Screening Library Enrichment : Criteria that matter, timely manner

Jérôme Amaudrut, Fabrice Guillier

Inventiva, 50 rue de Dijon, 21121 Daix, France (www.inventivapharma.com)

1 - INTRODUCTION

The acquisition of new screening molecules is of key importance for any pharmaceutical company. These compounds are the potential hits and starting point of our future research programs, hence they should ideally have all the qualities required to enter the hit to lead phase. Here we describe the selection process that was used for our latest corporate screening collection (IVALib) enrichment campaign from several vendor's catalogues. The key criteria were **chemical diversity** and **complementarity** to our collection as well as **3D shape** and **molecular complexity**. The desired modulation of the overall properties of the library was achieved with a relatively modest increase in size (9%).

2 - DESCRIPTION OF THE PROCESS

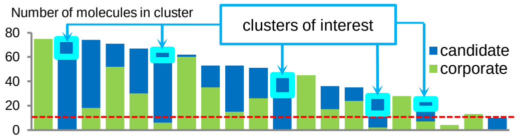
Each vendor collection goes through the following steps:

-**Filtering**

Compounds are excluded from selection based on simple criteria such as high MW, high logP or the presence of an unwanted chemical motif. Over 1000 substructures defined in-house or in the litterature[1,2] are included in this filter.

-**Clustering**

A virtual library composed by the reunion of both vendor's collection and IVALib is clustered using FCFP_4 fingerprint and the "cluster molecules" component of Pipeline Pilot.[3] Special care is taken to perform this step with no *a-priori* knowledge of the final number of clusters, reducing the risks of fingerprint collisions and remaining time efficient. Clusters containing too many molecules from our corporate collection are excluded.



-**Scoring**

Each molecule is given a score representing its desirability in terms of MW, logP, complexity (measured as the FCFP_4 size)[4] and 3D shape (see part 3). To each criteria is associated a score in the range [0-1] that can easily be given more/less weight in the final score calculation.

-**Ranking**

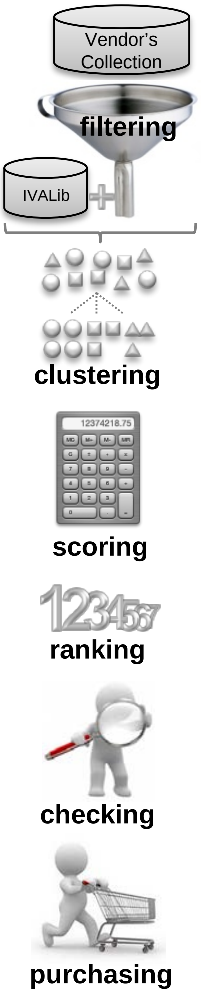
First, only the top 10 molecules of each cluster are considered. Clusters are scored as the average value of these top 10 members. Complete ranking of all the molecules is achieved by sorting by cluster scores and then by molecule scores.

-**Checking**

All molecules selected in silico are reviewed by the medicinal chemist's eye and a further 1/3 are excluded. This shows that there is still room for improvement in the definition of the filters.

-**Purchasing**

Having an open-ended list of molecules ranked by desirability is very efficient at this stage as it allows dealing with out-of-stock molecules by simply moving on to the next best one from the list.



3 - ACCOUNTING FOR THE 3D SHAPE

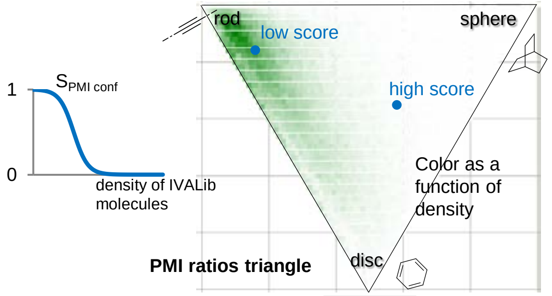
The 3D shape of molecules is recognized as a key parameter for drug likeness in general and its effect on aqueous solubility in particular.[5]

We selected the Principal Moments of Inertia (PMI)[6] to represent the 3D shape. A caveat is that this method is very sensitive to the conformation used. For example, linear alkyl chains are considered as needle like in their lowest energy conformation.

To circumvent this issue we calculate a desirability score based on the 3D shape with the following method: we generate up to 15 conformers with their Boltzmann population (calculated from their relative energies). PMI are calculated for each conformer which then in turn gets a score depending on the density of population of corporate molecules in that area of the PMI ratios triangle. The score for a molecule is the sum of all its conformers scores weighted by their Boltzmann population.

$$S_{PMImol} = \sum_{conf} Bpop_{conf} \times S_{PMIconf}$$

Using this description of the shape requires longer calculation times considering the 3D conformers generation step but thanks to Omega[7] (fast, distributable on many cpus), this is not a road block. Biases from 2D descriptors are avoided with this approach which provides a clear way of "escape from flatland"[5].

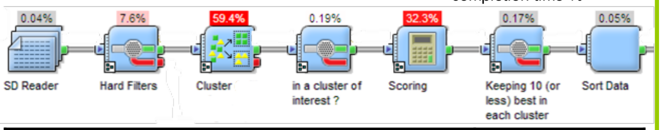


4 - RESULTS

All steps until the final visual check are automated in a Pipeline Pilot protocol that ran in 2013.

Clustering is the most time consuming step.

completion time %

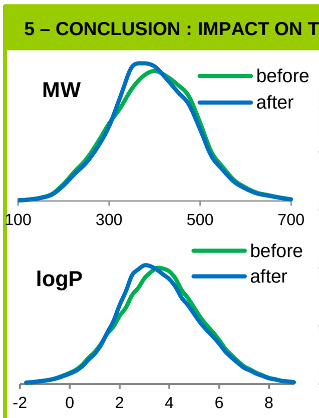


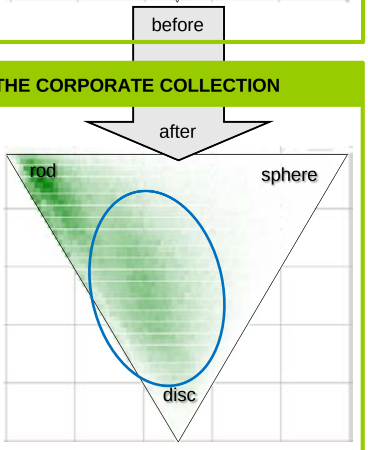
Vendors' collections	5
Total number of molecules in input	~4 700 000
Number of molecules clustered (5 independent runs)	~4 900 000
Number of molecules for which to generate 3D conformers	~3 900 000
CPUs in linux cluster (3D conformers generation)	16
Total protocol running time	167 hours (< 7 days)

REFENCES

1) Baell J. B. et al.; *J. Med. Chem.* **2010**, *53*, 2719–2740.
2) Bruns R. F. et al.; *J. Med. Chem.* **2012**, *55*, 9763−9772.
3) Pipeline Pilot, v8.5 Accelrys Software Inc.
4) Schuffenhauer, et al ; *J. Chem. Inf. Model.*, **2006**, *46*, 525-535.
5) Lovering, F. et al. ; *J. Med. Chem.* **2009**, *52*, 6752–6756.
6) Sauer and Schwarz, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987-1003.
7) OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM.
8) Bemis G. W. et al. ; *J. Med. Chem.* **1996**, *39,* 2887-2893.

5 – CONCLUSION : IMPACT ON THE CORPORATE COLLECTION





Chemotypes	Number of Bemis Murcko Ring Assemblies[8]	Number of clusters (FCFP_4, Maxsim = 0.5)
Before	67237	22289
After	80117	24453

Inventiva's screening library IVALib (~218K compounds) was enriched by 9%. The impact on the 3D shape, physicochemical properties profiles and number of chemotypes is obvious. Follow-up enrichment campaigns focusing on specific themes are ongoing with similar methodology.