

# ScreeningAssistant: a Free Software for Managing Chemical Databases



Aurélien Monge, Alban Arrault, Christophe Marot et Luc Morin-Allory

Institut de Chimie Organique et Analytique, UMR CNRS 6005,  
Université d'Orléans BP 6759, 45067 ORLEANS Cedex 2, France.

[aurelien.monge@univ-orleans.fr](mailto:aurelien.monge@univ-orleans.fr)

<http://screenassistant.sourceforge.net/>



Managing a database of screening compounds is a tedious task. *ScreeningAssistant* is a platform dedicated to make this operation easier. It can use the SDF libraries provided by the suppliers to create a chemical database. The software is designed to insert only the new chemical structures, but keeping the references of all the providers for a given compound. Analyses of the libraries, using molecular frameworks, scaffolds, diversity, and several physicochemical properties are available. Furthermore, the compounds can be represented in a 2D chemical space.

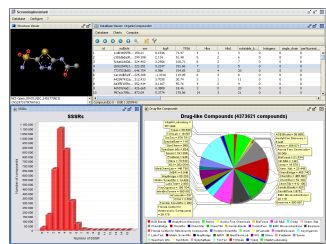
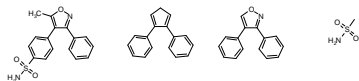
Technically, this platform is based on free tools allowing any research team to use it. Furthermore, the software is provided with the source code under GPL license, allowing to be improved by other developers. However, interface to some commercial softwares is provided, allowing *ScreeningAssistant* to access to other functionalities such as improved molecular display (Marvin), 2D to 3D conversion (Corina), and conformers generations (Omega).

## Software

*ScreeningAssistant* is developed in Java. It is based on JOELib [1], a cheminformatics library under the GPL license, and MySQL, an open source database. InChI [2] is used to generate unique code for the structures and to remove duplicates. *ScreeningAssistant* can use Corina and Omega to generate 3D structures and conformers. Automatic charts generation allows to easily analyze the databases.

- Suppliers' names
- Descriptors: PDL, PLL, MW, LogP, TPSA, HBA, HBD...
- Fingerprints: 001101100001...
- Structure, frameworks, scaffolds and side chains

MySQL



## Filtering

- Progressive "Drug-Like" (PDL) and Progressive "Lead-Like" (PLL)

"Progressive Drug-Like" (PDL) and "Progressive Lead-Like" (PLL) scores [3] based on progressive limits on physicochemical properties (molecular weight, logP, H bond acceptors, H bond donors, rotatable bonds, SSSR, maximum ring size and halogens) are computed for all the compounds. On the basis of these scores, a new personalized score, "Cleaning For My Screening" (CFMS), can be computed. Depending on the choice of the user, reactive functions, warheads agents, promiscuous aggregating inhibitors, single chains, perfluorinated chain, absence of N or O ... can be used – or not - to add additional penalties to the compounds. CFMS can be used, in combination with filters for any other descriptors computed by *ScreeningAssistant*, to select compounds for real or virtual screenings.

- Frequent hitters

Reactive functions (cause false positives by covalent binding) and warheads (can cause false positives by no covalent binding) are detected. The user can edit/add substructures in these two lists.

In addition, the system is able to identify 48 known promiscuous aggregating inhibitors.

- Privileged structures

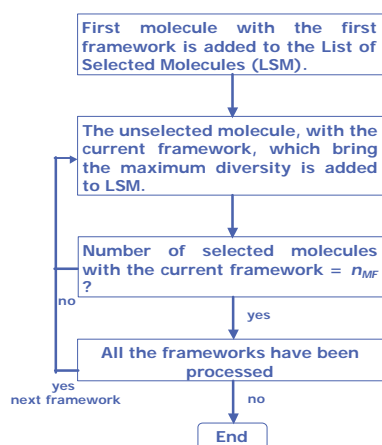
A privileged structure is a molecular scaffold able to provide good ligands for different biological targets.

These scaffolds are known to have good drug-like properties. Privileged structures can be used to design focused libraries (GPCR, kinases...). Another approach is to select compounds without privileged structure in order to get original compounds with a probable good selectivity.

## Diversity

The first criteria of the diversity algorithm developed for *ScreeningAssistant* is the proportional repartition of the frameworks. It is important to have a good diversity of frameworks in the selection, because it is very interesting to get hits with different scaffolds. Then, the compounds for a given framework are selected according to the SKey3DS fingerprints. The diversity of compounds can be visualized in a 2D factorial space.

This algorithm was successfully tested to select 500 000 compounds.

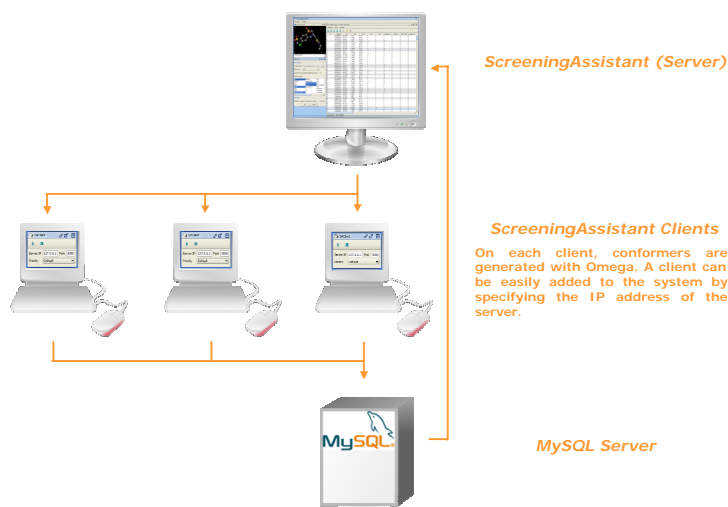


$$n_{MF} = \text{nb of molecule with the framework} \times \frac{\text{nb of compounds to select}}{\text{nb of compounds in the database}}$$

## Distributed Conformers Generation

In the next version, *ScreeningAssistant* will use Omega to generate conformers. As our system is dedicated to manage huge chemical databases, the computational time for conformers generation can be high. We have developed a grid-computing system based on Java's sockets. With this system, *ScreeningAssistant* can generate the conformers using several Windows computers.

The data of the computations are stored in the MySQL database. In consequence, any computer of the platform can crash without losing any information.



## Perspectives

*ScreeningAssistant* is freely available under the GPL license. It is able to manage successfully the virtual chemical database of our laboratory containing 5 million compounds. It is used to conceive screening sets using physico-chemical properties, frequent hitters filters and diversity. The software was also used to conceive databases for HTS projects with very good results.

In the next version of *ScreeningAssistant*, prediction of mutagenicity based on toxicophores will be added. Other possible developments can be the addition of *de-novo* design functionalities based on Genetic Algorithms and RECAP retrosynthetic rules. An other field to develop can be the addition of QSAR functionalities.

1 - Wegner, J.K., JOELib, <http://joelib.sourceforge.net/>

2 - The IUPAC International Chemical Identifier. <http://www.iupac.org/inchi/>

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. Divers.* 2006, in press.