par défaut, car malgré le fait qu'il ne soit plus maintenu, nous avons constaté (entre 2008 et 2011) qu'il était beaucoup plus stable et plus rapide que le CDK.

4.4.2 Descripteurs

Les descripteurs listés dans l'annexe [A.1](#) sont calculés et insérés automatiquement dans la base lors de l'import d'une nouvelle molécule. Bien sûr, les valeurs obtenues pour notamment le LogP, ou le nombre d'accepteurs et de donneurs de liaisons hydrogènes dépendent du *Molecular Handler* utilisé, chacun d'entre eux implémentant ses propres modèles. Outre des propriétés de base telles que le poids moléculaire ou le LogP, certaines propriétés ont été calculées dans l'optique d'optimiser la recherche par sous-structure, comme décrit plus loin.

4.4.3 Marquage des molécules

Les molécules sont automatiquement marquées pour chacun des critères suivants :

- La molécule est perçue comme réactive. On utilise les 15 SMARTS définis par Rishton [\[30\]](#).

- La molécule est perçue comme warhead. On utilise les 20 SMARTS définis par Rishton [30].
- La molécule est perçue comme formant des agrégats (PAI). On utilise les SMARTS définis par Bael [11]
- La molécule contient plus d'un fragment connecté.
- La molécule contient des atomes non-reconnus par le *Molecular Handler*.

Ces marquages sont stockés dans la table principale et sont consultables dans une vue dédiée représentant les molécules sous forme de table et coloriant la colonne correspondant à la marque si celle-ci est présente. Ils peuvent également être utilisés comme critères pour la création de bibliothèques filtrées.

4.4.4 Squelettes moléculaires

La notion de squelette moléculaire est essentielle en chémoinformatique et en chimie médicinale. Elle a pour but d'obtenir une représentation simplifiée des molécules et représente un moyen simple et intuitif de les regrouper. Elle peut également être utilisée comme base de mesure de diversité, comme nous le décrirons plus loin. Lors de l'import d'une nouvelle molécule, deux types de squelettes seront calculés et associés à celle-ci :

Le scaffold

Il est défini comme étant l'ensemble des cycles et des atomes reliant ces cycles. Les chaînes latérales et les atomes terminaux (n'étant liés qu'à un seul autre atome) sont donc supprimés, à l'exception des doubles liaisons exocycliques et terminales.

Le framework

Il représente un niveau de simplification supplémentaire par rapport au scaffold. A partir du scaffold, les atomes sont transformés en carbones, et les doubles ou triples liaisons sont transformées en liaisons simples, à l'exception des liaisons aromatiques. Tout atome terminal est de plus supprimé. La conservation des liaisons aromatique nous permet de rendre un peu plus spécifique les frameworks calculés et on notera que cette règle n'avait pas été utilisée par l'étude de Bemis et Murcko [12].

La figure IV.6 donne un exemple concret. Les scaffolds et les frameworks sont stockés dans des tables dédiées (*Scaffold* et *Framework*). La table de connectivité au format MOL y est stockée et l'unicité est également gérée à l'aide du code InChI.

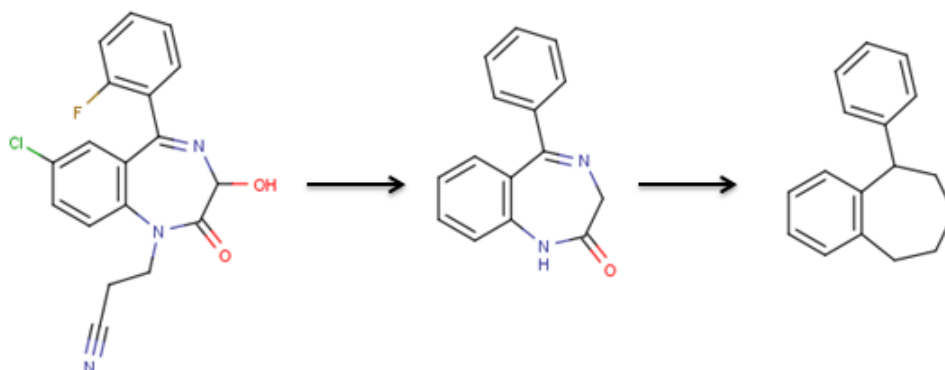


Figure IV.6 – Exemple de scaffold et de framework généré pour la molécule de Cinolazepam (Drugbank ID : DB01594).

4.4.5 Travailleurs

La notion de travailleurs (ou *Workers*) a été introduite afin de représenter des traitements additionnels effectués lors de l'insertion des molécules. Ils sont représentés par des classes JAVA implémentant un service nommé *MolWorker* qui, à partir d'une molécule, effectue un traitement quelconque. Une fois les opérations de base exécutées (calculs des descripteurs, des scaffolds, etc.), les travailleurs vont à leur tour effectuer les traitements pour lesquels ils ont été définis. Ils sont tous optionnels et peuvent être lancés, soit lors de l'import des molécules, soit plus tard une fois que l'ensemble des molécules a été chargé. SA2 propose actuellement quatre travailleurs :

- **CDK Descriptors** : calcule l'ensemble des descripteurs disponibles dans le CDK.
- **CDK Fingerprints** : calcule les fingerprints disponibles dans le CDK.
- **Indigo Fingerprints** : calcule le fingerprint implémenté dans Indigo et dédié à la recherche de similarité.
- **JOELib / SA1 Descriptors (JOELib)** : calcule l'ensemble des descripteurs et fingerprints disponibles dans SA1.

Bien que les travailleurs actuellement implémentés soient tous dédiés au calcul de descripteurs ou de fingerprints, il est parfaitement possible d'effectuer toute autre opération. Du point de vue du développeur, il s'agit d'implémenter le service *MolWorker* en créant un nouveau module dans le projet. Une fois fait, il apparaîtra automatiquement dans la liste des travailleurs disponibles à l'import ou dans l'interface de calcul exposée par l'application.

4.5 Gestion de sous-ensembles de molécules

La gestion de sous-ensembles de molécules nous a semblé indispensable pour un logiciel proposant de gérer et comparer des chimiothèques et d'en extraire des sous-ensembles d'intérêt. SA2 permet de gérer explicitement deux types de sous-ensembles :

Les fournisseurs

Les molécules étaient à l'origine (dans SA1) organisées par **fournisseurs** (*Providers*). Cette notion a pour but de représenter l'origine des molécules insérées dans SA, typiquement le nom d'un fournisseur commercial ou d'un laboratoire proposant des chimiothèques destinées au criblage. Toute molécule importée dans SA2 doit ainsi être associée à un fournisseur.

Molécules associées à un fournisseur
Molécules associées à un Framework
Molécules associées à un Scaffold
Molécules appartenant à d'autres librairies (union et intersection)
Molécules n'appartenant pas à une librairie
Molécules issues d'une recherche par sous-structure
Molécules issues d'une recherche par SMARTS
Molécules issues d'une recherche par similarité
Molécules issues d'une recherche par filtres
Molécules issues d'une sélection de sous-ensemble par diversité
Molécules sélectionnées dans l'une des vues de SA2 (espaces chimiques 2D, tables, etc.)

Tableau IV.3 – Liste des méthodes permettant de créer de nouvelles librairies dans SA2.

Les librairies

Une notion similaire a été ajoutée afin de faciliter la gestion de sous-ensembles générés dans le logiciel. Les molécules peuvent ainsi être regroupées sous forme de **librairies**. La notion de librairie est très similaire à la notion de fournisseur, en ce sens qu'elle permet également de grouper les molécules. Tandis que les fournisseurs utilisent la provenance des molécules, la notion de librairie se veut plus générique : il s'agit simplement d'offrir la possibilité de créer des sous-ensembles de molécules au sein de la base, quelle que soit la méthode utilisée. Il est

d'ailleurs possible de créer une librairie à partir d'un fournisseur. Le tableau [IV.3](#) liste les différentes manières de créer de nouvelles librairies. De même que pour les fournisseurs, nous verrons que la plupart des analyses disponibles dans SA2 peuvent se faire à la fois globalement pour l'ensemble de la base de données, mais également pour une ou plusieurs librairies.

Ces deux notions sont modélisées de manière identique dans la base. Les tables *MolProvider* et *MolLibrary* sont respectivement dédiées à l'association entre molécules et fournisseurs, et molécules et librairies. Les tables *Provider* et *Library* permettent de stocker les informations de base pour ces deux entités (identifiant, nom, description).

4.6 Gestion des descripteurs

Comme suggéré dans les sections précédentes, SA2 permet de stocker et d'associer à chaque molécule un nombre quelconque de descripteurs et de fingerprints. Offrir cette possibilité à travers une base de données n'est pas sans difficultés. Les SGBD sont des systèmes assez rigides qui sont parfaitement adaptés à des modèles de données peu sujets au changement. Ici, on souhaite laisser la liberté à l'utilisateur d'associer ses propres propriétés à chaque molécule. Cela signifie qu'un modèle rigide n'est pas une solution envisageable, puisqu'on ne peut deviner à l'avance ce que l'utilisateur souhaitera stocker dans la base.

Classiquement, le modèle *Entité-Attribut-Valeur* est utilisé pour répondre à ce type de problématique. L'idée est de stocker l'ensemble des données dans une seule table sous la forme d'un triplet associant l'entité (ici la molécule), l'attribut (ici ses propriétés) et la valeur associée. Ce modèle a pour mérite d'être simple à implémenter et surtout de permettre de répondre au problème tout en conservant un modèle stable. En revanche, il est beaucoup moins adapté lorsque des gros volumes de données sont en jeux. Ainsi par exemple, l'insertion de 1 million de molécules associées à 100 descripteurs implique l'insertion de 100 millions de lignes dans cette table. De plus, récupérer un sous-ensemble de descripteurs associé à une molécule nécessite certaines opérations de pivots qui impactent de manière dramatique les performances.

Pour ces raisons, nous avons choisi de stocker les descripteurs dans de nouvelles tables qui sont créées dynamiquement, soit au moment de l'import de molécules ou de descripteurs, soit en utilisant une interface dédiée. Notons également que chaque table peut être assignée à une catégorie. Cela permet de regrouper des tables suivant le type de descripteur y étant stocké

(2D, 3D, etc.), ou encore l'outil utilisé pour réaliser le calcul. Les données peuvent ainsi être organisées de la manière suivante :

- Une table représentera un groupe logique de propriétés. Par exemple, une table MOE2D est créée automatiquement lors de la création d'une nouvelle base SA2, et sera dédiée au stockage des descripteurs MOE2D.
- Une propriété sera modélisée sous la forme d'une colonne associée à une table. Quatre types de propriétés sont actuellement supportés : entiers, flottants, booléens et texte.
- Une catégorie ou sous-catégorie représentera un groupe logique de tables. Par exemple, un groupe MOE est automatiquement créé, dans lequel les tables MOE2D et MOE3D sont référencées.

Le même principe s'applique pour les fingerprints, à ceci près qu'une table ne peut en stocker qu'un seul. Le stockage de fingerprints en base de données n'est par ailleurs pas immédiatement réalisable et pose généralement problème. Il existe de nombreux moyens de stocker un fingerprint (et à notre connaissance aucun format de fichier standard, bien qu'Andrew Dalke ait proposé il y a peu un nouveau format d'échange [10]) mais peu d'entre eux sont adaptés à une représentation en base de données. Afin d'obtenir un bon compromis entre performance et stockage, nous avons tout d'abord choisi de modéliser les fingerprints sous la forme d'un nombre fixe de bits. Pour les fingerprints structuraux classiques, cette représentation est bien adaptée puisque le nombre de bits du fingerprint est en général fixe. Pour les fingerprints hachés en revanche, il n'existe aucune limite sur le nombre de bits pouvant être retrouvés. SA2 ne gère pas ce type de fingerprint et il appartient donc à l'utilisateur de le transformer en un fingerprint de taille fixe.

Chaque fingerprint sera représenté par un nombre limité d'entiers codés sur 32 bits. Chaque entier est donc interprété sous sa forme binaire pour permettre la conversion vers une liste de bits (et vice versa). La figure IV.7 illustre le principe. L'utilisation d'entiers permet un stockage plus compact des données et évite la création d'une colonne par bit (à la place on crée une colonne pour 32 bits). Ceci est d'autant plus important que MySQL et le moteur InnoDB posent une limite de 1000 colonnes par table.

4.7.1 Optimisation de la recherche par sous-structure

La recherche de sous-structures implique l'utilisation d'algorithmes avancés de recherche de sous-graphes. La recherche d'isomorphisme de graphes étant un problème NP-complets, les implémentations existantes sont en général gourmandes en temps de calcul. A titre d'exemple il faut environ 35 secondes à Indigo (le plus rapide comparé au CDK et JOELib) pour rechercher l'azulène parmi 10 000 molécules prises au hasard dans notre base de 7 millions de molécules. Pourtant, Indigo utilise une version modifiée de l'algorithme VF2 de Cordella et al. [13]. Cet algorithme de recherche de sous-graphe est probablement ce qui se fait de mieux, comme démontré récemment par Ehrlich et al. [14]. Il devient donc nécessaire de trouver un moyen d'accélérer la recherche par sous-structure afin d'obtenir les résultats plus rapidement. Notre approche est classique et simple : nous commençons par pré-filtrer les molécules à tester suivant les critères d'optimisation décrits ci-après, puis nous appliquerons l'algorithme lui-même sur les molécules restantes. Nous avons considéré trois critères d'optimisation fonctionnant tous sur le même principe, à savoir qu'il est inutile de tester une molécule ne possédant pas une propriété retrouvée dans la sous-structure requête :

Optimisation 1 :

Si une sous-structure requête correspond à un scaffold stocké dans la base de données, on peut d'ores et déjà considérer l'ensemble des molécules associées à ce scaffold comme étant des résultats valides de la recherche. Il n'est donc pas nécessaire de tester ces molécules, qui seront directement retournées de manière quasiment instantanée (suivant le nombre de molécules en jeu bien entendu).

Optimisation 2 :

Lorsque l'on recherche une sous-structure contenant N_{query} atomes, il n'est pas nécessaire de tester les molécules ayant $N_{target} < N_{query}$ atomes lourds. Nous avons appliqué ce principe en utilisant le nombre d'atomes lourds et le nombre d'halogènes. Un index a donc été créé sur ces colonnes afin d'accélérer cette étape.

Optimisation 3 :

Suivant le même principe que la première optimisation décrite précédemment, il n'est pas nécessaire de tester des composés ne contenant pas des fragments présents dans la sous-structure à rechercher. A cet effet, Indigo propose un fingerprint ayant été optimisé pour la recherche de sous-structure. Ce fingerprint est notamment utilisé dans la cartouche moléculaire Bingo. Aucune description officielle n'est néanmoins disponible. Lors de l'import de nouvelles molécules,

ce fingerprint sera calculé et inséré dans une table dédiée suivant le même format que les fingerprints classiques stockés dans SA2. Lors de la recherche, un nouveau fingerprint sera calculé comme étant issu d’une opération de ET logique entre le fingerprint de la sous-structure requête et celui des molécules en base. Seules les molécules induisant alors un fingerprint identique à celui de la sous-structure requête seront ensuite considérées pour la recherche de sous-structure. Autrement dit, on s’assure que les molécules possèdent au moins l’ensemble des fragments (tels que définis par le fingerprint d’Indigo) détectés dans la sous-structure requête.

Au final, une fois les optimisations mises en place, la recherche de l’azulène décrite précédemment devient instantanée. Sur de plus grosses bases de données (plusieurs centaines de milliers de molécules), les recherches prennent en moyenne entre 2 et 10 secondes suivant la spécificité de la recherche et peuvent aller jusqu’à plusieurs minutes pour des recherches très peu spécifiques. Il est évident que rechercher un Benzène au sein d’une base de molécule prendra beaucoup plus de temps puisque ce type de fragment est peu discriminatoire que l’on retrouve dans un grand nombre de molécules.

4.8 Espaces chimiques

La notion d’espace chimique est essentielle dans l’optique de comparer plusieurs chimiothèques et d’évaluer leur diversité. SA2 permet une gestion flexible d’espaces chimiques, en proposant à la fois d’en créer mais également de projeter les molécules dans un graphique interactif. Nous avons également introduit la méthodologie DRCS dans le logiciel afin de faciliter son utilisation.

4.8.1 Création d’espaces chimiques

L’analyse en composantes principales est utilisée afin de créer de nouveaux espaces chimiques à partir de l’ensemble de la base de molécules ou d’un sous-ensemble (librairies). Le choix des descripteurs est effectué par l’utilisateur, qui peut sélectionner n’importe quel descripteur parmi ceux présents dans la base. Il est bien entendu nécessaire que des valeurs soient associées à chaque descripteur et ce pour chaque molécule. Dans le cas contraire, toute molécule ayant une valeur manquante ou invalide sera ignorée dans le calcul. De même, tout descripteur ayant une variance égale à 0 sera ignoré. Par défaut, les descripteurs sont centrés et réduits afin de s’assurer qu’ils seront tous sur la même échelle, mais cette opération peut être désactivée si nécessaire. L’espace chimique ainsi généré sera alors sauvegardé dans la base de données dans des tables dédiées (voir figure [A.1](#)) et seules les 6 premières composantes seront sauvegardées.

Optionnellement, il est également possible de calculer l'enveloppe convexe de l'espace ainsi créé en utilisant la méthodologie des DRCS décrite dans les chapitres 2 et 3. Il est ensuite possible de calculer d'autres enveloppes convexes une fois l'espace créé, sur l'ensemble de la base ou sur les molécules associées à une librairie.

4.8.2 Visualisation d'espaces chimiques

Les molécules peuvent être projetées dans tout espace chimique créé dans SA2. Le module de visualisation permet de visionner les molécules sous la forme d'un nuage de points représenté en deux dimensions. Il est possible de sélectionner les composantes principales à utiliser pour le tracé. Un soin particulier a été apporté quant aux interactions possibles avec l'utilisateur. Les molécules peuvent ainsi être sélectionnées de manière interactive, cette sélection pouvant être ensuite réutilisée pour, par exemple, créer une nouvelle librairie. Afin de faciliter la comparaison de plusieurs librairies, il est également possible, soit de limiter les molécules représentées à une ou plusieurs librairies, soit de colorier ces molécules suivant leur appartenance à une ou plusieurs librairies ou fournisseurs. Il est également possible de colorier suivant la valeur d'un descripteur choisi au sein de la base.

A noter qu'il existe également une fenêtre permettant de réaliser les mêmes opérations pour cette fois tracer plus simplement une propriété en fonction d'une autre.

4.8.3 DRCS

Outre l'introduction de la méthodologie DRCS dans SA2, nous avons également ajouté les trois espaces chimiques de référence décrits dans le chapitre 3. Ces trois espaces chimiques sont basés sur des descripteurs MOE2D, MOE3D et CDK2D et ont été calculés sur notre base de produits contenant plus de 6 millions de molécules. Ils sont insérés automatiquement lors de la création d'une nouvelle base de données SA2 et les enveloppes convexes délimitant différents types de sous-espaces (HTS, "drug-like", "lead-like", etc.) sont également insérées.

4.9 Comparaison de chimiothèques et mesures de diversité

La diversité, comme discuté dans l'introduction de cette thèse, n'est pas une notion absolue et il n'existe ainsi aucune métrique unique permettant de statuer définitivement sur la diversité relative et absolue de plusieurs chimiothèques. Pour cette raison, nous avons souhaité introduire

différentes manières de comparer deux chimiothèques et d'évaluer leur diversité.

La projection de chimiothèques dans un espace chimique 2D (qu'il soit basé sur une ACP ou sur deux propriétés) donne une bonne indication visuelle du recouvrement de deux chimiothèques. De manière complémentaire, nous avons ajouté d'autres mesures permettant de se placer d'un point de vue plus global pour évaluer la diversité moléculaire.

4.9.1 Mesures basées sur les scaffolds

Les mesures décrites ci-après restent bien entendu subjectives, puisqu'elles dépendent de la définition d'un scaffold utilisée par SA2, mais peuvent déjà donner une bonne indication sur la diversité (ou plutôt sur la non-diversité) d'une chimiothèque donnée. L'ensemble de ces mesures peut être calculé en se basant sur les scaffolds ou sur les frameworks.

Nombre et proportion de scaffolds : calcule le nombre et la proportion de scaffolds pour chaque sous-ensemble disponible (librairies ou fournisseurs). La proportion de scaffolds est définie ici comme étant le rapport entre le nombre de scaffolds et le nombre de molécules associées au sous-ensemble. Une telle métrique permet d'obtenir une première idée de la diversité interne de chaque sous-ensemble, une faible proportion de scaffolds suggérant une faible diversité.

Proportion cumulative de scaffolds : on la définit comme la proportion de scaffolds nécessaire pour représenter 50 ou 80% de molécules. A nouveau, une faible proportion suggère également une faible diversité. Cette métrique a été utilisée dans de nombreuses études évaluant la diversité de scaffolds dans des ensembles de molécules [20, 23].

Nombre de scaffolds singletons : le nombre et la proportion de scaffolds singletons, c'est à dire de scaffold associé à une seule molécule. Un nombre élevé suggère une bonne diversité en termes de scaffolds. On note néanmoins que certains expérimentateurs souhaitent, pour chaque molécule, ajouter un ou plusieurs analogues afin de confirmer une éventuelle activité, auquel cas cette mesure doit être analysée en tenant compte de ce fait.

Unicité de scaffolds : calcule le nombre et la proportion de scaffolds appartenant exclusivement à chaque sous-ensemble (librairies ou fournisseurs) défini dans la base. La proportion est ici calculée comme étant le rapport entre le nombre de scaffolds exclusifs et le nombre de scaffolds associés au sous-ensemble. Cette mesure est donc relative. Elle permet de quantifier l'originalité de chaque sous-ensemble en termes de scaffolds.

Ces mesures sont toutes complémentaires. Ainsi, un nombre de scaffolds singletons égal à 0 n'est pas nécessairement synonyme d'un manque de diversité, comme indiqué précédemment. En se rapportant aux autres mesures, et notamment à la proportion de scaffolds, on peut alors obtenir une information supplémentaire qui permettra de mieux appréhender les résultats obtenus pour les autres métriques. De même, le calcul de l'unicité des scaffolds permet lui, d'évaluer l'originalité de chaque librairie par rapport aux autres.

4.9.2 Mesures basées sur les fingerprints

Les mesures basées sur les scaffolds présentent l'avantage d'être simples à interpréter et donnent déjà une première information importante sur la composition d'une chimiothèque. Elle atteignent néanmoins leur limite lorsque les scaffolds sont tous différents bien que très similaires entre eux, même si dans de tels cas les mesures équivalentes calculées sur les frameworks permettront d'identifier ce type de biais. De plus, elles ne donnent pas d'information sur la similarité qui peut exister entre les molécules elles-mêmes (deux molécules peuvent avoir un scaffold identique mais être très différentes). Nous avons ainsi introduit des mesures classiques et complémentaires de celles décrites précédemment.

Similarité au plus proche voisin : moyenne des similarités au plus proche voisin, ainsi que la valeur minimale et maximale observées.

Similarité moyenne : similarité moyenne entre chaque paire de molécules, ainsi que la similarité moyenne minimale et maximale observées.

Histogramme : L'histogramme de ces deux métriques est également tracé afin de donner une vue plus détaillée de la distribution de ces deux mesures.

Ces métriques peuvent être calculées sur une librairie en utilisant un des fingerprints disponibles dans la base. Par défaut, le coefficient de Tanimoto est utilisé pour comparer les fingerprints.

4.9.3 Histogrammes

Les librairies peuvent également être comparées de manière plus simple, en traçant simplement l'histogramme d'une propriété disponible dans SA2. Il est ainsi possible de tracer sur

la même figure, les histogrammes de plusieurs librairies pour une propriété donnée. L'histogramme en lui-même est paramétrable (nombre de bins, valeurs minimale et maximale) et sera automatiquement tracé lorsqu'une propriété est sélectionnée dans la fenêtre faisant l'inventaire des tables et des propriétés disponibles. Ce type de représentation permet d'identifier rapidement des différences de distributions sur des propriétés interprétables (poids moléculaire, logP, nombre d'accepteurs d'hydrogènes, etc.).

4.10 Génération de sous-ensembles divers

Nous avons conservé l'algorithme utilisé dans Screening Assistant 1 pour la génération de sous-ensembles divers et nous y avons apporté une amélioration supplémentaire (optionnelle). On utilise donc une approche basée sur l'algorithme Stochastic Clustering Analysis (SCA) [28] en utilisant les scaffolds (ou les frameworks) pour identifier les groupes de molécules. La méthode fonctionne comme suit pour la génération d'une librairie diverse LD_n contenant n molécules :

1. On récupère tout d'abord l'ensemble des scaffolds, soit de manière aléatoire, soit en les ordonnant par ordre décroissant du nombre de molécules associées. Cette dernière méthode permet d'introduire un léger biais si l'on souhaite obtenir un échantillonnage un minimum représentatif de la composition de la base en assurant que les scaffolds les plus peuplés seront bien représentés dans LD_n .
2. Une première molécule est ajoutée à LD_n à partir du premier groupe de molécules représentées par le premier scaffold de la liste générée à la première étape. Cette molécule est sélectionnée comme étant la molécule la plus similaire à un fingerprint moyen calculé sur le groupe de molécules sélectionné.
3. Tant que le nombre de molécules désiré n'a pas été atteint, la molécule la plus dissimilaire aux molécules de LD_n est récupérée. Cette molécule sera choisie parmi les molécules associées au scaffold courant.
 - (a) Si aucune valeur maximale de similarité n'a été définie, on ajoute la molécule à LD_n .
 - (b) Si une valeur maximale de similarité a été définie, on vérifie que la similarité minimale observée ne dépasse pas cette valeur. Il est en effet possible que la valeur minimale de similarité à une étape donnée soit de 1, ce qui pénaliserait la diversité de la librairie

générée. En définissant une valeur maximale, on évite ainsi d'ajouter des molécules ayant un scaffold différent mais étant malgré tout très similaires à au moins une molécule de LD_n . Si la valeur maximale de similarité est atteinte pour la molécule courante, on ignore la molécule et on passe au scaffold suivant.

4. L'algorithme s'arrête lorsque $\text{card}(LD_n) = n$.

Notons qu'il est possible que n ne soit pas atteint si la valeur maximale de similarité est définie et fixée (si aucune molécule n'est ajoutée après avoir testé l'ensemble des scaffolds). Dans un tel cas, sa valeur est incrémentée de 0,05 et l'algorithme est relancé.

Cet algorithme possède deux points forts principaux : (1) il est rapide et peut donc être utilisé sur des grosses bases de données et (2) il est aisément interprétable, le premier critère de sélection étant l'appartenance d'une molécule à un scaffold. Bien sûr, l'utilisation d'une valeur maximale de similarité impacte de manière plus ou moins significative la vitesse d'exécution de l'algorithme. Mais il permet également de minimiser la redondance au sein de la librairie diverse, un critère qui nous semble être essentiel dans ce contexte.

5 Discussions et perspectives

Nous avons présenté dans cette section le logiciel Screening Assistant 2. En successeur logique de la première version du logiciel développé au sein du laboratoire, cette nouvelle mouture ajoute de nombreuses fonctionnalités ayant pour objectif de rendre l'outil plus flexible et plus polyvalent. L'utilisation d'un framework de programmation tel que la plateforme NetBeans offre par ailleurs un confort d'utilisation accru à la fois en termes d'expérience utilisateur, mais également du point de vue du développeur.

Du point de vue de l'analyse de chimiothèques, un grand nombre de méthodes ont été introduites, à la fois pour gérer et créer de nouveaux sous-ensembles de molécules, mais également pour les analyser et les comparer. Les méthodes décrites dans ce chapitre nous semblent être toutes complémentaires et permettent ainsi d'obtenir une vision certes subjective mais globalement pertinente sur le contenu d'une ou plusieurs chimiothèques. L'ajout de recherches par sous-structure, similarité, ou SMARTS, la possibilité de stocker et d'organiser propriétés et fingerprints, ainsi que les diverses informations pouvant être affichées dans SA2 (tables 2D, scaffolds et frameworks) simplifient également certaines opérations plus routinières mais très

utiles dans le quotidien du chimoinformaticien.

Bien entendu, le logiciel reste très largement perfectible et de nombreuses améliorations peuvent être envisagées, tant du point de vue technique (architecture logicielle, optimisation du code, documentation) que du point de vue fonctionnel. Nous ne listerons pas tout ici, mais nous décrirons néanmoins les 4 axes de développement qui semblent les plus intéressants :

Normalisation

Comme indiqué dans ce chapitre, SA2 n'opère aucune transformation sur les molécules en entrée. Ce choix a été justifié précédemment, mais reste discutable pour plusieurs raisons :

- Il existe très peu d'outils libres permettant de normaliser un ensemble de molécules. Indigo et RDKit proposent des méthodes permettant de réaliser l'essentiel des opérations, mais nécessitent des compétences importantes en programmation, ce qui restreint leur utilisation. A notre connaissance, seul Chemaxon propose un outil complet de normalisation gratuit pour les laboratoires académiques.
- Nos espaces DRCS, disponibles dans SA2, ont été créés à partir de molécules standardisées dans Pipeline Pilot. Bien que nous ayons mis à disposition la procédure détaillée ainsi que le protocole lors de la publication de l'article, l'utilisateur doit tout de même, soit posséder Pipeline Pilot, la version gratuite pour les académiques n'étant plus disponible depuis 2011, soit réimplémenter l'ensemble de nos étapes à l'aide d'un autre outil.

Il serait donc intéressant de proposer en option la possibilité d'appliquer des protocoles de normalisation lors de l'import des molécules. Ces protocoles seraient, soit déjà définis dans SA2 (DRCS), soit définis par l'utilisateur sous la forme d'options (kékulisation, suppression des hydrogènes, neutralisation) et de fichiers SMIRKS. Lors du développement de SA2, les bibliothèques JAVA de chimoinformatique possédaient peu de fonctionnalités pouvant être utilisées à cet effet. Aujourd'hui, Indigo semble être la meilleure option pour ce type de traitement. La bibliothèque supporte en effet un certain nombre de fonctionnalités utiles (et non supportées par le CDK) et notamment la possibilité de transformer les molécules sous forme aromatique vers leur forme Kékulé. Plus important encore, le support des SMIRKS a été récemment ajouté à Indigo, ce qui offre à l'utilisateur averti une grande flexibilité pour la normalisation des molécules. Avec le recul, Indigo nous paraît être le choix le plus adapté pour la plus grande partie des traitements réalisés dans SA2.

L'utilisation d'une cartouche moléculaire

Bien que les performances soient relativement acceptables pour les différents types de recherches disponibles dans SA2, il serait tout de même intéressant d'utiliser une cartouche moléculaire afin, d'une part évaluer les différences de performances ainsi obtenues par rapport à la version actuelle de SA2 et d'autre part, déléguer le gros des traitements au serveur. MyChem étant la seule cartouche disponible pour SA2, et cet outil étant toujours en développement, peu de solutions s'offrent à nous pour ce SGBD. La cartouche Bingo de GGASoftware semble être de notre point de vue l'alternative la plus crédible, mais n'est disponible que pour Oracle et PostgreSQL, et bien que de nombreux efforts aient été dédiés à la modularité du code de SA2, un changement de SGBD ne se fait jamais sans douleur.

Espaces chimiques et méthodes de réduction de dimensions

SA2 ne propose que l'ACP pour la création et la visualisation d'espaces chimique. Bien qu'elle soit considérée comme indispensable dans le domaine, elle possède comme n'importe quelle autre ses avantages et ses inconvénients. Il serait donc intéressant de proposer d'autres moyens de générer des espaces chimiques. Les cartes de Kohonen ou les GTM par exemple, de part leur non-linéarité, peuvent apporter une information complémentaire de celle fournie par l'ACP, bien que leur paramétrisation soit particulièrement délicate.

La gestion des plaques et des résultats de criblage

De ce point de vue, il serait intéressant de proposer une gestion complète de plaques de criblage, afin d'augmenter le domaine d'applicabilité de SA2. La définition des plaques, le positionnement des produits dans celles-ci, la gestion des concentrations, des mélanges, des contrôles sont les étapes qui suivent logiquement la sélection des molécules à cribler et il existe peu d'outils libres permettant de réaliser ces opérations. Si l'on souhaite aller plus loin, il serait intéressant de se rapprocher encore plus de la partie expérimentale en permettant une gestion de résultats de criblage plus évoluée que le simple stockage de données dans des tables. Récemment, Visser et al. [34] ont décrit une ontologie dédiée au domaine du criblage expérimental. L'utilisation d'un vocabulaire contrôlé permet notamment de standardiser les processus de traitement des données tout en minimisant les risques d'erreurs de saisie.

Nous pouvons également ajouter que le développement du logiciel s'est fait en fil rouge tout au long de cette thèse. Le projet SA2 contient 608 classes JAVA, pour un total de 83468 lignes de code et 37299 lignes de commentaires. Il a été téléchargé 1004 fois (au 6 août 2013) pour la dernière version, 1445 fois au total depuis la première version publique (juillet 2011). Il est

téléchargeable à l'adresse suivante : [http ://sa2.sourceforge.net/](http://sa2.sourceforge.net/)

Références bibliographiques

- [1] The apache software fundation. <http://www.apache.org/>. 199
- [2] The bingo database cartridge. <http://ggasoftware.com/opensource/bingo>. 201
- [3] The indigo toolkit, gga software services. <http://ggasoftware.com/opensource/indigo>. 200
- [4] Joelib, a computational chemistry java library. <http://joelib.sourceforge.net/>. 199, 200
- [5] Mx - essential cheminformatics. <https://code.google.com/p/mx-java/>. 199
- [6] The mychem database cartridge. <http://mychem.sourceforge.net/>. 201
- [7] An open standard for chemical structure representation - the iupac chemical identifier. <http://stage.iupac.org/inchi/Stein-2003-ref1.html>. 200
- [8] The pgchem : :tigress database cartridge. <http://pgfoundry.org/projects/pgchem/>. 201
- [9] Rdkit : Open-source cheminformatics. <http://www.rdkit.org>. 200, 201
- [10] Spécifications du format fps créé par andrew dalkes. <https://code.google.com/p/chem-fingerprints/wiki/FPS>. 218
- [11] J. B. Baell and G. A. Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7) :2719–2740, 2010. 214
- [12] G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15) :2887–2893, 1996. 214
- [13] L. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10) :1367–1372, 2004. 220
- [14] H.-C. Ehrlich and M. Rarey. Systematic benchmark of substructure search in molecular graphs - from ullmann to VF2. *Journal of Cheminformatics*, 4(1) :13, 2012. 220

Références bibliographiques

- [15] P. Ertl, A. Schuffenhauer, and S. Renner. The scaffold tree : an efficient navigation in the scaffold universe. *Methods in molecular biology (Clifton, N.J.)*, 672 :245–260, 2011. [203](#)
- [16] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E. L. Willighagen. The blue obelisk-interoperability in chemical informatics. *Journal of chemical information and modeling*, 46(3) :991–998, 2006. [199](#)
- [17] S. Heller and A. McNaught. The status of the InChI project and the InChI trust. *Journal of Cheminformatics*, 2(Suppl 1) :P2, 2010. [200](#)
- [18] D. C. Ince, L. Hatton, and J. Graham-Cumming. The case for open computer programs. *Nature*, 482(7386) :485–488, Feb. 2012. [198](#)
- [19] N. Jeliaskova and V. Jeliaskov. AMBIT RESTful web services : an implementation of the Open-Tox application programming interface. *Journal of Cheminformatics*, 3(1) :18, 2011. [202](#)
- [20] M. Krier, G. Bret, and D. Rognan. Assessing the scaffold diversity of screening libraries. *Journal of chemical information and modeling*, 46(2) :512–524, 2006. [223](#)
- [21] T. Kuhn, E. L. Willighagen, A. Zielesny, and C. Steinbeck. CDK-Taverna : an open workflow environment for cheminformatics. *BMC Bioinformatics*, 11(1) :159, 2010. [202](#)
- [22] G. A. Landrum and N. Stiefl. Is that a scientific publication or an advertisement ? reproducibility, source code and data in the computational chemistry literature. *Future medicinal chemistry*, 4(15) :1885–1887, 2012. [198](#)
- [23] S. R. Langdon, N. Brown, and J. Blagg. Scaffold diversity of exemplified medicinal chemistry space. *Journal of chemical information and modeling*, 51(9) :2174–2185, 2011. [223](#)
- [24] A. Lanzén and T. Oinn. The taverna interaction service : enabling manual interaction in workflows. *Bioinformatics*, 24(8) :1118–1120, 2008. [202](#)
- [25] A. Monge. Création et utilisation de chimiothèques optimisées pour la recherche in silico de nouveaux composés bioactifs, 2006. [195](#), [196](#), [205](#)
- [26] N. M. O’Boyle, R. Guha, E. L. Willighagen, S. E. Adams, J. Alvarsson, J.-C. Bradley, I. V. Filippov, R. M. Hanson, M. D. Hanwell, G. R. Hutchison, C. A. James, N. Jeliaskova, A. S. Lang, K. M. Langner, D. C. Lonie, D. M. Lowe, J. Pansanel, D. Pavlov, O. Spjuth, C. Steinbeck, A. L. Tenderholt, K. J. Theisen, and P. Murray-Rust. Open data, open source and open standards in chemistry : The blue obelisk five years on. *Journal of Cheminformatics*, 3 :37, 2011. [199](#), [200](#)
- [27] N. M. O’Boyle and G. R. Hutchison. Cinfony - combining open source cheminformatics toolkits behind a common interface. *Chemistry Central Journal*, 2(1) :24, Dec. 2008. [201](#)
- [28] C. H. Reynolds, R. Druker, and L. B. Pfahler. Lead discovery using stockastic cluster analysis : A new method for clustering structurally similar compounds. *Journal of Chemical Information and Computer Sciences*, 38(2) :305–312, 1998. [225](#)
- [29] M. Rijnbeek and C. Steinbeck. OrChem - an open source chemistry search engine for oracle®. *Journal of Cheminformatics*, 1(1) :17, 2009. [202](#)

- [30] G. M. Rishton. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today*, 8(2) :86–96, 2003. [213](#), [214](#)
- [31] O. Spjuth, J. Alvarsson, A. Berg, M. Eklund, S. Kuhn, C. Masak, G. Torrance, J. Wagener, E. Willighagen, C. Steinbeck, and J. Wikberg. Bioclipse 2 : A scriptable integration platform for the life sciences. *BMC Bioinformatics*, 10(1) :397, 2009. [202](#)
- [32] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (CDK) : an open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2) :493–500, 2003. [200](#)
- [33] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Current Pharmaceutical Design*, 12(17) :2111–2120, 2006. [200](#)
- [34] U. Visser, S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon, and S. C. SchÄ¼rer. BioAssay ontology (BAO) : a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, 12(1) :257, 2011. [228](#)
- [35] W. P. Walters. Modeling, informatics, and the quest for reproducibility. *Journal of Chemical Information and Modeling*, 53(7) :1529–1530, July 2013. [198](#), [199](#)
- [36] S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel, and H. Waldmann. Interactive exploration of chemical space with scaffold hunter. *Nature chemical biology*, 5(8) :581–583, 2009. [202](#)

6 Mining collections of compounds with Screening Assistant 2

SOFTWARE

Open Access

Mining collections of compounds with Screening Assistant 2

Vincent Le Guilloux^{1*}, Alban Arrault², Lionel Colliandre¹, Stéphane Bourg³,
Philippe Vayer² and Luc Morin-Allory^{1*}

Abstract

Background: High-throughput screening assays have become the starting point of many drug discovery programs for large pharmaceutical companies as well as academic organisations. Despite the increasing throughput of screening technologies, the almost infinite chemical space remains out of reach, calling for tools dedicated to the analysis and selection of the compound collections intended to be screened.

Results: We present Screening Assistant 2 (SA2), an open-source JAVA software dedicated to the storage and analysis of small to very large chemical libraries. SA2 stores unique molecules in a MySQL database, and encapsulates several cheminformatics methods, among which: providers management, interactive visualisation, scaffold analysis, diverse subset creation, descriptors calculation, sub-structure / SMART search, similarity search and filtering. We illustrate the use of SA2 by analysing the composition of a database of 15 million compounds collected from 73 providers, in terms of scaffolds, frameworks, and undesired properties as defined by recently proposed HTS SMARTS filters. We also show how the software can be used to create diverse libraries based on existing ones.

Conclusions: Screening Assistant 2 is a user-friendly, open-source software that can be used to manage collections of compounds and perform simple to advanced cheminformatics analyses. Its modular design and growing documentation facilitate the addition of new functionalities, calling for contributions from the community. The software can be downloaded at <http://sa2.sourceforge.net/>.

Keywords: Chemical libraries, Molecular diversity, DRCS

Background

Exploring biology through the activity of small molecules is an established paradigm used in drug research for several decades now [1,2]. Today, a state of the art drug discovery program often begins with screening campaigns aiming at the identification of novel biologically active molecules. In the recent years, the rise of High Throughput Screening (HTS), combinatorial chemistry and the availability of large compound collections has led to a dramatic increase in the size of screening libraries, for both private companies and public organisations [3,4]. Yet, despite these constantly increasing capabilities, various authors have stressed the need to design better instead of

larger screening libraries [5-9]. Chemical space is indeed known to be almost infinite, and selecting the appropriate regions to explore for the problem at hand remains a challenging task.

What we call library design and analysis actually aims to increase the likelihood of screening collections to contain potentially active compounds, while ensuring that any of these represent an acceptable starting point for lead optimisation. In terms of biological activity, the concept of molecular diversity has proven useful to design libraries containing diverse chemotypes and hence increase the ratio of hits [10,11]. In terms of resource optimisation, the main difficulty lies in the removal of those nuisance compounds that are unlikely to be developed into effective drugs, especially using biochemical assays. For instance, reactives, warheads, promiscuous aggregating inhibitors, or more simply non-drug-like compounds are usually filtered out to avoid a waste of resources [12,13]. Most

*Correspondence: vchem@users.sourceforge.net;

luc.morin-allory@univ-orleans.fr

¹Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 7311 B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, France
Full list of author information is available at the end of the article

of these undesired properties are usually represented by either simple rule-based flags derived from physico-chemical properties (e.g. the Lipinski Rule of 5 [14]), or structural features encoded using the SMARTS notation [15].

The domain of chemoinformatics provides a plethora of methods that can be used to tackle various aspects of library analysis. Despite the growing diversity of available modeling and chemoinformatics tools, there are few software tools specifically dedicated to the management and the analysis of screening libraries. One of the main reasons for this no doubt is the specificity of each screening platform (e.g. plates format, automatic collection of results), which requires more specific developments to collect the results and associate them with tested molecules. Typically, screening collections are stored internally using in-house, usually web-based proprietary software programmes, some of which have been described in the literature [16-18]. Other general-purpose, proprietary packages also propose methods to handle chemical libraries. In particular, the Instant JChem application [19] from Chemaxon encapsulates various chemoinformatics functionalities, and makes it possible to store libraries in a database environment. Another example in this category is the CACTVS toolkit [20], an extensible distributed client/server system for the computation, management, analysis and visualisation of chemical information.

In the open-source literature, there is a growing range of tools that can deal with chemical libraries, each of them having its own specific applicability domain, ranging from general purpose to highly specific software. Workflow solutions allow the automation of numerous recurrent tasks, such as data reading / writing, filtering or visualisation, and some of them integrate chemistry functionalities. KNIME [21] in particular, is distributed with a set of chemistry nodes available for the Chemistry Development Kit (CDK) [22,23] and other chemoinformatics packages such as RDKit [24] and Indigo [25]. Various other advanced features have been encapsulated in KNIME nodes, and it is thus possible to e.g. compute molecular descriptors, perform substructure searches, or extract scaffolds. Recently, the CDK was also integrated in the Taverna workflow solution [26,27] through a set of more than 160 different workers handling chemoinformatics tasks. The combination of chemoinformatics functionalities with data-mining methods typically available in workflow solutions makes it possible to use more advanced strategies to analyse the content of chemical libraries.

Other more general-purpose software tools have been recently published by the chemoinformatics open-source community. Bioclipse [28] for example, is a general purpose modeling software that combines bioinformatics and

chemoinformatics functionalities in a modular and extensible workbench. In the context of the present work, Bioclipse can be used to perform simple routine tasks such as chemical file reading and visualisation, descriptor calculation, and SMART matching. More closely related to this work, AMBIT XT [29,30] is a chemoinformatics data management software, which consists in a MySQL database and a set of functional modules, allowing a variety of queries, data mining and predictive model building and application. Although not specifically dedicated to the management of screening libraries, it contains a set of chemoinformatics and data-mining facilities that make it usable for analysing collections of compounds. The software was recently enhanced by providing a set of OpenTox API [31] compliant REST web service interfaces to most of its functionalities, hence promoting collaborative development and data sharing [32]. More specific tools are clearly beyond the scope of this work, but it is worth mentioning the Scaffold Hunter [33], which allows one to display a chemical library in the form of an interactive scaffold tree, or the SARANEA package [34], which allows one to derive similarity graphs that can be used to perform structure-activity Relationship analysis.

This article presents Screening Assistant 2 (SA2), an open-source desktop software for chemical library management. SA2 stores unique chemical structures and properties in a MySQL database, and allows a variety of advanced chemoinformatics analyses and datamining queries to be performed. It has been designed to handle small to very large (up to millions of molecules) collections, and to integrate external sparse data in a flexible way (e.g. molecular descriptors, biological activities...). Besides various search and visualisation capabilities, SA2 can also be used to manage the provenance of stored compounds, and to create new subsets of molecules using many different methods, e.g. filtering, merging or diversity. SA2 was developed for facilitating the analysis of screening libraries, and to regroup multiple ways of mining these collections using chemoinformatics methods. Broadly speaking, it can also be used to quickly and interactively analyse the content of any chemical dataset.

Implementation

SA2 is open-source software, which means complete transparency and scalability in terms of algorithm implementation. A first version of the software was available on sourceforge and on the web-site of our laboratory [35,36]. This new version has been re-designed from scratch, keeping most of the concepts and features that were available in the first version. In this section, the general architecture of the software will be described, as well as the most important features and algorithms that make up its originality.

Architecture

SA2 is a desktop application based on the NetBeans Platform [37], a generic framework for JAVA / SWING based software. The NetBeans Platform allows applications to be developed in a modular fashion, thereby promoting good software engineering and programming practices. It contains a set of basic modules that can be used to handle various aspects of software development (e.g. fully dockable windowing system, module versioning or automatic updates), that would be otherwise time-consuming to (re)develop. This modular architecture makes it easier to add new menus, actions, extension points and modules without the need to go deep into the existing source code of the application. It is written in pure JAVA, and is therefore expected to be crossplatform. So far, it has been successfully tested on Windows XP and 7, Linux Ubuntu 10.4, CentOS 5. Issues related to the NetBeans Platform were found on some MacOSX operating systems which make the software currently incompatible with it.

Storage engine

All the data are stored in a MySQL database [38] using the InnoDB engine, which ensures data integrity. MySQL is a widely adopted database engine, extensively documented,

and with a wide user community. The choice of a database engine makes it possible to manage very large libraries - a database of around 7 million unique molecules has been successfully set up in our lab - and perform various routine tasks (e.g. filtering) more efficiently than using a simple file system. A mandatory requirement when using SA2 is therefore to have a MySQL server [38] installed and running on a server (or a simple desktop computer) that can be reached through a local network, along with a valid user account.

Input / Output

SA2 databases must be fed with MDL Mol-formatted input files (.sd or .sdf files) containing the full structure of the molecules. The full import workflow is shown in Figure 1. A step-by-step wizard is available to help new molecules import (and properties), as described in the documentation. Original molecules' names, if any, can be associated with each molecule, along with their CAS number. An internal unique database ID will also be automatically generated for each molecule. Text-delimited input files are also supported, and any number / kind of property can be associated with each existing molecule. These properties are stored in existing or new tables / fields

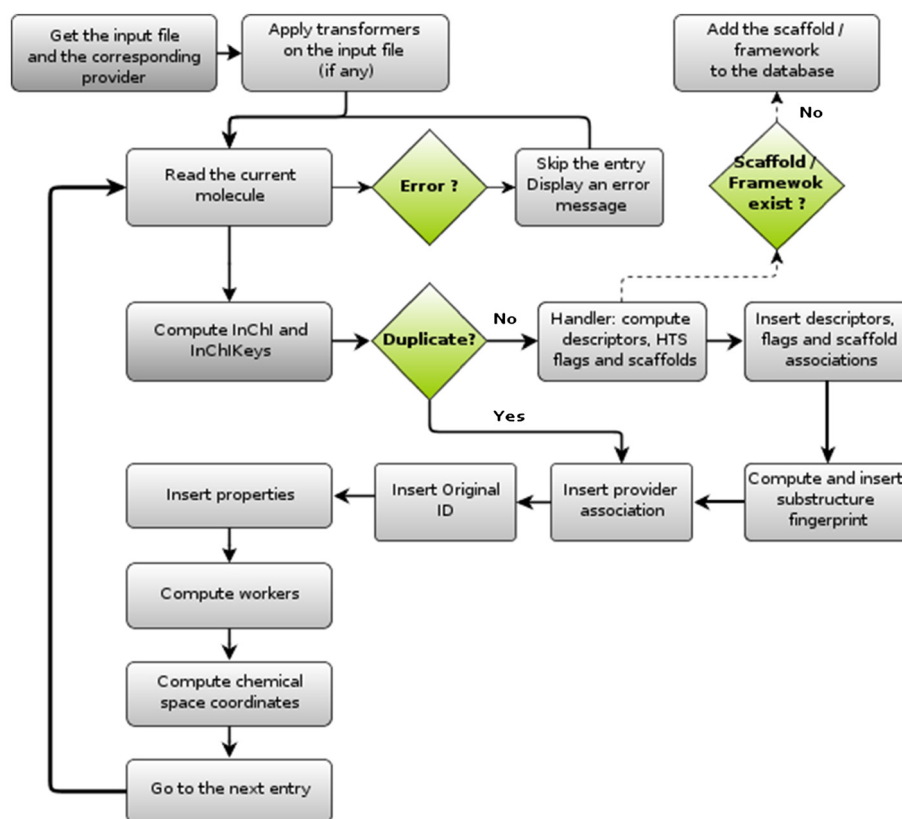


Figure 1 Full workflow of the SDF import process in SA2.

that can be created either directly during the import process, or using the dedicated window in SA2. The entire database (or a subset of it - see *Providers and libraries*) can be exported in the MDL Mol-format and / or in text-delimited format.

Storing and perceiving molecules

SA2 stores unique molecules under the form of a connection table in MDL Mol-format. Duplicates are removed using the IUPAC InChI identifier v1.0.4, taking into account stereochemistry. In the standard version of SA2, no particular pre-processing is applied on the input molecules, i.e. salts are not removed, protonation states and stereochemistry are kept as defined in the input file, etc. Although the standardisation of molecules is of prime importance, this procedure is usually problem-dependent, and we chose not to introduce any particular (possibly undesired) modification on the input molecules. It is however possible to code one's own standardisation services, referred to as "Transformers", that will be executed in a defined order - see the developer documentation for further details.

When a new set of molecules is imported into a database, various properties are automatically computed and associated with each molecule. To this end, SA2 introduces the notion of Molecular handler. A Molecular handler is an entity (typically a chemistry-aware programming library) that is responsible for the perception of

molecules. For each newly imported molecule, the handler reads the molecule and perceives atom types, calculates some simple properties (listed in Table 1), computes the scaffold and framework of the molecule, and computes various HTS-related flags. A molecule that cannot be loaded by the handler (e.g. invalid format) will not be imported into the database (an error message informs the user on each problematic molecule). Each database must be associated with a particular handler, which cannot be changed once the database has been created. Two different handlers are available in the current version of SA2: JOELib handler and the CDK handler, respectively based on the JOELib library [39] and the CDK library [22,23]. For computational performance reasons, the JOELib handler is used by default for new databases.

Table 1 provides the full list of properties that are computed for new molecules. Simple physicochemical descriptors will systematically be computed and stored in the main table of the database. Several binary flags can also be computed to provide simple ways of filtering compounds using widely adopted rules. The Lipinski drug-like [14] flag and the Rule of 3 flag [40] will be associated with each molecule. A set of 5 SMARTS-based flags is also available to provide warnings on potentially problematic compounds that typically contain undesired substructure patterns. Reactive and warhead compounds as defined in [12] are flagged as such. Three additional flags which account for Pan Assay Interference Compounds (PAINS)

Table 1 List of properties and flags that are automatically calculated when importing new molecules

Type	Name	Description	Ref.
SMARTS	Reactive	Reactive compounds (15)	[12]
	Warhead	Warhead compounds (20)	[12]
	PAINS < 15	Pan Assay Interference Compounds (409)	[13]
	PAINS < 150	Pan Assay Interference Compounds (55)	[13]
	PAINS > 150	Pan Assay Interference Compounds (16)	[13]
Flags	RO5	Lipinski's rule of 5 *	[14]
	RO3	Fragment rule of 3 *	[40]
	Exotic	Unrecognised atom type *	-
	Salt	Disconnected structures	-
Descriptors	Weight	Molecular weight	-
	LogP	Calculated logP *	-
	Heavy atoms	Number of heavy atoms	-
	HBA	Number of hydrogen bond acceptors *	-
	HBD	Number of hydrogen bond donors *	-
	Halogens	Number of halogen atoms	-
	Rot. Bonds	Number of rotatable bonds *	-
	Ring count	Number of rings (SSSR)	-
	Max Ring size	Maximum size of rings	-

The definition of descriptors marked by an asterisk is handler-specific.

are also available. These filters were defined in a recent paper [13] that highlights the need to remove a variety of classes of compound that are likely to be characterised as false positives in biochemical screening. The SMARTS version of these classes of compounds was extracted by Rajarshi Guha and made available to the community on his blog [41].

SMARTS-based flags are defined by a set of SMARTS queries which are stored in the database. The value of each flag for each molecule will be 1 if the molecule matches any of the SMARTS for a particular flag, 0 otherwise. The user can add / deactivate / delete any of these SMARTS, and hence recalculate the value of all the flags at any moment. An additional flag was also created to contain only user-defined SMARTS queries.

Properties and fingerprints

There are basically two ways of storing properties and fingerprints in SA2: importing them from an external source, and computing them directly within SA2, when possible. Properties and fingerprints in SA2 are organised in tables, and each table is assigned to a category. Various tables are available in the default version of SA2, e.g. tables for CDK descriptors / fingerprints and or MOE [42] descriptors. For new properties, the user must create new tables, either manually using the appropriate wizard, or directly when importing new molecules / properties. Fingerprints on the other hand, can only be imported using text-delimited files, in an fingerprint table which must be created using a dedicated editor prior to the import process. Two input fingerprint formats are currently supported: simple binary strings, and index strings where only the index of each bit set to 1 is recorded.

Several descriptors can also be computed directly within SA2. These include almost all the descriptors and fingerprints available in the CDK, several descriptors based on the JOELib library that were already available in the previous version of Screening Assistant, the Klekota and Roth fingerprint derived from [43], and the Indigo similarity fingerprint [25]. All these descriptors / fingerprints are stored in pre-installed tables. The reader should also note that two tables are available for MOE descriptors, which must however be computed externally and imported back into the database using SDF or text files. These two tables are automatically inserted because various PCA-based chemical spaces available in SA2 are based on these descriptors (see [44] and the next sections for details).

Providers and Libraries

When importing a new set of molecules, the user will be asked to associate these molecules with a so-called *provider*. A *provider* was primarily intended to represent commercial vendors that propose collections of compounds. In practice however, there is no restriction on

what a *provider* can represent. The notion of *provider* can hence be thought of as "where your compounds come from", e.g. a commercial vendor, a specific medicinal chemistry project, etc.

libraries on the other hand, represent subsets of molecules within a particular SA2 database. There are many ways of creating / modifying *libraries*: simple filtering rules using any descriptor(s), merging two existing libraries, creating a diverse library, creating a scaffold-based library, or saving search results. This concept is probably the most important one in SA2, as it provides great flexibility in the management of new sets of molecules. Once a new library has been created, it can be further analysed using the various visualisation and chemoinformatics facilities included in SA2, or simply exported in any output format available. Moreover, many tasks (e.g. searches, creation of diverse subsets) can be performed either on the whole database, or restricted to a particular library.

Scaffolds / Frameworks

A scaffold is a particular substructure that can be obtained based on the full structure of a query molecule. The best known Scaffold definition is that of Bemis and Murcko [45], who defined the scaffold of a molecule as being the union of ring systems and linkers.

In SA2, each newly imported molecule is associated with a unique Scaffold and a unique Framework. The scaffold retains all rings and linkers between rings, and removes all lateral chains, with the exception of exocyclic double bonds. The Framework of a molecule is the same as the scaffold, except that atom types are removed and bond orders are set to 1, leaving all atoms to SP3 carbons. The only exception here is 6-membered aromatic rings, for which the bond order is kept to retain aromaticity [35,36]. All remaining lateral chains, resulting from the exocyclic double bonds kept in the scaffold, are also removed.

Two windows in SA2 are dedicated to displaying the scaffold / framework of a molecule selected in any other window. All the molecules that belong to a particular scaffold can then be easily saved as a new library, added to an existing library, or removed from the database. Scaffolds are also used by diversity selection algorithms available in SA2, as described in the following sections. A report can also be generated, which will retrieve the most populated scaffolds, along with other information such as the total number of scaffolds in the database, the average number of molecules per scaffold and so on. This report can be generated for the whole database or for a particular library.

Visualisation

The concept of chemical space has been widely adopted by the chemoinformatics community as a way to represent and compare sets of molecules. SA2 provides various ways

of visualising multiple molecules. Simple X-Y plots can be used to draw molecules in an interactive panel using two selected properties. A similarity graph view can also be used, whereby molecules are drawn in the form of a graph. Nodes then represent molecules, and edges are drawn between two nodes if the corresponding molecules have a similarity higher than a given threshold.

SA2 also encapsulates the Delimited Reference Chemical Subspaces (DRCS) methodology described in a recent article [44]. DRCS are defined by the combination of a Principal Component Analysis (PCA) model (DRCS model), and one or several subspace(s) delimitation(s) (DRCS contour) intended to encompass the most populated part spanned by a particular library. This delimitation is computed on a reduced (2D) space obtained using the PCA model, and is based on the calculation of an average convex hull. Isolated compounds (outliers) are excluded prior to the creation of this delimitation, which finally represents the most populated (dense) subspace spanned by the reference library.

Three pre-computed DRCS models are available in new SA2 databases. These DRCS have been described in [44], and make use of different sets of descriptors. For each model, several subspaces are also included, which represent the densest part of different types of collections, e.g. general purpose HTS molecules, Oprea Leadlike molecules [46], Pharmaceutical molecules or fragment molecules (Rule of 3 [40]).

Molecular diversity

SA2 provides the possibility of extracting diverse subsets of molecules using scaffold-based min-max algorithms. The diversity selection can be performed on the whole database or on an existing library. This way, one can restrict the search to a carefully selected set of molecules. Briefly, the base algorithm is designed to ensure the presence of one molecule per scaffold (or framework, depending on the user's choice). It starts by retrieving all scaffolds within the database (or the selected library). These scaffolds are either randomly shuffled, or ordered by decreasing number of associated molecules. The first molecule is added to the library as being the molecule that is the most similar to an average fingerprint computed on all the molecules that belong to the first selected scaffold. The similarity between two molecules is defined by any similarity coefficient (e.g. Tanimoto) available in SA2 applied to the selected fingerprint. Next, for each remaining scaffold, the molecule having the lowest similarity to the currently selected molecules is added to the library.

A maximum similarity cutoff can also be defined. For a particular scaffold, all candidate molecules that have a similarity to the already selected molecules greater than this cutoff are not accepted, thereby ensuring that similar scaffolds are not over-represented in the library. The

counter part of this is a higher computational complexity if the similarity cutoff is defined too small.

Once all the scaffolds have been processed, the final number of molecules may still be lower than the desired size of the library. Two reasons can lead to this situation: (1) the number of scaffolds in the database is lower than the required number of molecules, and (2) the similarity cutoff used is too small. In both cases, the entire selection process is just repeated. In the second case, the cutoff is automatically increased for each new run. The selection process finally stops when *N* molecules have been selected.

Searches

Various search capabilities are available in the software, such as exact structure, similarity, substructure, SMARTS searches, or simpler searches using the name or database ID. Similarity searches can be performed using any of the fingerprints available in the database. The entire database is scanned within the application, and a bitwise comparison is performed using the selected similarity metric (Tanimoto coefficient by default). A similarity search using a query molecule that does not exist in the database nevertheless requires the use of a fingerprint that can be calculated within SA2 (i.e. not an external fingerprint that has been imported into the database). SMARTS search is performed by retrieving molecules from the database and applying a SMARTS matching algorithm (referred to as SMARTS engine in the application) to detect matches. It is also possible, with a working internet connection, to visualise a SMARTS query using the SMARTS viewer service provided by the bioinformatics center of the university of Hamburg [47,48].

Substructure search is performed as a two-step process, with a prescreening step followed by a graph isomorphism test. The first step is done using a database query that filters out the molecules that cannot match the query. This query makes use of two different types of information: (1) a small set of basic properties (number of heavy atoms, number of SSSR, and number of halogens) that are indexed in the main table, and (2) a fingerprint that is calculated for each molecule upon import. This fingerprint is computed using the Indigo library, which provides an implementation of a specific substructure fingerprint. The fingerprint is encoded as a set of 33 unsigned integers of 32 bits, which means that a maximum of 1056 bits is accepted. The fingerprint calculation can be replaced by one's own implementation (see Extension points), and the values associated with each molecule can be subsequently updated by recomputing the fingerprint on the entire database.

For any type of search, the results obtained are displayed in a specific window, and can be saved as a new

library, added to an existing library, or removed from the database.

Extension points

The Netbeans Platform makes it possible to define extension points, referred to as Services, using a NetBeans Platform-specific mechanism. A Service is a JAVA class that is able to provide a specific functionality as defined by the Service facade it corresponds to (a JAVA interface). For example, a *FingerprintSimilarityMetric* service provides a floating number ranging from 0 to 1, based on two fingerprints representing two molecules. The *SubstructureFingerprint* service on the other hand, returns a fingerprint based on a SDF text representing a molecule. Based on such a mechanism, one can easily add new services by simply implementing the corresponding service interface, and registering it using a single line annotation. Various extensions points were actually mentioned previously, e.g. molecular handlers, transformers, SMARTS engine, substructure fingerprint, or substructure engine. The only requirement to setup your own service implementation is to know the full list of services along with their specifications, and provide an implementation of it in a new module; there is usually no need to change or even know the source-code of the original application that makes use of this service. For example, in a new SA2 module, adding a new similarity metric can be performed by adding a single JAVA class implementing the corresponding interface, i.e:

```
@ServiceProvider(service=FingerprintMetric.class)
public class FPTanimotoMetric implements
FingerprintMetric
{
    @Override
    public String getName() {
        return "Tanimoto" ;
    }

    @Override
    public String getDescription() {
        return "Tanimoto_coefficient_for_
fingerprints." ;
    }

    public float getSim(BitSet m1, BitSet m2) {
        // Implementation...
    }
    (...)
}
```

The most important line here is the `@ServiceProvider`. Using this JAVA annotation, one register this class as a fingerprint metric service. No more operation is required

except implementing the methods defined by the *FingerprintMetric* interface: the new metric will automatically appear everywhere the *FingerprintMetric* service is required, e.g. for similarity search, for diverse subset creation... The developer documentation of SA2 provides detailed examples on this.

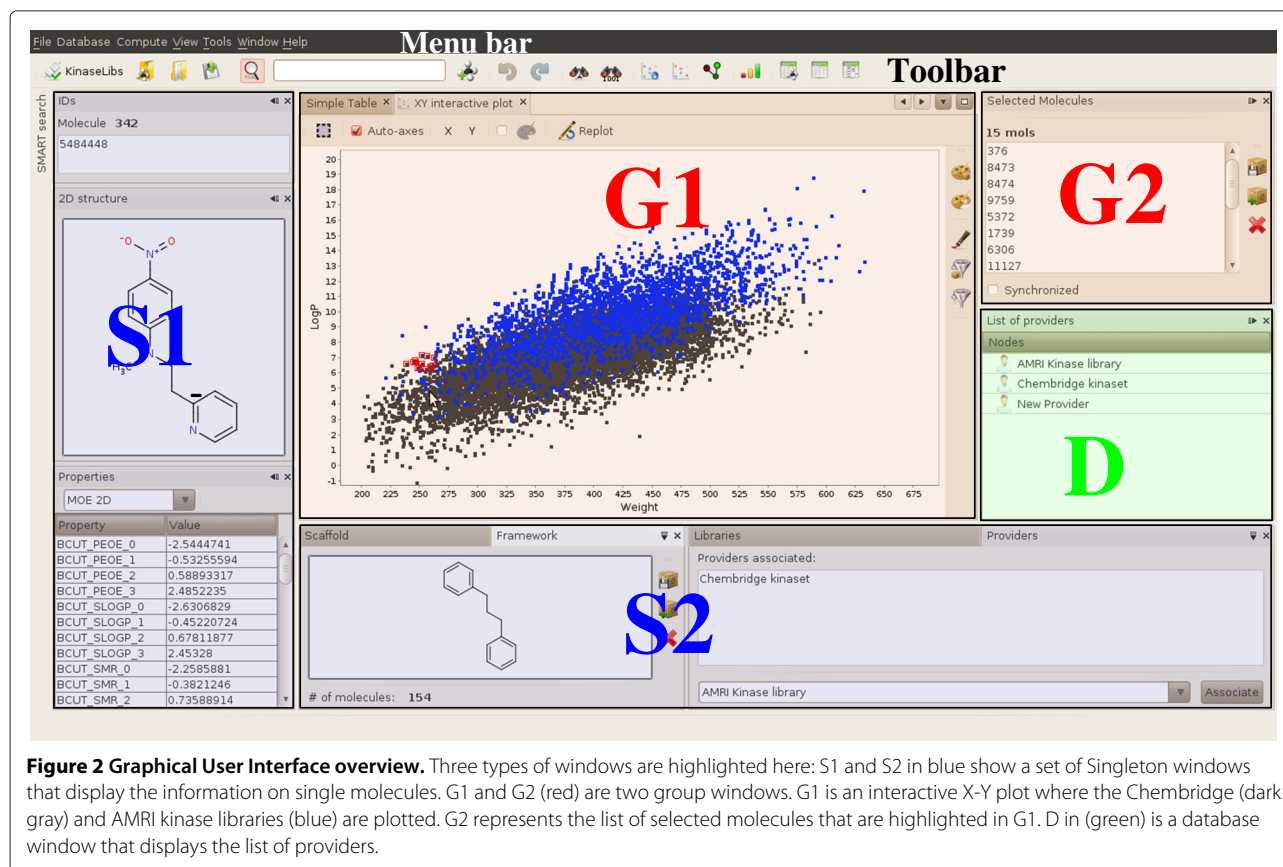
One particular service should be emphasised here. The *MolWorker* service can be implemented to compute any operation of an input molecule. *MolWorkers* can be activated either directly when importing new molecules, or can be run afterwards using the dedicated menu. This service offers the possibility of adding any additional calculation in a completely transparent way. From the point of view of a developer, the only requirement (as for all services) is to create a new module, and to provide a registered implementation of the *MolWorker* interface. There is no restriction on the operation that a *MolWorker* should do. In the 1.0 version of SA2, three workers are available: the CDK worker, which calculates CDK fingerprints and descriptors, the JOELib worker, which does the same using the JOELib library, and the Indigo worker, which makes it possible to compute two fingerprints.

Results

To illustrate the use of SA2, various collections of different sizes have been analysed. A brief description of the Graphical User Interface (GUI) is first provided as an introduction. In the first case study, an in-house database of 6.3 million unique, standardised (see [44] for details) compounds collected from 73 commercial vendors was used. In the second case study, two focused commercial screening libraries were compared, and SA2 was used to select a diverse subset from a combination of the two libraries. Finally, various possible improvements and perspectives will be discussed. All the data and figures presented here have been generated with the version 1.1.0 RC1 of SA2.

Graphical User Interface (GUI)

An overview of the GUI is given in Figure 2. The GUI is composed of four main types of window: (1) Singleton windows, where the information on a single molecule is displayed (e.g. 2D structure, list of associated providers...); (2) Group windows, where multiple molecules can be displayed in a certain form (e.g. 2D plot, simple tables...); (3) Database windows, which usually display a list of specific entities stored in the database (e.g. providers, libraries, SMART flags...); and (4) other windows which do not fit into any of these categories. All the windows are accessible through the main menu located at the top of the main window, and shortcuts are also available for various views in the toolbar. Each window can be opened, closed, reduced, undocked or maximised, which provides great flexibility in selecting the relevant information that needs to be displayed. The relative positioning of windows



is also completely customisable using intuitive drag and drop, and is restored upon startup.

In Group windows, the user can select the molecules of its choice interactively. When a single molecule is selected, each Singleton window is updated to display the information corresponding to this molecule. When multiple molecules are selected, they are highlighted in the corresponding view, and a specific window is updated to display the list of molecules that forms the current selection (as illustrated in Figure 2). Based on this list, various operations can be performed, such as removing these molecules from the database, or creating a new library based on this selection. It is also possible to synchronise the selection, whereby molecules selected in one Group window are automatically selected in all the other Group windows that are open.

Most of the algorithms are available in the Compute menu. This menu includes all the advanced algorithms and search capabilities that can be applied on the database, such as diversity selection, creation of new DRCS models, similarity search, etc.

Case study 1: large scale analysis

A collection of 73 SDF files representing 73 different commercial vendors was imported into an SA2 database.

The precise collection and pre-processing of these data is described elsewhere [44]. It took around 4 weeks to import all the data into the database. This was mainly due to the large number of molecules to be processed (15 million altogether): for each molecule, more than 400 SMARTS have to be matched, dozens of descriptors are calculated, the scaffolds and frameworks are computed along with their InChI and InChIKey, and all these data must then be stored and indexed within the MySQL database.

In this illustrative analysis, the basic properties and HTS flags listed in Table 1 were computed, as well as the scaffolds and frameworks of each molecule. Various reports were generated within SA2, and the results are summarised in Tables 2, 3, 4, Additional file 1: Table S1 and S2 for detailed values, and in Figure 3.

Scaffolds and frameworks

Table 2 outlines the results obtained by generating a Scaffold and a Framework report in SA2. A total of 1 084 411 scaffolds and 247 689 frameworks can be found in the database, representing respectively 15.24% and 3.47% of the compounds. Only 1.06% and 0.18% of the scaffolds are needed to obtain 50% of the compounds in the database, which shows that only a very few scaffolds and

Table 2 Scaffold and framework representativity

	Scaffold		Framework	
	Count	Percent	Count	Percent
Database	1 084 411	15.24%	247 689	3.47%
Singleton	647 260	9%	104 250	1.5%
Cumulative freq. (50%)	11 565	1.06%	447	0.18%
Cumulative freq. (80%)	150 777	13.87%	7 741	3.13%

Singletons are core structures that are associated with only one molecule. Percentages for Database and Singleton rows are expressed as the number of scaffolds (Count column) divided by the total number of molecules in the database. Cumulative frequency values represent the number (resp. proportion) of scaffolds that are needed to obtain a certain percentage of compounds (in brackets) in the database.

frameworks represent a very large part of the database. Figure 3 shows the 20 most populated scaffolds and frameworks as extracted from the Scaffold report of SA2. Interestingly, the Benzene core structure is the most populated for both scaffolds and frameworks, with almost 3% of the compounds associated with it. There is also a significant proportion of acyclic compounds, which are represented here by a single carbon atom. On the other hand, the uneven distribution of scaffolds and frameworks is highlighted with 59.52% and 42.09% of scaffolds / frameworks which are associated with a single molecule. This corresponds to 9% and 1.5% of the molecules in the database that have a unique scaffold / framework.

Additional file 1: Table S2 presents the detailed results in terms of compound unicity, and scaffolds / frameworks composition for each provider. An outline of these results can be found in Table 3. The reader should note that low percentages of compounds unicity (CU), scaffold unicity (SU) and Framework unicity (FU) have to be analysed carefully: some providers include the libraries of other providers in their own collection, which obviously biases the results. Large unicity values on the other hand are more informative.

As summarised in Table 3, around half of the providers have a CU greater than 10%, and only 15 greater than

50%. In terms of the proportion of scaffolds and frameworks, almost all providers contain at least 10% of unique scaffolds, but this number drops to only 26 providers having more than 10% of frameworks, which suggests that some providers have a significant proportion of scaffolds that share the same graph and differ only in their heteroatoms composition. On average, providers contain 24.7% of scaffolds and 11.6% of frameworks. Providers with the largest proportion of scaffolds are generally those that contain a small number of compounds. There are however, some top-populated providers that have a large proportion of scaffolds and high unicity as well, e.g. ChemDiv (CU = 54.97%, SU = 45.23% and FU = 31.49%). The Chimiothèque Nationale, which federates collections of synthesis products available in French academic laboratories, also contains a large percentage of original compounds, (CU = 85.24%, SU = 64.23% and FU = 37.37%), hence highlighting the potential interest of academic screening collections.

Drug-likeness

The database was also analysed in terms of Drug-like properties. Various reports were generated with SA2 for the Lipinski rule of 5, the fragment Rule of 3, Reactive and PAINS flags. The detailed results can be found in Additional file 1: Table S2, and an outline is provided in Table 4. These data show that there is a fairly low percentage of potentially problematic compounds in screening libraries. On average, 5.7% of compounds fail the Rule of 5, while 6.9% of compounds are found reactive and 5.9% might be PAINS compounds. These average values nevertheless mask some differences between providers, with some of them having up to 20% percent of potentially problematic compounds. The percentage of fragment-like compounds is on the other hand more evenly distributed, with many providers containing more than 50% of fragments. Besides the availability of libraries specifically designed for fragment-based screening, these high percentages can also be explained by the presence of building

Table 3 Summary of the scaffold composition and unicity analysis

	Unicity (CU)	Scaffolds		Frameworks	
		Proportion	Unicity (SU)	Proportion	Unicity (FU)
Min	0%	6.3%	0%	2.2%	0%
Max	100%	84.2%	87.9%	57.8%	49%
Average	24.9%	24.7%	13.3%	11.6%	5.9%
> 10%	37	69	27	26	13
> 20%	33	37	17	11	8
> 50%	15	4	4	1	0

Unicity is defined as the proportion of molecules (or scaffolds / frameworks) that are exclusive to a given provider, i.e. that cannot be found in any other provider. The proportion of scaffolds / frameworks are expressed as the number of molecules divided by the number of scaffolds / frameworks associated with a given provider. The minimum, maximum and average values through all vendors are given in this table. The number of vendors having one of these indices up to a given threshold is given in the second part of the table.

Table 4 Summary of the drug-like analysis

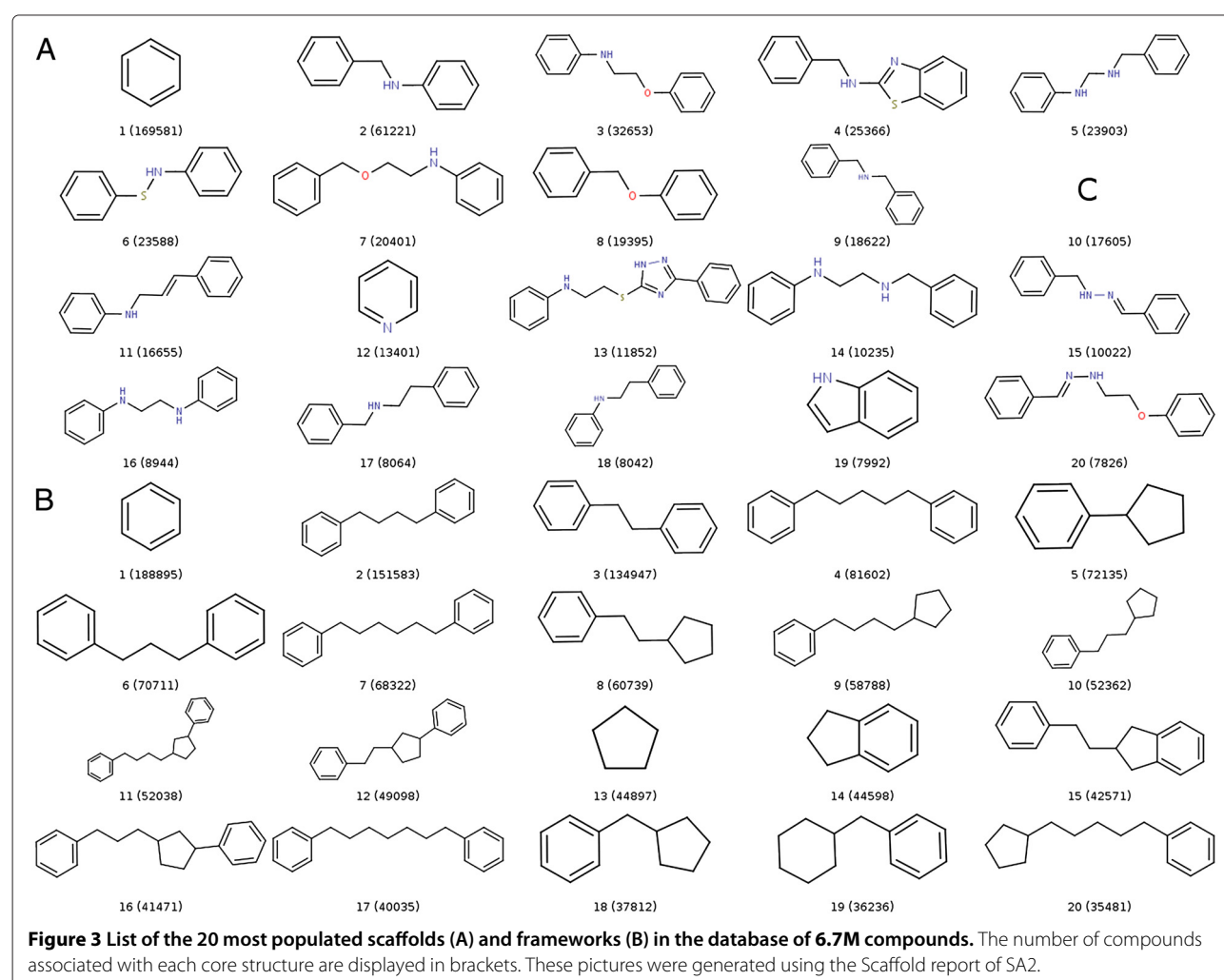
	Fragment-like	Non-Drug-like	Undesired		Global
			Reactive	PAINS	
Min	1.3%	0%	0%	0%	0%
Max	86.7%	25.6%	19.5%	27.5%	29%
Average	27.7%	5.7%	6.9%	5.9%	12.3%
> 5%	59	36	36	35	66
> 10%	47	13	21	14	48
> 20%	34	1	0	1	7

Percentages are expressed as the minimum, maximum and average proportion of molecules that are flagged for each criterion, computed over each provider. The number of providers having one of these indices up to a given threshold is given in the second part of the table.

blocks. Providers containing a large number of molecules usually contain a small percentage of fragment-like compounds (the 15 providers containing the largest number of molecules have less than 10% of fragments), which are much more heavily represented in small to medium-sized libraries.

Conclusion

In conclusion, a detailed picture of the composition of compound collections from various points of view can be obtained using SA2. In this example, a large number of available collections was analysed in terms of scaffolds and drug-like properties. Such an analysis can be



used to provide a detailed picture of the chemical space spanned by the available screening libraries, and to identify those collections that can be of great interest in terms of originality.

Case study 2: kinase libraries

For this second illustrative example, we assume the need to create a diverse subset of potential kinase inhibitors based on two existing focussed libraries: the AMRI kinase library (3232 molecules), and the Chembridge Kinase library (11501 molecules). Both libraries were downloaded from their respective vendors' websites, pre-processed as described in [44], and imported into a SA2 database (a different provider were associated with each input file), leading to 14732 unique molecules. SA2 was subsequently used to analyse both libraries in order to guide the creation of a diverse subset that contains 1500 molecules (around 10% of the entire database).

A simple analysis was first conducted to compare the distribution of various physico-chemical descriptors in the two libraries. Using the Property Stats window of SA2, the distribution of any property can be plotted and compared between any set of libraries (or on the entire database). Figure 4 shows a comparison of the distribution of some physico-chemical descriptors for both

libraries. What can be concluded from this simple analysis is that the two libraries seem fairly complementary, with some properties showing different distributions. The same observation can be made when comparing the scaffold composition of the two libraries. The Charts window was used here to evaluate the overlap between the libraries in terms of compounds and scaffolds. Table 5 shows that there is a small overlap between the two libraries in terms of compounds, scaffolds and frameworks. Only one compound is shared between the two libraries, and they both contain a large percentage of exclusive scaffolds and frameworks. All these observations are further supported by the projection of both libraries in reduced chemical spaces. Figure 5 shows the projection of the two libraries in a PCA space computed on the entire database using the CDK BCUT descriptors, and in the DRCS-MOE2D reduced space. Despite some visible overlaps, there are obviously some parts of each space that are covered by only one of the two SA libraries.

The first step in creating a diverse library is to remove all the potentially problematic compounds. To this end, two filtered libraries were created for each provider (i.e. each original library). The filters were defined to remove all reactive, warhead and PAINS compounds in both libraries. The AMRI and Chembridge libraries contained

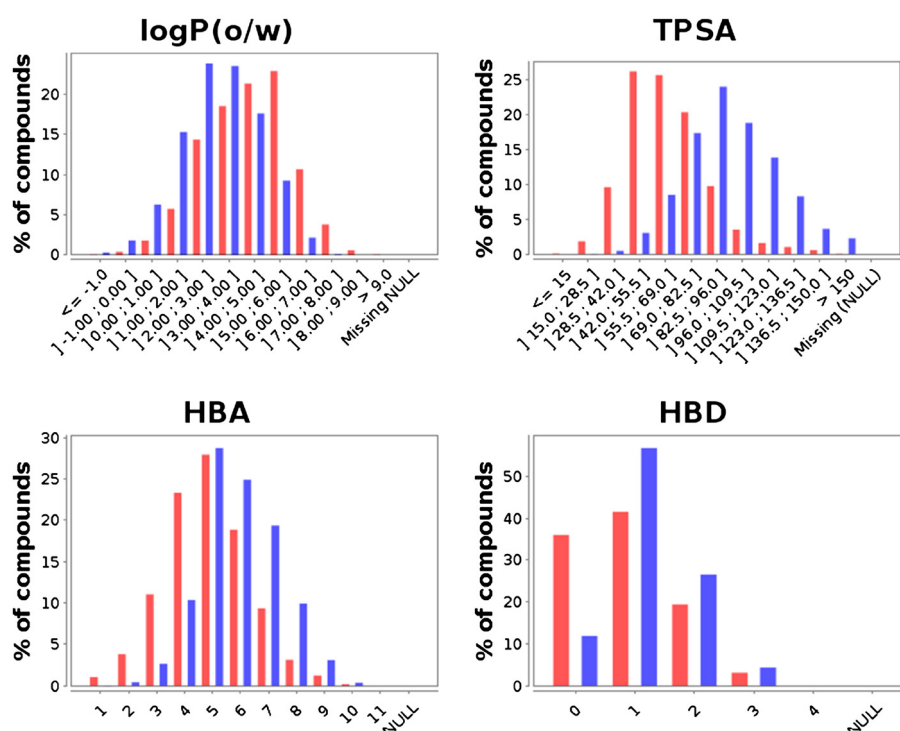


Figure 4 Comparative distribution of some physico-chemical properties. Chembridge Kinaset (red) and the AMRI kinase library (blue). HBA (resp. HBD) stands for Hydrogen Bond Acceptor (resp. Donor). These histograms can be obtained by simply clicking on the property to analyse in the *Properties* window of SA2.

Table 5 Scaffold, framework and compound originality of the AMRI and Chembridge kinase libraries

	AMRI		Chembridge	
	Count	Percent	Count	Percent
Frameworks	873	27.0%	2 204	19.2%
Frameworks unicity	747	85.6%	2 078	94.3%
Scaffolds	1 053	32.6%	4 036	35.1%
Scaffolds unicity	1 008	95.7%	3 991	98.9%
Overlap (compounds)	1 molecule			

The proportion of scaffolds (resp. frameworks) is expressed as the number of unique scaffolds (resp. frameworks) divided by the number of molecules in the library. The scaffolds (resp. frameworks) unicity is expressed as the number of scaffolds (resp. frameworks) unique to the library divided by the total number of scaffolds (resp. frameworks) in the library.

respectively 159 and 587 such compounds. A third library was then created by merging the two previously created libraries. This new library was subsequently used to create a scaffold-based diverse subset of 1500 molecules using the default parameters (scaffold-based, MACCS fingerprint, Tanimoto metric with a similarity cutoff of 0.6 increased by 0.1 at each new run, frequency-based ordering of core structure). The diverse subset creation procedure took around 4 seconds to complete.

Once a new diverse subset has been obtained, SA2 offers various ways of assessing its diversity. A first overview can be obtained by plotting the entire database and highlighting the subset in one or several reduced spaces available in SA2. Figure 6 shows the projection of the diverse library within the two spaces used previously. Such a plot provides a good overview of the chemical space coverage of our library. One can easily see that it covers the space spanned by the entire database fairly well, although some parts of the space remain rather poorly represented (Figure 6b). This can be explained by the fact that there is a significant difference in the size of the two original libraries. As a consequence, the Chembridge diverset was

clearly oversampled. The composition of the diverse subsets which contain 78% of molecules coming from the Chembridge library. A simple way to obtain a more balanced selection would be to extract a random subset from the Chembridge library, and perform the diversity selection on the union of this random subset and the AMRI library. Repeating the process to test this hypothesis, a new library was obtained which contained 60% of Chembridge compounds. As suggested by the remaining bias, it seems that overall, the Chembridge library offers more diversity.

Several reports available in SA2 can provide further insight into the diversity of a library. In particular, the Similarity report (*Compute-Similarity-Similarity report*) can be used to calculate various diversity indices based on any of the fingerprints available in SA2, along with their distribution. The percentage of scaffolds can also be easily obtained, as shown previously. Table 6 shows various diversity indices computed in SA2 on the diverse library, and two random subsets of the same size. Obviously, the diverse library shows greater diversity in terms of scaffolds, frameworks, and similarity. The difference between random and diverse subsets becomes however less visible when using different fingerprints than that used to create the diverse subset. This behavior is to a certain extent expected as the diverse subset has been optimised using the MACCS fingerprint. The diverse library however, remains the most diverse in all cases, and in particular for the average nearest neighbor (Avg. NN) similarity indice, which is known to be efficient in discriminating between to libraries of the same size [49].

Discussion

Performance and possible improvements

An SA2 database is indexed and optimized to obtain a good compromise between the time needed to import new molecules, and the time needed to perform all the analyses available in the software. Most of the calculations

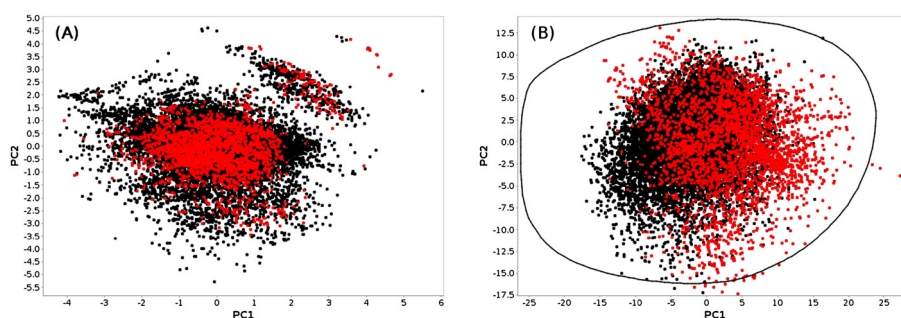


Figure 5 PCA projections of the Chembridge (black dots) and AMRI (red dots) libraries. The first reduced space (A) has been computed within SA2 on the entire kinase database using the CDK BCUT descriptors which were computed upon import. The second reduced space (B) is the DRCS-MOE2D space which is already available in new SA2 databases, and for which descriptor values were imported. The contour shown in black encompasses the densest region spanned by HTS compounds (see [44] for details).

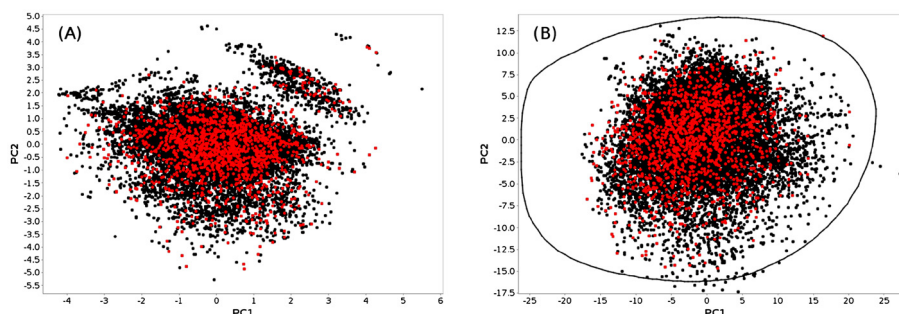


Figure 6 PCA projections of the diverse subset (red dots) in the two spaces described previously. The remaining molecules (AMRI + Chembridge) are drawn in black. The figure has been generated with the *DRCS plot* window of SA2.

are therefore quite fast: plotting several tens of thousands of molecules in a reduced or an X-Y plot can be achieved within few seconds depending on the hardware, keeping the interactive selection completely functional. The scaffold and framework reports illustrated in the “Results” section can be obtained within around two minutes for 7 million compounds. For newly imported molecules, and despite the large number of calculations involved, the process is fast enough to rapidly set up new databases for small to medium-sized datasets. Of course, the time needed to import new molecules will increase with the number of workers selected. From this point of view, the only possible improvement would be to split each

input file and distribute the computation through several threads. Although it does not pose any substantial technical challenge, special care would nevertheless be required as to the integrity of the data as the database will be accessed by multiple processes. The diverse subset algorithms also performs very efficiently. As illustrated previously, a diverse subset can be obtained rapidly using the default parameters. This performance is due to the heuristic nature of the algorithm, which was designed to obtain a good compromise between an optimum diversity and an algorithm that can be used on very large databases. The performance of the algorithm obviously depends on the number of molecules required as well as the similarity cutoff used. Higher value of the similarity cutoff lead to more diverse libraries, but require more time to complete.

Despite the use of a database engine as a backend storage solution, SA2 is not a chemistry database cartridge, nor it is based on any particular existing one. Despite this matter of fact, most of the search capabilities of SA2 also perform fairly well. The name and exact structure search are fully optimised, and the results can be retrieved almost immediately, even for very large databases. Similarity search is of course slower, but the performance is still acceptable. A similarity search launched in medium-sized databases (e.g. 100 000 molecules) retrieves the results within one or two seconds, depending on the hardware. Similarity searches in the 7 million database, using the Tanimoto metric (cutoff = 0.7) and the Indigo similarity fingerprint (512 bits), took less than 3 minutes to scan the entire database on a modern computer through a local network. Typically, around 50 000 structures per seconds are processed for such a search. Furthermore, the search results are updated for each new hit, making it possible to browse them as they are retrieved. The performance of similarity search actually depends on three main factors, in decreasing order of importance: the similarity cutoff (the lower it is, the slower the search will be, as much more results have to be retrieved), the size of the fingerprint used, and the location of the database.

Table 6 Diversity evaluation for diverse and random subsets

	Diverse	Random 1	Random 2
Scaffold%	84%	61%	63%
Framework%	52%	44%	44%
MACCS			
Avg. pairwise	0.44	0.48	0.48
Avg. NN	0.76	0.88	0.88
Max. sim.	0.80	1.00	1.00
Pubchem			
Avg. pairwise	0.48	0.50	0.50
Avg. NN	0.82	0.87	0.87
Max. sim.	0.98	1.00	1.00
Indigo			
Avg. pairwise	0.26	0.29	0.29
Avg. NN	0.70	0.81	0.81
Max. sim.	1.00	1.00	1.00

The percentage of scaffolds and frameworks are reported for each library. The Tanimoto metric and three different fingerprints were also used to compute average pairwise similarity (Avg. pairwise), average nearest neighbor similarity (Avg. NN), and maximum pairwise similarity (Max. sim) within each library, using 3 different fingerprints that can be computed directly within SA2. These data were generated with the *Similarity report* and the *Scaffold report* of SA2.

Substructure search has also been optimised, but is still subject to some limitations in specific cases. For instance, the pre-filtering step can be almost useless if the query substructure is too generic, e.g. a simple benzene. Using scaffolds provides a way to retrieve a subset of the results almost immediately, but the remaining of the database still have to be scanned. Two possible improvements could be rapidly made to accelerate substructure search: (1) improve the fingerprint used as filtering, but this would not solve the problem mentioned previously and (2) store a set of generic small structures, and flag all compounds containing these structures when importing them into the database. This second solution may provide the advantage of having the results of low-specificity substructure searches directly available. On the other hand, it would certainly impact the time needed to import new molecules, and require a substantial amount of additional disk space. Finally, although it would require more development efforts, the integration of an existing cheminformatics database cartridge, such as Bingo [50], MyChem [51], PgChem [52] or OrChem [53], may certainly represent an interesting direction. As SA2 currently only support MySQL, it would be of particular interest to integrate cartridges for other database engine, e.g. Bingo or OrChem for PostgreSQL.

Perspectives

SA2 was initially designed to manage and analyse screening libraries. The most important missing piece is probably the possibility to manage screening projects (including plates, activity types, targets...), and to integrate the results of HTS assays in a more specific way. The diversity of HTS assays and screening results types pose substantial challenges as to the organisation and integration of the resulting data. Currently, SA2 offers the possibility of associating any kind of property to each molecule, but in a non-specific and uncontrolled way. This provides great flexibility, but it would no doubt be advantageous to integrate controlled vocabulary (typically an ontology) in order to organise the data more appropriately. In a recent article, Visser et al. [54] described a novel approach to standardise, organise and semantically define biological assays and screening results. Ontologies can be truly valuable in the mining of HTS data, and open up exciting perspectives for tools like Screening Assistant 2.

Although at first restricted to the domain of screening, SA2 is also moving toward becoming a more general-purpose software that deals with chemical libraries in a broad sense. Hence, there is again considerable room for adding new cheminformatics, datamining or Structure-activity relationship features that are more specific to the analysis of small to medium-sized datasets. The first step toward this was taken by adding a graph similarity view (see "Implementation" section, or the official

documentation), which is primarily useful to perform SAR on small (one or two thousand molecules) datasets. A Self-organizing map module is also on its way, which will provide a complementary non-linear method to the PCA currently available in SA2.

Conclusions

Screening Assistant 2 complements the growing ecosystem of modeling tools by providing a set of cheminformatics facilities integrated in a database environment. It facilitates the management of chemical libraries through an intuitive and interactive graphical interface, and provides a set of advanced methods to analyse and exploit their content. As with any new software, there are still many improvements that can be made, and probably even more directions to take. Special care was taken to provide a comprehensive documentation for both users and developers. We therefore encourage anyone to feed the project with remarks, new ideas and features, and hope that the software will be useful to the community.

Availability and requirements

Project name: SA2

Project home page: <http://sa2.sourceforge.net/>

Operating system(s): Platform independent

Programming language: JAVA / SQL

Other requirements: Java 1.6.0 or higher <http://java.sun.com/>, MySQL 5.1 or higher <http://dev.mysql.com/downloads/mysql/>, and the NetBeans Platform 6.9.1 for developers willing to add new modules <http://netbeans.org/features/platform/>.

License

Screening Assistant 2 is released under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

Additional file

Additional file 1: [sumpinf/providers.pdf](#). Two large tables containing detailed values for the provider analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VLG designed and implemented the software, and drafted most of the manuscript. LMA initiated and supervised the development of Screening Assistant. PV and AA supervised specific parts of the project, fed it with new ideas, and participated in testing the software. LC and SB made extensive tests, and participated in documenting the software. All authors approved the final manuscript.

Acknowledgements

The authors would like to thank Peter Schmidtke for providing the Mac-OSX InChI binaries and for fruitful comments on the manuscript, and the CDK and

Indigo teams for their help in using their respective libraries. VLG thanks the Conseil général du Loiret for funding his Ph.D.

Author details

¹Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 7311 B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, France.
²Bioinformatic Modelling Department, Technologie Servier, 45007 Orléans Cedex 1, France. ³Fédération de Recherche, Physique et Chimie du Vivant, Université d'Orléans-CNRS; FR 2708, avenue Charles Sadron, 45071 Orléans Cedex 2, France.

Received: 1 June 2012 Accepted: 6 August 2012
Published: 31 August 2012

References

- Mayr LM, Bojanic D: **Novel trends in high-throughput screening.** *Curr Opin Pharmacol* 2009, **9**(5):580–588.
- Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DVS, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS: **Impact of high-throughput screening in biomedical research.** *Nat Rev Drug Discovery* 2011, **10**(3):188–195.
- Gibbon P: **High-throughput hit finding and compound-profiling technologies for academic drug discovery.** *Drug Discovery Today* 2009, **5**:e3–e7.
- Frearson JA, Collie IT: **HTS and hit finding in academia—from chemical genomics to drug discovery.** *Drug Discovery Today* 2009, **14**(23-24):1150–1158.
- Walters WP, Namchuk M: **Designing screens: how to make your hits a hit.** *Nat Rev Drug Discovery* 2003, **2**(4):259–266.
- Harper G, Pickett SD, Green DVS: **Design of a compound screening collection for use in high throughput screening.** *Comb Chem & High Throughput Screening* 2004, **7**:63–70.
- Gillet VJ: **New directions in library design and analysis.** *Curr Opin Chem Biol* 2008, **12**(3):372–378.
- Compound Profiling: Size Impact on Primary Screening Libraries.** [http://ddw.net-genie.co.uk/currentissue/487302/compound_profiling_size_impact_on_primary_screening_libraries.html].
- Hajduk PJ, Galloway WRJD, Spring DR: **Drug discovery: A question of library design.** *Nature* 2011, **470**(7332):42–43.
- Yeap SK, Walley RJ, Snarey M, van Hoorn WP, Mason JS: **Designing compound subsets: comparison of random and rational approaches using statistical simulation.** *J Chem Inf and Model* 2007, **47**(6):2149–2158.
- Sukuru SCK, Jenkins JL, Beckwith RE, Scheiber J, Bender A, Mikhailov D, Davies JW, Glick M: **Plate-Based Diversity Selection Based on Empirical HTS Data to Enhance the Number of Hits and Their Chemical Diversity.** *J Biomol Screening* 2009, **14**(6):690–699.
- Rishton GM: **Nonleadlikeness and leadlikeness in biochemical screening.** *Drug Discovery Today* 2003, **8**(2):86–96.
- Baell JB, Holloway GA: **New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.** *J Med Chem* 2010, **53**(7):2719–2740.
- Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Delivery Rev* 2001, **46**(1-3):3–26.
- Daylight Chemical Information Systems Manual.** [http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html].
- Leach AR, Bradshaw J, Green DV, Hann MM, 3rd Delany JJ: **Implementation of a system for reagent selection and library enumeration, profiling, and design.** *J Chem Inf and Comput Sci* 1999, **39**(6):1161–1172.
- Yasri A, Berthelot D, Gijzen H, Thielemans T, Marichal P, Engels M, Hoflack J: **REALIS: A Medicinal Chemistry-Oriented Reagent Selection, Library Design, and Profiling Platform.** *J Chem Inf and Model* 2004, **44**(6):2199–2206.
- Mosley RT, Culberson JC, Kraker B, Feuston BP, Sheridan RP, Conway JF, Forbes JK, Chakravorty SJ, Kearsley SK: **Reagent Selector: using Synthon Analysis to visualize reagent properties and assist in combinatorial library design.** *J Chem Inf and Model* 2005, **45**(5):1439–1446.
- Instant JChem , Chemaxon.** [http://www.chemaxon.com].
- Ihlenfeldt WD, Takahashi Y, Abe H, Sasaki S: **Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility.** *J Chem Inf and Comput Sci* 1994, **34**:109–116.
- Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B: **KNIME: The Konstanz Information Miner.** In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Heidelberg: Springer; 2007.
- Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E: **The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics.** *J Chem Inf and Comput Sci* 2003, **43**(2):493–500.
- Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL: **Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics.** *Curr Pharm Des* 2006, **12**(17):2111–2120.
- RDKit: Open-source cheminformatics.** [http://www.rdkit.org].
- The Indigo toolkit, GGA Software Services.** [http://ggasoftware.com/opensource/indigo].
- Lanzén A, Oinn T: **The Taverna Interaction Service: enabling manual interaction in workflows.** *Bioinformatics* 2008, **24**(8):1118–1120.
- Kuhn T, Willighagen EL, Zielesny A, Steinbeck C: **CDK-Taverna: an open workflow environment for cheminformatics.** *BMC Bioinf* 2010, **11**:159.
- Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C, Torrance G, Wagener J, Willighagen E, Steinbeck C, Wikberg J: **Bioclipse 2: A scriptable integration platform for the life sciences.** *BMC Bioinf* 2009, **10**:397.
- AMBIT project.** [http://ambit.sourceforge.net/].
- Jeliazkova N, Jaworska J, Worth A: **Open Source Tools for Read-Across and Category Formation.** In *In Silico Toxicology: Principles and Applications*. Edited by Lewin RA. Cambridge: RSC Publishing: Cronin M. and Madden J; 2010:408–445.
- Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, Girschick T, Buchwald F, Wicker J, Karwath A, Gutlein M, Maunz A, Sarimveis H, Melagraki G, Afantitis A, Sopasakis P, Gallagher D, Porokov V, Filimonov D, Zakharov A, Lagunin A, Glorizova T, Novikov S, Skvortsova N, Druzhilovsky D, Chawla S, Ghosh I, Ray S, Patel H, Escher S: **Collaborative development of predictive toxicology applications.** *J Cheminformatics* 2010, **2**:7.
- Jeliazkova N, Jeliazkov V: **AMBIT RESTful web services: an implementation of the OpenTox application programming interface.** *J Cheminformatics* 2011, **3**:18.
- Wetzel S, Klein K, Renner S, Rauh D, Oprea TI, Mutzel P, Waldmann H: **Interactive exploration of chemical space with Scaffold Hunter.** *Nat Chem Biol* 2009, **5**(8):581–583.
- Lounkine E, Wawer M, Wassermann AM, Bajorath J: **SARANEA: A Freely Available Program To Mine Structure-Activity and Structure-Selectivity Relationship Information in Compound Data Sets.** *J Chem Inf and Model* 2011, **50**:68–78.
- Monge A, Arrault A, Marot C, Morin-Allory L: **Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers.** *Mol Diversity* 2006, **10**(3):389–403.
- Screening Assistant (previous version); Developed by Aurelien Monge.** [http://www.univ-orleans.fr/icoa/screeningassistant/].
- The NetBeans official website regrouping all ressources around the Platform. SA2 is based on the 6.9.1 version of the platform.** [http://netbeans.org/features/platform/].
- The MySQL official website. SA2 has been tested on 5.* versions of the MySQL.** [http://www.mysql.com/products/enterprise/database/].
- JOELib, a computational chemistry JAVA library.** [http://joelib.sourceforge.net/].
- Congreve M, Carr R, Murray C, Jhoti H: **A 'rule of three' for fragment-based lead discovery?** *Drug Discovery Today* 2003, **8**(19):876–877.
- R. Guha, PAINS Substructure Filters as, SMARTS, 2010-11-14.** [http://blog.rguha.net/?p=850].
- MOE, version 2009.10 Chemical Computing Group (CCG): Montreal, Canada, 2009.** [http://www.chemcomp.com/software.html].
- Klekota J, Roth FP: **Chemical substructures that enrich for biological activity.** *Bioinformatics* 2008, **24**(21):2518–2525.

44. Le Guilloux V, Colliandre L, Bourg S, Guénégou G, Dubois-Chevalier J, Morin-Allory L: **Visual characterization and diversity quantification of chemical libraries: 1. creation of delimited reference chemical subspaces.** *J Chem Inf and Model* 2011, **51**(8):1762–1774.
45. Bemis GW, Murcko MA: **The Properties of Known Drugs. 1. Molecular Frameworks.** *J Med Chem* 1996, **39**(15):2887–2893.
46. Oprea TI, Davis AM, Teague SJ, Leeson PD: **Is there a difference between leads and drugs? A historical perspective.** *J Chem Inf and Comp Sci* 2001, **41**(5):1308–1315.
47. Schomburg K, Ehrlich H, Stierand K, Rarey M: **From Structure Diagrams to Visual Chemical Patterns.** *J Chem Inf and Model* 2011, **50**(9):1529–1535.
48. **The SMARTS viewer server, University of Hamburg.** [http://smartsview.zbh.uni-hamburg.de/].
49. Meini T, Ostermann C, Berthold MR: **Maximum-Score Diversity Selection for Early Drug Discovery.** *J Chem Inf and Model* 2011, **51**(2):237–247.
50. **The Bingo database cartridge.** [http://ggasoftware.com/opensource/bingo].
51. **The MyChem database cartridge.** [http://mychem.sourceforge.net/].
52. **The Pgchem::tigress database cartridge.** [http://pgfoundry.org/projects/pgchem/].
53. Rijnbeek M, Steinbeck C: **OrChem - An open source chemistry search engine for Oracle®.** *J Cheminformatics* 2009, **1**:17.
54. Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schürer SC: **BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results.** *BMC Bioinf* 2011, **12**:257.

doi:10.1186/1758-2946-4-20

Cite this article as: Le Guilloux et al.: Mining collections of compounds with Screening Assistant 2. *Journal of Cheminformatics* 2012 **4**:20.

Publish with **ChemistryCentral** and every scientist can read your work free of charge

“Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge.”

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.chemistrycentral.com/manuscript/



ChemistryCentral

Profiling collections of compounds with Screening Assistant2

Vincent Le Guilloux^{*1}, Alban Arrault², Lionel Colliandre¹, Stéphane Bourg³, Philippe Vayer², and Luc Morin-Allory^{*1}

¹ Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans, UMR CNRS 7311 B.P. 6759, rue de Chartres, 45067 Orléans Cedex 2, France.

² Bioinformatic Modelling Department, Technologie Servier, 45007 Orléans Cedex 1, France.

³ Fédération de Recherche, Physique et Chimie du Vivant, Université d'Orléans-CNRS; FR 2708, avenue Charles Sadron, 45071 Orléans Cedex 2, France

Email: Vincent Le Guilloux^{*} - vchem@users.sourceforge.net; Luc Morin-allory^{*} - luc.morin-allory@univ-orleans.fr;

^{*}Corresponding author

SUPPLEMENTARY INFORMATION

Table S1 - Proportion of flagged compounds by provider

Table S2 - Uniqueness and scaffold composition by provider

Table S1 - Proportion of flagged compounds by provider

The proportions are given for each provider, as the number of molecules that are flagged divided by the total number of molecules. The Undesired column represents the proportion of compounds that have been flagged by at least one of the 4 HTS flags (reactive and PAINS).

Provider	Size	RO5	RO3	Reactive	PAINS< 15	PAINS< 150	PAINS> 150	Undesired
Ambinter	5204292	92.75	5	2.81	0.23	1.19	4.46	8.38
Enamine	1516762	97.09	4.49	2.95	0.1	0.9	2.53	6.3
ScientificExchange	1184178	85.77	5.42	2.46	0.14	3.29	8.28	13.52
Princeton	971576	92.33	7.44	2.13	0.34	2.1	7.07	11.14
VitasMLab	839835	91.48	6.48	2.22	0.51	2.35	7.92	12.43
ChemDiv	663386	90.59	2.28	1.86	0.21	1.74	6.59	9.57
Chembridge	633235	97.72	7.26	1.7	0.61	1.35	5.6	8.89
Specs	471247	92.24	9.05	2.13	0.37	2.73	6.81	11.39
InterBioScreen	463396	92.86	5.11	2.56	0.24	2.39	8.95	13.22
Asinex	379455	92.44	6.21	1.72	0.25	1.59	4.78	8.02
LifeChemical	333633	96.22	3.44	1.26	0.06	0.41	3.61	5.19
AMRI	254611	74.4	1.4	1.41	0	0.64	3.93	5.9
AlindaChemical	253867	93.98	9.43	1.71	0.29	1.57	3.58	6.9
AsisChem	234085	91.4	4.44	1.72	0.36	4.05	10.86	16.07
Pharmeks	227343	86.91	3.86	3.1	0.34	2.41	10.79	15.72
TimTec	204835	92.66	13.62	3.47	0.65	2.19	5.97	11.76
ChemTI	170752	93.64	9.99	1.97	0.29	1.67	2.16	5.84
ChemicalBlock	125109	90.25	7.8	2.86	0.87	2.04	6.32	11.57
Otava	120101	95.72	6.89	2.27	0.08	1.78	6.67	10.39
Labotest	105237	94.14	20.73	5.37	0.34	3.04	8.45	16.34
Nanosyn	65041	93.52	12.51	2.41	0.43	2.99	9.25	14.28
Maybridge	60391	97.38	20.05	3.27	0.26	1.43	3.33	7.67
SynthonLab	52713	93.02	6.76	3.78	1.08	7.62	18.79	29.04
Key-Organics	42647	97.05	10.98	6.56	0.74	1.47	3.84	11.86
Bionet	42640	97.05	11.02	6.54	0.74	1.48	3.85	11.86
ChimiothequeNationale	41950	93	27.14	9.27	0.03	1.13	2.99	12.93
Apollo Scientific	38981	98.31	58.16	17.21	0.12	1.05	1.34	19.19
MatrixScientific	38321	99.43	55.11	12.52	0.11	2.37	2.59	17.1
FluoroChem	34377	98.37	59.1	13.94	0.1	1.22	1.18	16.17
Intermed	32041	81.26	1.33	0.09	0.03	2.82	7.37	9.57
Arkive	30511	95.16	35.31	6.33	0.79	0.87	5.76	13.03
EMC	28097	84.7	3.04	1.57	0	0.54	2.27	4.39
Analyticon Discovery	26427	85.9	2.26	3.76	0.02	1.52	2.38	7.5
IS-Chemical-Technology	25041	98.21	61.46	12.9	0.13	1.05	1.14	14.78
Aronis	23707	96.04	12.35	3.75	0.4	3.74	3.24	10.82
TOSLab	16557	80.92	2.87	2.55	1.55	4.28	11.45	18.11
Peakdale	14621	94.07	4.86	1.01	0.06	0.04	0.74	1.51
TorontoResearchChemicals	13249	88.5	25.87	12.41	0.1	1.44	2.88	16.27
Oakwood	13208	99.08	66.5	13.95	0.09	0.95	0.74	15.66

Table S1 (continued)

Provider	Size	RO5	RO3	Reactive	PAINS< 15	PAINS< 150	PAINS> 150	Undesired
ALBChemical	9998	99.19	10.37	2.91	0.03	1.82	9.48	13.64
ARVI	9960	98.65	33.99	8.43	0.13	1.6	2.4	11.9
MyriaScreen	9934	99.3	14.75	2.68	0.34	1.96	4.64	9.14
Spectrum	8673	90.89	4.44	2.43	0.07	2.4	8.68	13.48
CombiBlock	7087	99.84	59.22	4.74	0	0.08	0.17	4.98
Menai	4012	98.48	12.59	12.34	0.25	8.28	5.81	23.58
Sinova	3791	98.87	61.12	8.92	0.11	0.79	0.76	10.31
ChemiK	3760	99.92	83.32	19.34	0.08	0.72	0.27	20.11
ACBBlocks	3232	100	69.89	14.36	0	2.66	3.25	20.2
Szintekon	2664	98.46	39.08	10.47	0	2.06	2.55	14.08
PepTech	2376	99.03	24.33	1.73	0	0.17	0.63	2.53
FocusSynthesis	2321	98.54	55.75	8.32	0	0.34	0.52	9.09
ExclusiveChemistry	2272	89.74	24.34	3.65	0	0.53	1.63	5.5
Sequoia	2132	89.82	17.73	10.79	0.38	4.17	3.05	17.4
Biosynth	1804	91.19	37.31	12.58	0.11	1.83	1.66	15.63
Cayman	1777	88.91	14.24	18.8	0.06	2.59	1.8	22.85
FrontierScientific	1677	96.42	72.39	17.05	0	0.12	0	17.17
SynChem	1546	100	58.47	19.47	0.26	1.81	2.26	23.29
InFarmatik	1373	99.85	45.67	7.79	1.38	0.66	1.53	11.22
Biotrend	1261	87.63	19.11	5.55	0.16	4.04	3.81	12.61
Prestwick	1181	92.97	22.44	6.27	0.51	2.03	2.96	11.26
Adesis	1173	99.83	46.63	6.73	0	0.09	0	6.82
Chemivate	1108	98.74	1.71	1.71	0	0	0	1.71
Endeavour	813	99.88	81.55	14.88	0	0.49	1.11	15.62
Endotherm	677	91.43	33.68	11.23	0	1.92	0.15	13.15
KaironKem	647	100	58.27	14.22	0.31	5.26	5.72	23.18
GreenPharma	647	93.35	23.65	8.04	0	0.46	0.15	8.65
AFChemPharm	617	96.27	48.3	8.1	0.16	1.46	1.13	10.21
Chess	563	100	54.71	7.46	0	0.36	1.42	9.24
Synphabase	490	88.98	37.55	14.49	0.41	0.41	3.88	18.37
Pyxis	317	100	77.6	0	0	0	0.63	0.63
Sinof	263	100	86.69	13.69	0	0	0	13.69
EnzoLifeSciences	202	90.1	16.83	18.32	0	1.49	0	19.8
Azasynt	67	100	14.93	0	0	0	0	0

Table S2 - Uniqueness and scaffold composition by provider

Unicity is defined as the proportion of molecules (or scaffolds / frameworks) that are exclusive to a given provider (that cannot be found in any other provider). The proportion of scaffolds / frameworks are expressed as the number of scaffolds / frameworks divided by the number of molecules associated with a given provider.

Provider	Size	Unicity (%)	Scaffolds		Frameworks	
			Proportion (%)	Unicity (%)	Proportion (%)	Unicity (%)
Ambinter	5204292	29.41	16.63	14.23	3.88	11.09
Enamine	1516762	9.3	28.9	8.96	6.17	3.13
ScientificExchange	1184178	56.7	9.58	29.43	3.11	21.6
Princeton	971576	14.7	13.47	10.18	4.3	6.69
VitasMLab	839835	0.15	16.66	0.04	5.45	0.03
ChemDiv	663386	54.97	17.98	45.23	7.1	31.49
Chembridge	633235	2.79	25.05	3.9	7.9	1.04
Specs	471247	4.44	16.88	1.76	6.07	1.41
InterBioScreen	463396	1.41	18.67	1.41	6.93	0.3
Asinex	379455	36.13	20.8	34.13	7.47	22.26
LifeChemical	333633	2.77	18.25	3.78	6.48	1.02
AMRI	254611	1.09	11.5	0.31	6.23	0.1
AlindaChemical	253867	2.43	16.35	0.34	5.52	0.22
AsisChem	234085	14.31	12.37	4.69	4.82	1.92
Pharmeks	227343	0.02	23.11	0	9.93	0
TimTec	204835	0.65	25.94	0.36	9.38	0.17
ChemTI	170752	3.07	17.08	3.57	6.36	3.55
ChemicalBlock	125109	0.89	29.05	0.71	11.87	0.22
Otava	120101	10	17.83	4.77	7.3	1.64
Labotest	105237	36.94	24.44	20.99	8.86	9.33
Nanosyn	65041	20.07	23.6	10.14	9.61	5.33
Maybridge	60391	50.52	30.23	23.52	9.08	7.83
SynthonLab	52713	1.8	11.33	0.77	5.01	0.38
Key-Organics	42647	0.14	26.48	0.11	8.99	0.05
Bionet	42640	0	26.45	0	8.96	0
ChimiothequeNationale	41950	85.24	28.24	64.23	12.39	37.37
Apollo Scientific	38981	0.26	12.27	0	3.6	0
MatrixScientific	38321	13.39	10.76	4.63	2.48	1.89
FluoroChem	34377	20.69	10.36	8.54	3.2	6.73
Intermed	32041	0.02	7.05	0	3.7	0
Arkive	30511	0.09	28.85	0.05	11.42	0
EMC	28097	98.91	14.81	87.93	8	45.22
Analyticon Discovery	26427	0.01	32.09	0	15.79	0
IS-Chemical-Technology	25041	26.11	13.21	20.34	5.19	12.23
Aronis	23707	0	16.25	0	6.68	0
TOSLab	16557	0.49	36.2	0.2	21.18	0.11
Peakdale	14621	0	36.13	0	16.05	0
TorontoResearchChemicals	13249	64.87	27.07	39.18	14.15	27.25
Oakwood	13208	1.25	10.49	0.29	3.1	0.24

Table 2 (continued)

Provider	Size	Unicity (%)	Scaffolds		Frameworks	
			Proportion (%)	Unicity (%)	Proportion (%)	Unicity (%)
ALBChemical	9998	0.07	20.33	0	8.83	0
ARVI	9960	4.5	25.38	1.19	12.11	0.25
MyriaScreen	9934	11.15	45.14	5.73	18.7	0.75
Spectrum	8673	2.24	37.59	0.95	21.93	0.37
CombiBlock	7087	29.24	10.44	13.92	3.01	2.35
Menai	4012	0	33.85	0	16.13	0
Sinova	3791	62.49	11.37	14.62	6.07	4.35
ChemiK	3760	6.41	6.3	2.53	2.18	2.44
ACBBlocks	3232	1.39	18.25	1.02	6.03	0
Szintekon	2664	65.73	25.26	43.24	11.26	14.33
PepTech	2376	48.95	11.7	14.03	5.01	0
FocusSynthesis	2321	44.68	34.17	29.63	15.68	11.81
ExclusiveChemistry	2272	67.21	19.94	18.54	8.93	7.88
Sequoia	2132	22.84	44.93	6.58	27.77	4.05
Biosynth	1804	42.18	18.02	13.23	10.31	7.53
Cayman	1777	67.08	34.83	32.15	22.06	14.8
FrontierScientific	1677	20.57	14.13	3.38	4.65	3.85
SynChem	1546	28.27	19.66	8.55	6.34	1.02
InFarmatik	1373	0	34.38	0	15.29	0
Biotrend	1261	35.21	60.75	18.54	41	9.86
Prestwick	1181	8.55	59.44	3.56	34.38	1.48
Adesis	1173	28.99	8.78	17.48	3.5	0
Chemivate	1108	98.92	42.69	86.26	22.56	49.2
Endeavour	813	20.79	12.05	2.04	2.83	0
Endotherm	677	72.23	38.4	23.85	18.46	8.8
KaironKem	647	38.64	84.23	0	57.81	0
GreenPharma	647	1.24	22.41	4.83	8.19	0
AFChemPharm	617	74.07	45.38	41.79	26.74	23.03
Chess	563	1.07	22.2	0	8.53	0
Synphabase	490	1.02	41.02	0	27.55	0
Pyxis	317	97.16	12.62	37.5	7.57	0
Sinof	263	5.32	12.17	0	5.7	0
EnzoLifeSciences	202	47.03	16.83	5.88	10.4	0
Azasynt	67	100	77.61	76.92	47.76	3.13

Conclusion générale

Dans cette thèse, nous nous sommes intéressé au développement de méthodes et d'outils chimoinformatiques appliqués à la gestion, la comparaison et l'analyse de chimiothèques. Nous avons proposé une nouvelle méthodologie permettant d'évaluer de manière visuelle et numérique la couverture et la diversité relative d'une ou plusieurs chimiothèques par rapport à un espace ACP de référence. Nous avons également développé plusieurs outils open-source permettant de mettre en oeuvre cette méthode. Enfin, nous avons développé une nouvelle version du logiciel Screening Assistant afin, d'une part proposer de nouvelles fonctionnalités, toujours dans l'optique de faciliter la gestion et l'analyse de chimiothèques et d'autre part, permettre l'utilisation des DRCS dans ce logiciel. Les résultats obtenus peuvent être résumés en trois principaux points :

La méthodologie des DRCS permet d'obtenir une délimitation de sous-espaces réduits à partir de la projection en deux dimensions de molécules en utilisant un modèle ACP. Cette délimitation se veut représentative du sous-espace le plus dense défini par une certaine proportion de composés extraits d'un ensemble de molécules de référence. Elle permet notamment de mieux délimiter l'espace le plus dense en ne prenant pas en compte les molécules exotiques situées aux extrémités d'un espace chimique. Nous avons d'une part, défini et validé la méthodologie permettant de construire ces délimitations en identifiant des jeux de paramètres adaptés et d'autre part, montré son intérêt pour la comparaison et l'analyse visuelle de chimiothèques. A partir d'une base de données de plus de 6 millions de molécules dédiées au criblage, nous avons utilisé la méthode afin de définir des espaces de référence utilisant trois jeux de descripteurs 2D et 3D. Des délimitations représentatives des types de composés suivants ont ensuite été créées : HTS, "drug-like", pharmaceutiques, adaptés au criblage par fragment. L'ensemble des programmes développés pour produire les résultats présentés dans

cette thèse ont par ailleurs été documentés et mis à disposition de tous sous license GPL. Ils sont téléchargeables à l'adresse suivante : <http://www.univ-orleans.fr/icoa/DRCS/index.html>

Les DRCS ont été appliqués au calcul d'indices relatifs de diversité. Les règles de Waldman ont tout d'abord été modifiées afin de prendre en compte les cas où des chimiothèques de taille différentes doivent être comparées. Nous avons ensuite réimplémenté 11 indices de diversité et nous en avons adapté certains aux DRCS afin de permettre un quadrillage des sous-espaces se voulant plus représentatif que les méthodes classiquement utilisées. Nous avons analysé le comportement de ces indices dans des situations fictives afin de mettre en avant leurs avantages et leurs inconvénients. Nous avons constaté qu'aucun d'entre eux ne permet à lui seul de statuer sur la diversité relative d'une chimiothèque et nous avons donc sélectionné cinq de ces indices pour leur complémentarité. Différents aspects de la diversité sont capturés par chacun d'entre eux et nous avons pu montrer que leur utilisation conjointe permet une analyse plus complète et plus pertinente.

Deux limites importantes peuvent être mises en avant pour ces deux parties :

1. Les délimitations proposées sont calculées sur un espace réduit à deux dimensions. Cela présente l'avantage non négligeable d'être visuellement interprétable, mais limite aussi l'information capturée par la délimitation, puisque deux composantes principales ne représentent qu'une certaine proportion de la variance expliquée. Deux pistes peuvent être envisagées pour pallier cela : (1) faire les calculs en utilisant d'autres composantes (PC2-PC3, PC3-PC4 par exemple), ce qui complique l'analyse mais permet également de prendre en considération d'autres propriétés représentées par chaque composante principale et (2) faire le calcul de l'enveloppe convexe sur plus de deux dimensions. En utilisant 3 dimensions, nous pourrions conserver l'interprétation visuelle tout en augmentant la quantité d'information représentée par la délimitation. L'utilisation de plus de 3 dimensions aura uniquement un intérêt pour le calcul d'indices de diversité.
2. Les délimitations proposées se basent sur une enveloppe convexe et sont donc bien adaptées à la linéarité de l'ACP. Elles sont en revanche moins adaptées à des méthodes non-linéaires, à des très petits jeux de données ou à des chimiothèques contenant des groupes de molécules bien distincts. Pour ces types de cas, le calcul de plusieurs enveloppes convexes peut s'avérer nécessaire. Le calcul d'une enveloppe concave peut repré-

senter une alternative intéressante, bien qu'elle ajouterait une difficulté supplémentaire puisqu'il peut exister plusieurs solutions à ce type de problème.

Une nouvelle version de Screening Assistant a été développée . Nous sommes reparti de zéro pour permettre l'ajout de nombreuses fonctionnalités facilitant son utilisation et permettant une analyse plus poussée et interactive de chimiothèques. Basée sur la plateforme NetBeans, l'architecture du logiciel a été totalement repensée à travers un développement modulaire facilitant l'ajout de nouvelles fonctionnalités. L'interface utilisateur proposée par cette plateforme apporte également un confort d'utilisation nettement accru. De nombreuses fonctionnalités ont également été ajoutées. Nous avons notamment permis l'utilisation des DRCS à travers la création et le stockage de modèles ACP et de délimitations telles que définies par la méthode décrite dans les chapitre 2 et 3. La visualisation des espaces chimiques a été rendue plus interactive, offrant la possibilité de sélectionner, filtrer ou encore colorier les molécules suivant la valeur d'une propriété. De nouvelles fonctionnalités de recherche (par sous-structure, par SMARTS, par similarité, par scaffold / framework) permettent aussi de créer de nouvelles chimiothèques focalisées. De nombreux rapports interactifs ont également été ajoutés afin de proposer plusieurs méthodes se voulant complémentaires pour l'analyse de chimiothèques et la détermination de leur diversité. Le logiciel a été mis à disposition de tous sous licence GPL, et est téléchargeable à l'adresse suivante : <http://sa2.sourceforge.net/>.

L'ensemble de ces travaux a pu être présenté sous la forme de trois publications, sept posters et deux logiciels open-source.

Annexes

1 Descripteurs calculés automatiquement par SA2

Type	Nom	Description	Ref
SMARTS	Reactive	Reactive compounds (15)	
	Warhead	Warhead compounds (20)	
	PAINS <15	Pan Assay Interference Compounds (409)	
	PAINS <150	Pan Assay Interference Compounds (55)	
	PAINS >150	Pan Assay Interference Compounds (16)	
Flags	RO5	Lipinski rule of 5 *	
	RO3	Fragment rule of 3 *	
	Exotic	Unrecognised atom type *	
	Salt	Disconnected structures	
Descriptors	Weight	Molecular weight	
	Heavy atoms	Disconnected structures	
	HBA Number	Disconnected structures	
	HBD Number	Disconnected structures	
	Rot. Bonds	Disconnected structures	
	Ring count	Disconnected structures	
	Max Ring size	Disconnected structures	

Tableau A.1 – Liste des propriétés calculées par SA2 lors de l’insertion des molécules. Les propriétés marquées par une étoile ont des valeurs spécifiques du Molecular Handler utilisé.

2 Schéma de base de données pour la gestion des descripteurs et des espaces chimiques

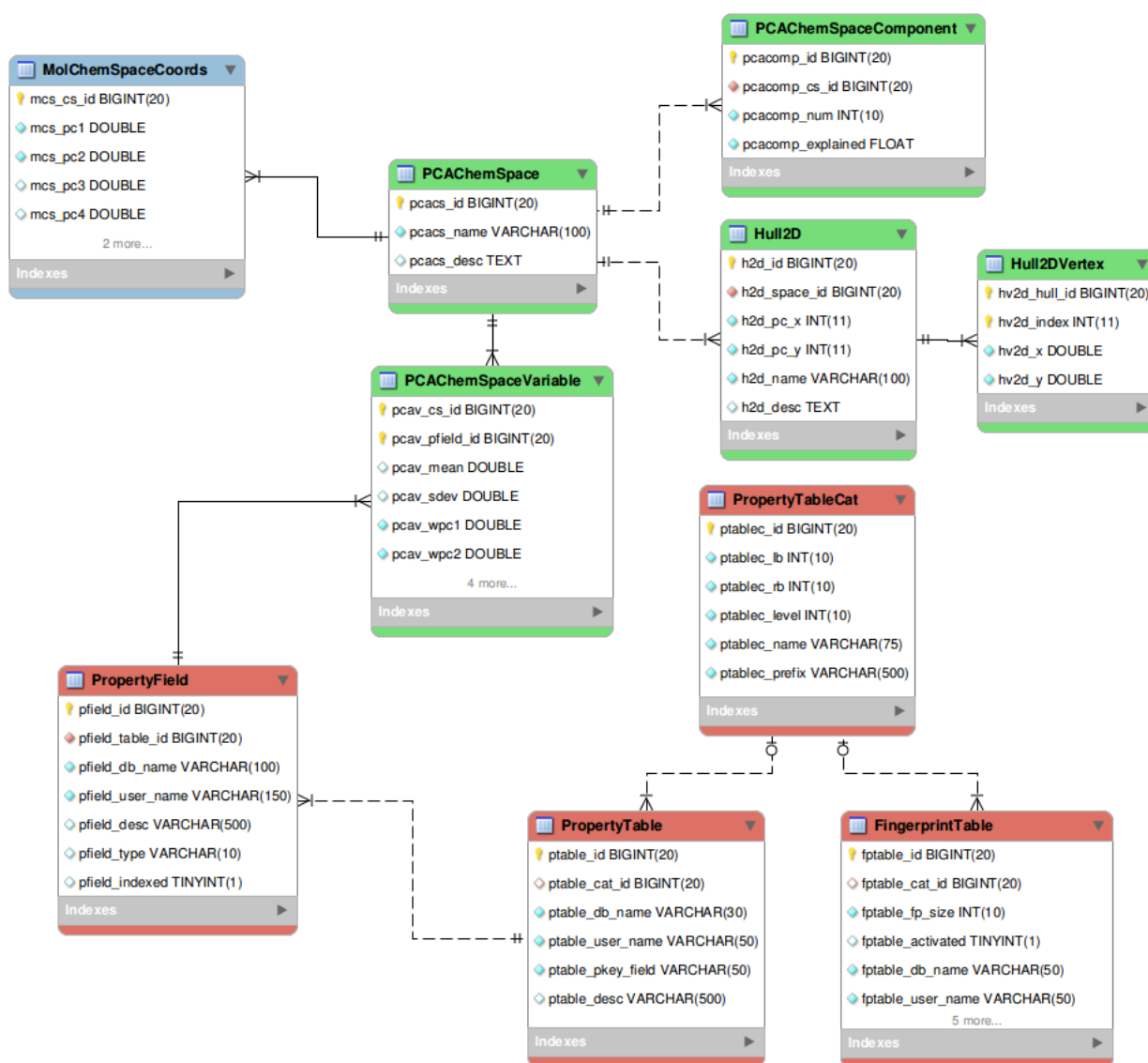


Figure A.1 – Tables permettant la gestion des propriétés et des espaces chimiques.

3 Guide de démarrage rapide SA2



Quickstart guide

Last edited: 06/10/2011

By: VLG

Sourceforge: [SA2 website](#)

Help: [SA2 forums](#)

Documentation [index](#)

This page will guide you through the basic utilisation of SA2. In this tutorial, you will learn how to perform simple and common tasks such as importing molecules, importing properties, viewing your molecules... A MySQL server must be installed on your machine or on a reachable network to go through this tutorial. See the [installations and requirements](#) section for all installation instructions. You may also want to read about the [important terminologies used by SA2](#) before starting this tutorial. At the end of this tutorial, you will be pointed to other sections of the documentation dealing with more specific functionalities of SA2.

Table of Contents

- A. [Introduction and data](#)
- B. [Starting SA2 and creating a new database](#)
- C. [User interface](#)
- D. [Importing new SDF files](#) (containing properties)
- E. [Importing CSV file](#)
- F. [Importing Fingerprint](#)
- G. [Basic visualisation](#)
- H. [A word on selection management](#)
- I. [Let's create new libraries !](#)
- J. [Conclusion and further readings](#)

Introduction and data - [top](#)

In this small tutorial, you will learn how to import molecules, visualise your molecules, and create new subsets of molecules ([Libraries](#)). We will use some SDF file that we have randomly extracted from a database of more than 6 millions molecules. Each molecule in these file has been standardized using a specific Pipeline Pilot protocol, and the 3D coordinates have been generated using Corina. We also provide you a set of MOE2D descriptors that we will use to illustrate the import of properties in the database. All the files used are located in the **sample-data/** directory, which is included in the SA archive since the 1.0.2b version.

- **sample-data/Provider1.sdf**: 100 molecules with MOE2D descriptors
- **sample-data/Provider1b.sdf**: 200 molecules with MOE2D descriptors
- **sample-data/Provider2.sdf**: 100 molecules without MOE2D descriptors
- **sample-data/Provider2_MOE2D.csv**: semi-column separated text file containing the MOE2D descriptors of molecules contained in the Provider2.sdf file.
- **sample-data/All_SSKey.csv**: semi-column separated text file containing a fingerprint that can be calculated / stored in SA2. It will be used to illustrate the import of fingerprints in the database.

SADB ID is the field corresponding to the identifier of each molecule.

Starting SA2 and creating a new database - [top](#)

OK lets start doing some SA2 stuffs. Run the executable binary of SA2 (located in \$SA/bin/), or use the shortcut installed if you used the automatic installer. **Note that you cannot run two (or more) instances of SA2.** After a few seconds, the following window should popup.

At this point you have to (1) connect to the MySQL server of your choice, and (2) create a new SA2 database. Let's detail a bit these two steps.

Connecting to the server

You can either connect to a local server (i.e. a server installed on the same computer as SA2) or to a server running on a different computer which can be reached by whatever network protocol. To connect to a server installed on your computer, enter "localhost" in the server text field. To connect to a distant server, enter a valid, reachable URL (e.g. an IP adress) in the server text field.

Once done, you will have to enter your user name and your password, that you had previously created just after the [installation of your MySQL server](#) in the appropriate fields. Click on the connect button to establish the connection.

Creating a new database

Once connected, a list of databases compatible with the current version of SA2 should appear in the "Existing database" table. In our case, this list should be empty, as this is the first time you've run SA2. Lets populate it then. Click on the "New" button.



In this window, you simply have to enter the name of the database and an optional description. You also have to choose which [handler](#) to use for your database (note that this choice is definitive). The name of the database must fit a specific pattern. It should not contain any space or special character; in such case, a warning message will be displayed as shown in the screenshot, and you won't be able to create your database until your name fits the pattern.

When you are done, click in the Finish button. The database will be created. Many operations will be performed, such as creating tables and inserting various pre-computed data (e.g. PCA-based reference chemical spaces - DRCS). One or two seconds will be needed for the database to be created, depending on how fast your computer is.

Once the database has been created, it will appear in the table. Select it, and click on the "Open" button to open it. Rendez-vous to the next section then.

User interface - [top](#)

Once you open a database, a set of windows opens automatically. When running SA2 for the first time, a default window organization (layout) will be setup. Note that you can completely reorganize your windows. Try to play around with opened windows (drag them, undock them using left click...) so you can get used to it and see all the possibilities offered by the windowing system. Here is an example of layout we often use in our lab:

A more detailed overview of the Graphical User Interface (GUI) can be found in the [dedicated section](#) of this documentation.

Importing SDF files - [top](#)

Let's import new molecules! Note that the full workflow describing what SA2 will do when importing a new molecule can be found in the [Import workflow](#) section of this documentation.

Click on the second button on the toolbar, or use File->Import SDF in the menubar. There are 4 configuration steps before starting the actual import process. Let's detail each of them.

Important note: we will import a SDF file containing properties here. If you are not interested in importing existing / new properties, you can skip the 4th step, thereby making the import process a bit faster.

1. Input file, properties and basic calculations

Steps

1. Input file & config.
2. Additional calculations
3. Provider
4. Properties

Input file & config.

Input file: /ince/SA2/sample-data/Provider1.sdf

ID field: SADB ID

CAS field: <None>

☒ Compute scaffolds & frameworks

☒ Compute Reactive flags

☒ Compute Warhead flags

☐ Compute PAIRS (<15) flags

☒ Compute PAIRS (<150) flags

☒ Compute PAIRS (>150) flags

< Back Next > Finish Cancel Help

1. Set the input file as being the **Provider1.sdf** file located in the *sample-data* directory.
2. Select the "SADB ID" field as the name field of each molecules (note that all other properties have been automatically detected as well). To do so, click on the selection drop-down box, and type S when the list of fields pops up. This should leave you directly to the right field, instead of scrolling the entire list of properties.

3. Leave the checkboxes just as is. In SA2, if you are not interested in either the HTS-related flags, or the scaffold / frameworks calculations (which are used for diversity analysis though!), just uncheck them; the import process will be faster then.
4. Click next (easy hu..!)

2. Providers / Libraries

During this step, you must inform SA2 on the origin of our compounds. If you are building a database dedicated to store chemical vendors collections, you will want to assign each collection to its dedicated provider. If you are importing a library that corresponds to a medicinal chemistry project, just create a new provider for this project. In our case, we will create a new dummy provider.

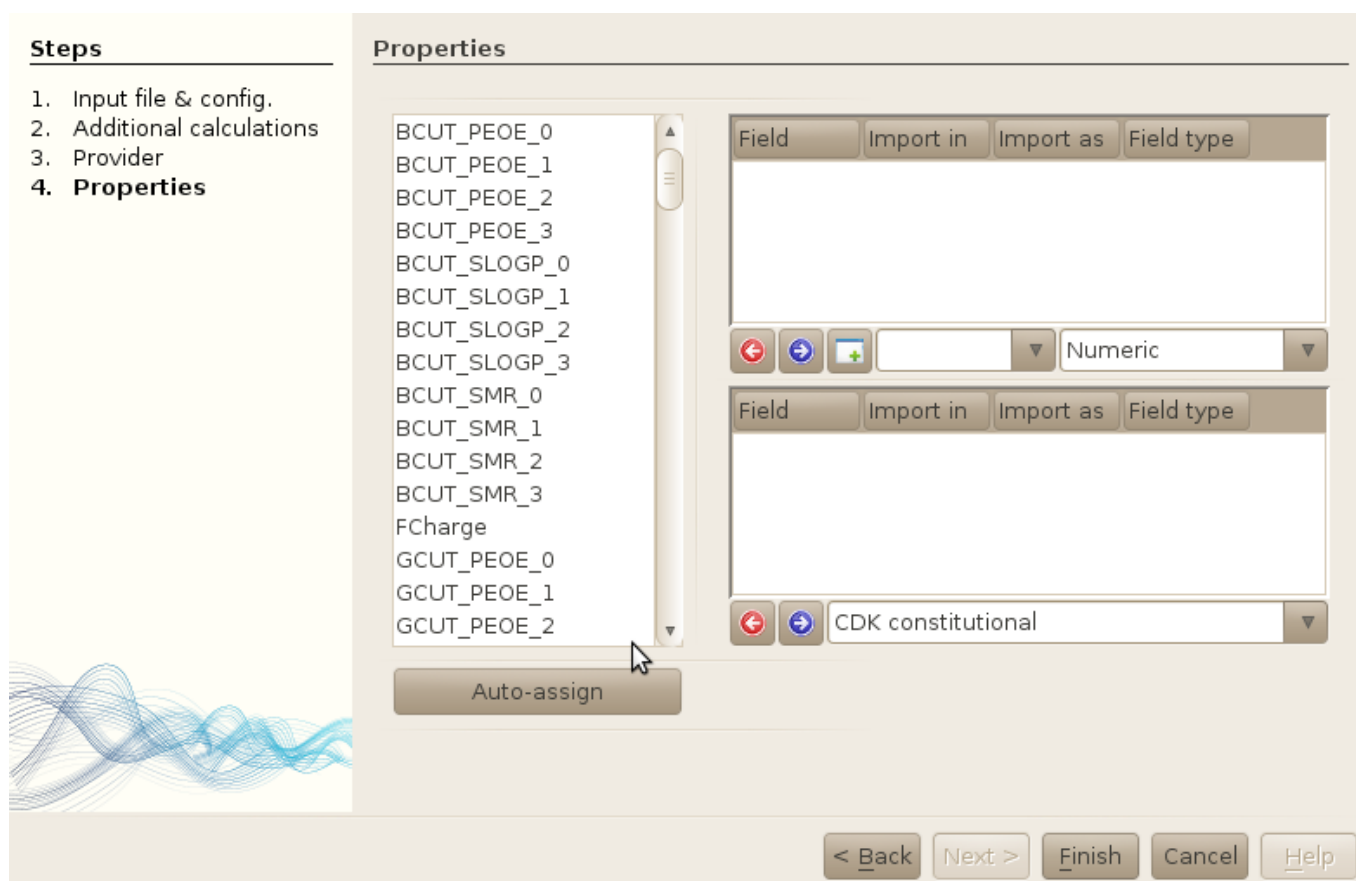
Note that during this step, you can also associate your molecules with a new or an existing libraries. [Libraries](#) are slightly different from [Providers](#): they represent subsets of molecules in the database, while providers represents the origine of molecules. Learn more about these simple (yet important) concepts in the dedicated section.

In this example, we will not create libraries.

1. Type "Provider 1" in the name field.
2. Leave the *Comments* text area empty (or write anything you want...)
3. Leave the Associate library option unchecked.
4. Click on the "Create" button. This should leave you to the previous screenshot. **Note that you can't reach the next step if you haven't selected an existing Provider or created a new one.**
5. Click next

3. Importing properties

Let's now import descriptors in the database. This step actually allows you to import properties available in your input file, in existing or new [Property Tables](#). In our case, we will import the MOE descriptors available in the input file, in the MOE table that is already available in SA2.



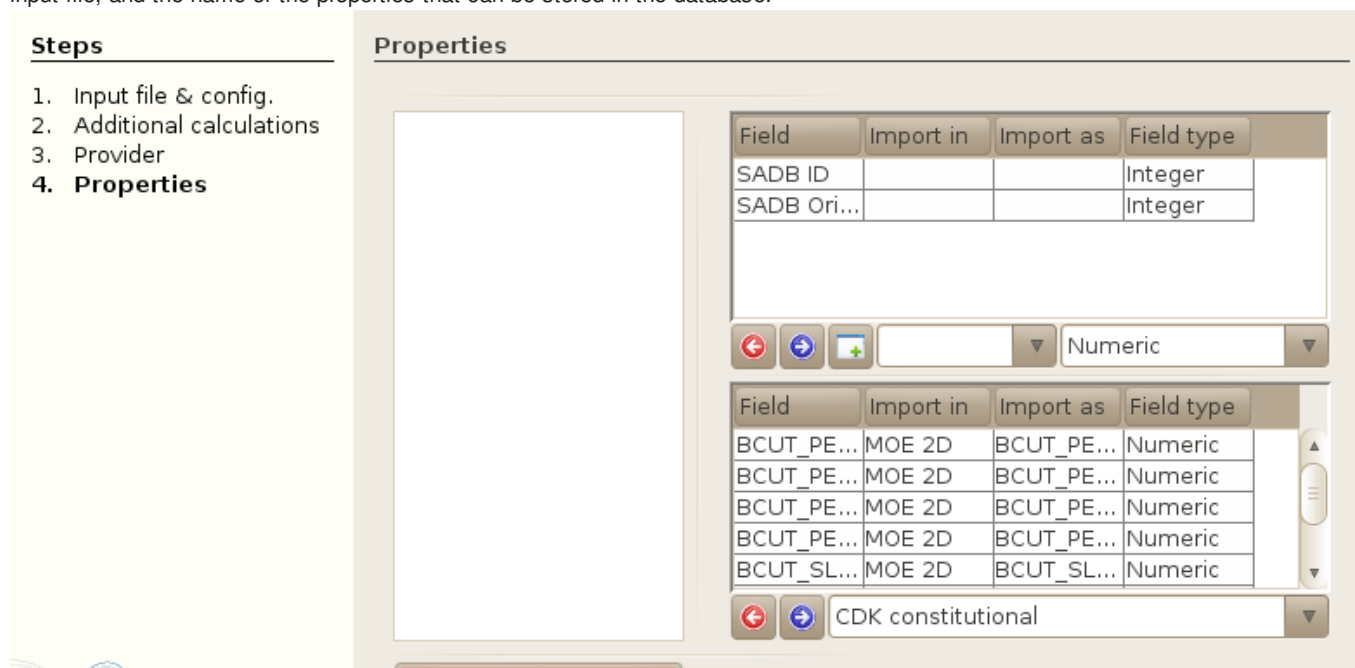
The window is splitted in 3 main parts: the left part is a simple list of properties that have been detected in the input file.

The right part is splitted in two more parts: on the top, you will find the properties that have been assigned to new tables, and on the bottom, you will find the list of properties that have been assigned to existing tables. In both parts, you can click on the arrow pointing to the left (red one) to remove a property from the import process, and on the right arrow (blue) to import a property in a new (top) or existing (bottom) table. The creation of new tables will be described later.

1. Click on any property in the property list. This will bring the focus on the list of properties.
2. Press CTRL and type A on your keyboard. This allows you to select all properties available in the list. Alternatively, select the first property, scroll down to the end of the list, press shift and click on the last property.

3. Assign properties

Click on the "Auto-assign" button. SA2 will automatically make the correspondance between the name of the properties in the input file, and the name of the properties that can be stored in the database.





Note: All unmatched properties will be assigned to the new table section, and you will have to either remove them from the import, or create a new table and assign this table to the properties in the top table.

4. Remove undesired properties

At this point, all properties have been assigned to either an existing field, or a new table (undefined). An error message is displayed to inform you that some properties have not been assigned. As you can see, two properties have been assigned to the new tables part of the window. We don't want to import these new properties, so select both of them, and click on the left arrow to remove them from the import process.

5. Check correspondance !



Everything seems fine now (the error message should not be here anymore), but actually it is not ! In the database, several properties might have the same user name ! This is the case for some MOE descriptors that have the same name as some SA1 / JOELib descriptors, e.g. the TPSA or Weight properties. Both properties have been assigned to the SA1 descriptors table, while we expect them to be associated with the MOE2D table. We could have changed this but we decided to keep it as a good illustrative example of things that you should pay attention to.

Let's ensure that all the properties have been assigned to the right fields. As we already know that all descriptors should be imported in the MOE2D table, we will assign all fields to this table in a single operation:

- Select all properties in the list of imported properties (do the CTRL+A stuffs).
- Click on the drop-down combo box right next to the blue arrow button, and select the "MOE2D" table.
- All the properties have been assigned to the MOE2D table.

Note: If a property was not assigned to any field available in the selected table, an error message would be displayed and you could not go further in the import process. In other words, you can't create new fields in an existing table during the import process :) (but you can do it using the [Properties window](#)).

Steps

1. Input file & config.
2. Additional calculations
3. Provider
- 4. Properties**

Properties

SADB ID
SADB Original ID

←
→
+

Field	Import in	Import as	Field type
BCUT_PE...	MOE 2D	BCUT_PE...	Numeric
BCUT_PE...	MOE 2D	BCUT_PE...	Numeric
BCUT_PE...	MOE 2D	BCUT_PE...	Numeric
BCUT_PE...	MOE 2D	BCUT_PE...	Numeric
BCUT_SL...	MOE 2D	BCUT_SL...	Numeric

←
→

MOE 2D

Auto-assign

< Back
Next >
Finish
Cancel
Help

4. Workers (additional calculations)

Last but not least, the additional calculation. If you haven't done yet, learn about workers in the [Terminologies](#) section or in the [specific section](#) where a detailed description of each available worker is provided.

Name	Use	Config
CDK descriptors	<input type="checkbox"/>	
CDK fingerprints	<input type="checkbox"/>	
Indigo fingerprints	<input type="checkbox"/>	
JOELib / SA1 (desc...)	<input type="checkbox"/>	

Here, just leave all workers unchecked. Note that each worker can be configured through the buttons located in the third column of this table. You will be able to select what the worker will do (which descriptors should be calculated...), and eventually set more specific parameters.

Click on Finish to start the import.

The import procedure should not be too long. An output window should open to inform you on the various steps and eventual errors detected, as well as a progress status bar located on the right bottom part of the main SA2 window.

Once the importation is finished, you may want to practice a bit, and repeat the process for the **Provider1b.sdf** input file. The only difference will be that you will assign to this input file the same provider as for the first imported file instead of creating a new one.

With a bit of practice, it takes me no more than 10 seconds to complete the 4 steps described previously.

Importing properties - [top](#)

We will now import another SDF file, but which does not contain any properties. Once done, we will import the corresponding MOE2D descriptors stored in a separate semi-colon separated text file.

1. Import the *sample-data/Provider2.sdf* in the database. Follow the steps described previously, but: (i) create a new provider (e.g. "Provider 2") for this file, and (ii) do not import any of the two properties. Also, make sure that you assign the "SADB ID" as the identifier of each molecule.
2. Once done, click on the Import properties button in the toolbar, or go to File->Import properties.
3. Select *sample-data/Provider2_MOE2D.csv* as the input file.

Steps

1. Input file & reader
2. Properties

Input file & reader

Input file:

Reader:

Fields detected

SADB ID;SADB Orig

Reader parameters

Parameter	Value
Delimiter	,
Quote	"

ID field:

Corresponds to:

☒ Update values if exists

- Once done, you will see that a default reader (CSV reader) will be assigned. However, the reader used a comma as a separator, and we want it to be a semi-column. Change this in the *Reader parameters* table. The list of detected fields should be properly reloaded now.
- Select SADB ID as the ID field, and select "Original ID" as the corresponding field in the database. See the [IDs section](#) of the documentation to know more about compounds identifiers in SA2.
- Leave the "Update values if exists" checkbox checked (it does not matter here, but you can choose not to update the properties if they are already stored in the database).

Steps

1. Input file & reader
2. Properties

Input file & reader

Input file:

Reader:

Fields detected

BCUT_PEOE_0
BCUT_PEOE_1
BCUT_PEOE_2

Reader parameters

Parameter	Value
Delimiter	;
Quote	"

ID field:

Corresponds to:

☒ Update values if exists

- Click next
- Import properties just as described previously. The user interface is exactly the same compared to the SDF import, so there is nothing new here to explain.
- Click finish to start the import of all properties.

Importing fingerprints - [top](#)

We will now import the values of a fingerprint available in the database: the SSKey fingerprint that was available in SA1 (and is still available in SA2). As mentioned previously, this fingerprint can be directly calculated using the [JOELib worker](#). As you will see, importing fingerprints is quite similar to importing properties. The main difference is that you can't create new storage capability for a new fingerprint during this process.

1. **File->Import fingerprints**
2. Browse your local drive and select the file *sample-data/All_SSKey.csv*
3. Update the parameters of the CSV reader to use semi-column as separator.
4. Select the ID field, and set the **Corresponds to** field to 'Original ID'.

Steps

1. Input file & reader
2. Fingerprints to import

Input file & reader

Input file: .0/sample-data/All_SSKey.csv

Reader: CSV reader

Fields detected: SA1 Fingerprint (JOELib), SADB Original ID

Reader parameters:

Parameter	Value
Delimiter	;
Quote	"

ID field: SADB Original ID

Corresponds to: Original ID

☒ Update values if exists

< Back Next > Finish Cancel Help

5. Click next.
6. Select the SA1 fingerprint (JOELib) and press on the blue arrow. The fingerprint will automatically be associated with the corresponding fingerprint in the database.

Note: if no correspondance was found (e.g. a different name was used in the input file), you can force the fingerprint to be imported in the table. If there is actually no fingerprint corresponding to the input file in the database, you will have to create the fingerprint table before importing the data. See the [Properties](#) section of this documentation for more information.

Steps

1. Input file & reader
2. **Fingerprints to import**

Fingerprints to import

SADB Original ID

Field	Import in ▲	FP. Size	Encoding
SA1 Fing...	SA1 Fingerpr...	54	Binary (1 and 0)

Import (auto)

< Back

Next >

Finish

Cancel

Help

7. Finish.

You fingerprints will be imported, and you will now be able to use them for e.g. similarity searching or diverse subset creation.

Lets now take a closer look at our compounds and properties.

Basic visualisation - [top](#)

We will now describe a very straightforward way of viewing our compounds. Before doing so, let's ensure that the appropriate windows are opened. Most of these windows should already be opened if you read the documentation about [setting up a better default layout](#) before running SA2 for the first time.

1. Open the following viewers, which are intended to represent different information on a single molecule selected in other views. Click in each link to have a more detailed description of what is shown in each view, but the basic name is enough to get a quick understanding of what will be displayed:

- [Window->Molecule->2D structure](#)
- [Window->Molecule->Scaffold](#)
- [Window->Molecule->Framework](#)
- [Window->Molecule->Properties](#)
- [Window->Molecule->List of IDs](#)

OK that's a lot of windows; remember that you can put each window pretty much anywhere you want (and even don't open them).

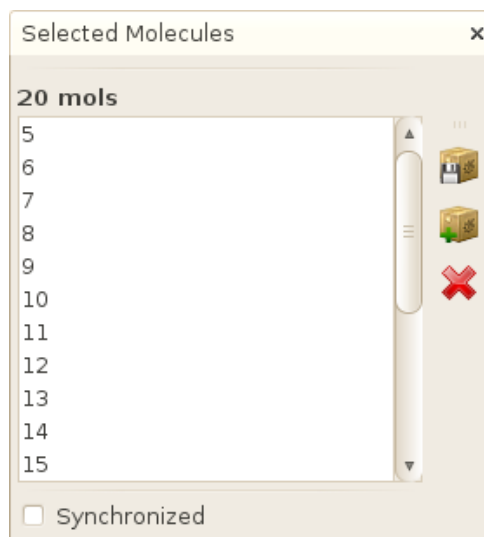
2. Open the following viewers, which will display molecules as simple lists along with various properties:

- [Window->MoleculeS->Basic table](#)
- [Window->MoleculeS->Flags table](#)
- [Window->Molecule->Selected molecules](#)

3. All these windows will open in the main centered area of the interface. You can now play around with each viewer. When you select a single molecule, the content of each so-called "single view" will be updated to display the information (structure / scaffold / properties / whatever...) associated with the selected molecule. Conversely, the ID(s) of the selected molecule(s) will systematically appear in the **Selection window** when a molecule is selected in whatever view allowing to perform selection. In the next sections, we will learn a bit more about this **Selection window**, and we will then create new libraries based on selected molecules.

A word on selection management - [top](#)

In simple table as well as in various plotting facilities, you have the possibility to select interactively one or several molecules. When doing so, the full list of selected molecules will appear in the **Selection window**, usually located in the right of the main SA2 window. You can subsequently perform various operations on these selected molecules using the vertical toolbar located on the right of the window.



As you will see, you can also synchronize the selection between the different views by checking the **Synchronized** checkbox. This way, when you select one or several molecules on either a table, or a plot, all opened views will be updated to select the same molecules (if available !).

Learn more about this view in the [GUI section](#) of this documentation.

Let's create new libraries ! - [top](#)

We will now create new libraries. For the recall, Libraries in SA2 represents subsets of molecules. We will illustrate this point using three simple approaches: (1) create a library based on selected molecules, (2) create a library using simple filtering rules, and (3) create a library grouping molecules that have a common scaffold (or framework). Note that there are other ways of creating libraries, e.g. by merging existing libraries, by complementing existing library, by using diversity algorithm etc.

(1) Using selection to create a new library

A simple way of creating a selection is to use the **Selection window**. Let's do this by creating a new selection containing all fragment (RO3 compliant) molecules.

1. Open the **Flags table** view: *Window->Molecules->Flags table*
2. Click on the RO3 column (that should be the second one) to order compounds based on this flag. You may have to click twice to obtain the descending ordering (RO3 compounds will be at the top of the table). The compounds that satisfy the RO3 should be coloured in blue.
3. Select all these compounds.
4. In the **Selection window**, you should see the database IDs of all your selected molecules. Click on the first button on the left of the window (pass your mouse over each button to see what they allow you to do, and select the one that allows to create a new library based on selected molecules).



If you select any of the row in the Selection window, the effect of each action available through the buttons



located on the right will apply to the selected molecules! In other words, you can select a subset of the selected molecules to save them in the database. If you want to save the entire selection, you must not select any row in the list (or you can select all of them).

5. Enter a name for your new library and press OK / finish

Your library has been saved. It should now be visible in various windows, including the [List of libraries](#) window, and in all other views (Flags table...) that allow you to view only one particular library.

(2) Creating filtered library

Let's now create a filtered library. We will create the exact same library as previously, but using a smarter way. Indeed, you probably noticed that the previous process is OK when you have only a few molecules, but becomes quite boring if you deal with a large database. Moreover, you don't want to use the sort capability of these table for large database. Let's makes the process a bit more automatic then.

1. **Compute->Library->Add filtered library**

Steps

1. **Name & description**
2. Property filter(s)
3. Providers

Name & description

Name:

Description:

Restrict to:

< Back Next > Finish Cancel Help

2. Type 'Fragment library (2)' as the name of your new library. Leave the "Restrict to" dropdown box as is. (changing this option, you can create a new filtered library based on an existing library !)
3. Click next.
4. Expand the property tree to find the RO3 property in the **Basic properties** table located at the root of the tree.
5. Select this property, and click on the blue arrow in the middle of the window
6. Enter the desired value for this property (we want fragments -> RO3 = 1)

Steps

1. Name & description
- 2. Property filter(s)**
3. Providers

Property filter(s)

Nodes

- PAINS (<150)
- PAINS (>150)
- Reactive
- RO3 (fragm...**
- Rot. bond co...
- SSSR count
- Max. ring size
- TPSA

→
✗

Prop.	Operator	Filter val.
RO3 (fra...	=	1

Logical operator

AND

< Back

Next >

Finish

Cancel

Help

7. Click next.
8. Leave the last step just as is. You also have the possibility to restrict the search to only one or several providers.

Steps

1. Name & description
2. Property filter(s)
- 3. Providers**

Providers

Unselect all

Select all

Provider	Include in library
Provider 1	<input checked="" type="checkbox"/>
Provider 2	<input checked="" type="checkbox"/>

< Back

Next >

Finish

Cancel

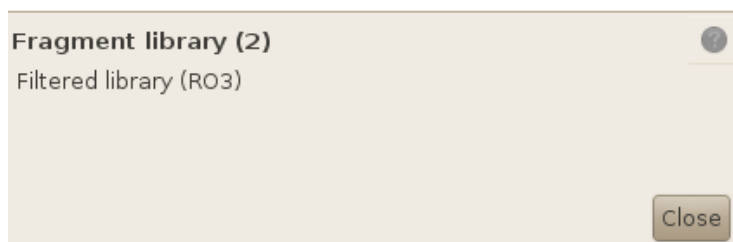
Help

9. Click on finish.

Your new library has been created.

If you are not convinced, open the [List of libraries](#) window. You should see your two libraries. Left click on each of them, and select **Properties**. You should see that they both contain the same number of compounds. This properties windows is available for most database entities (Providers, Libraries, Properties / Tables...), and must be used if you want to change the name / description of one particular entity, or see some interesting properties (e.g. the % of explained variance of each component for a DRCS model).

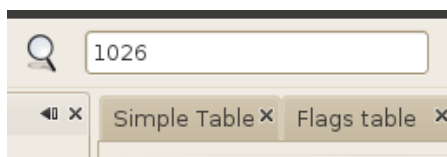
▼ Properties	
Name	Fragment library (2)
Description	Filtered library (RO3)
# of molecules	35



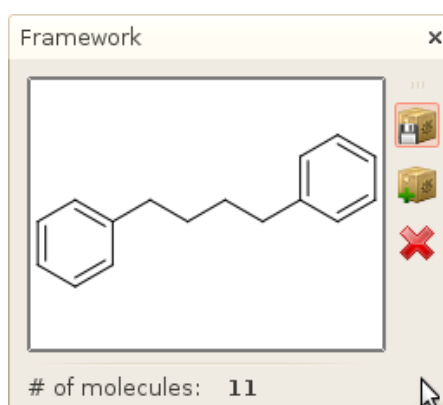
(3) Creating framework-based library

Let's finally create a library containing all molecules that belong to a particular framework.

1. If not already opened, open the Framework window that will show you the framework of the selected molecule: [Window->Molecule->Framework](#)
2. Next, in the search bar (toolbar on the top), type the following number: '1026', and press enter.



3. In the list of results, you should see one molecule. Click on this molecule in the search results window. The framework of this molecule should appear in the framework view, showing that 11 molecules are associated with this particular framework.



4. Click on the Save as new library button (the first one)
5. Enter whatever name you want, and save your library.

Go back in any of the table previously opened, and select the checkbox named "Lib". Select your newly created library and you should now see only the molecules contained by your library and having the same scaffold.

Conclusion - [top](#)

You've learned how to perform basic operations within SA2. There are plenty of other things that you can do with the software. The documentation is not completely exhaustive yet, but be patient, it will get improved with time. Here is a subset of interesting pages you may want to read:

- Learn how to [plot your molecules in 2D](#)
- Learn how to [search for molecules](#) in SA2 (similarity, substructure...)
- Learn how to [derive simple statistics](#) such as comparing property distributions between libraries...
- [Frequently asked questions](#)

[Go To Table of Contents](#)

