

Bionet Diversity Sets

Key Organics offer sets of 5000 and 2000 compounds as 1mg dry powders that sample the full range of diversity present within our complete screening catalogue of just under 54,000 compounds. In addition these sets have physicochemical properties favourable for use in drug research i.e. filtered to remove various reactive and/or toxic substructures and approximately 90% satisfy computational leadlike criteria.

Methods

Initially, compounds were filtered to ensure they satisfied the majority or all of the computational leadlike criteria:

- MW \leq 460
- ClogP \leq 4.2
- Rotatable bonds \leq 10
- Rings \leq 4
- H bond donor \leq 5
- H bond acceptor \leq 9

This database was then filtered by substructure searches for the following groups:

Carbazides	Acyl cyanides	Michael acceptor
Acid anhydrides	Sulfonyl cyanides	Beta halo carbonyl
Pentafluorophenyl esters	Cyanophosphonates	Peroxide
Paranitrophenyl esters	Azocyanamides	Phosphonic acid
Hydroxybenzotriazole esters	Azoalkanes	Phosphonic ester
Triflates	Polyenes	Phosphoric acid
Lawesson's reagent	Saponin derivatives	Phosphoric ester
Phosphoramides	Cytochalasin derivatives	Sulfonic acid
Aromatic azides	Cycloheximide derivatives	Sulfonic ester
Beta carbonyl quart nitrogen	Monensin derivatives	Triphenyl phosphene
Acylhydrazide	Acid halide	Unbranched chain (>4 atoms)
Quarternary C, Cl, I, P, or S	Cyanidin derivatives	Epoxide
Phosphoranes	Squalestatin derivatives	Hetero hetero
Chloramidines	Aldehyde	Sulfonyl halide
Nitroso	Alkyl halide	Halopyrimidine
P S Halides	Anhydride	Perhalo ketone
Carbodiimide	Azide	Methyl ketone
Isonitrile	Azo	Aziridine
Triacyloxime	Di-peptide	Imine
Cyanohydrins	Long aliphatic chain (>7 atoms)	Oxalyl

In addition, compounds were assessed for reactivity using the Oprea set (Oprea 2000), as implemented in MOE2005.06, and duplicate entries were removed from the database.

Finally, the solubility of the compounds was predicted using a linear atom contribution method (based on Hou 2004). The predicted solubilities range from -6.0 to 2.0 mol/L. Various molecular descriptors were then calculated based on the Drug-Likeness Index from Xu and Stephenson (Xu 2000). These are 25 2D physicochemical properties including the number of atoms, donors, acceptors, rotatable bonds, along with various cyclicity measures

and molecular graph descriptors. These 25 descriptors were reduced to 5 principle components (PCs), which were then used to cluster the database via calculation of binned non-parametric distances in this 5D space. This was performed using the QuaSAR-Cluster function (Labute) in MOE. This analysis resulted in 13042 clusters. A diverse subset (on the basis of the DLI descriptors) of the 21 thousand compounds was chosen by selecting one compound from each cluster. By plotting the first three principal components on a graph, the set of selected compounds can be visualised (figure 1).

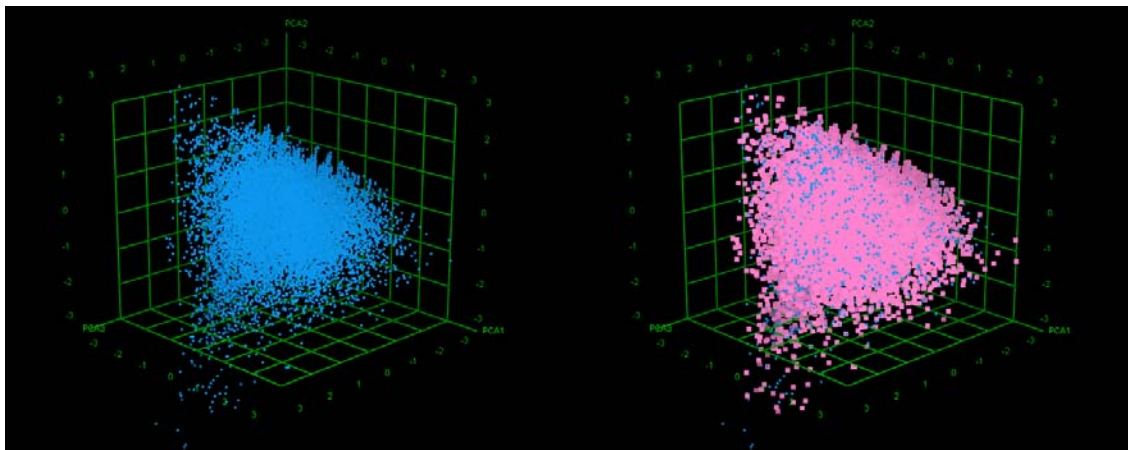
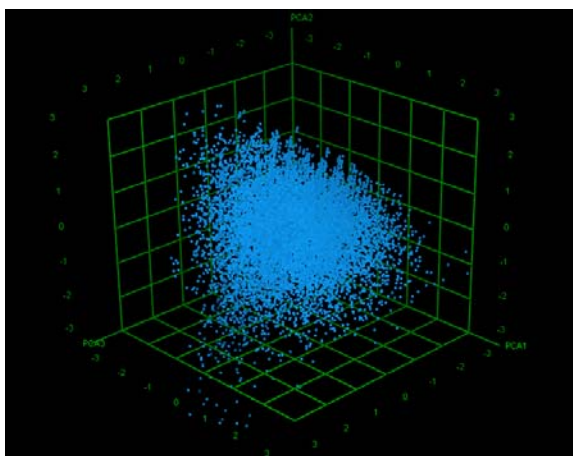


Figure 1. 21k set (left) and the 13k set highlighted (right).

To reduce these 13 thousand compounds to collections of the desired size, diverse subsets were chosen on the basis of the chemical groups that comprise the compounds. This was achieved by calculating bit-packed MACCS key fingerprints for each compound, then choosing a diverse subset based on Euclidean distance between the compounds in this fingerprint space. The diverse libraries were then selected by choosing the first five and two thousand most “distant” compounds. Again these selections can be visualised in PC space (figure 2).



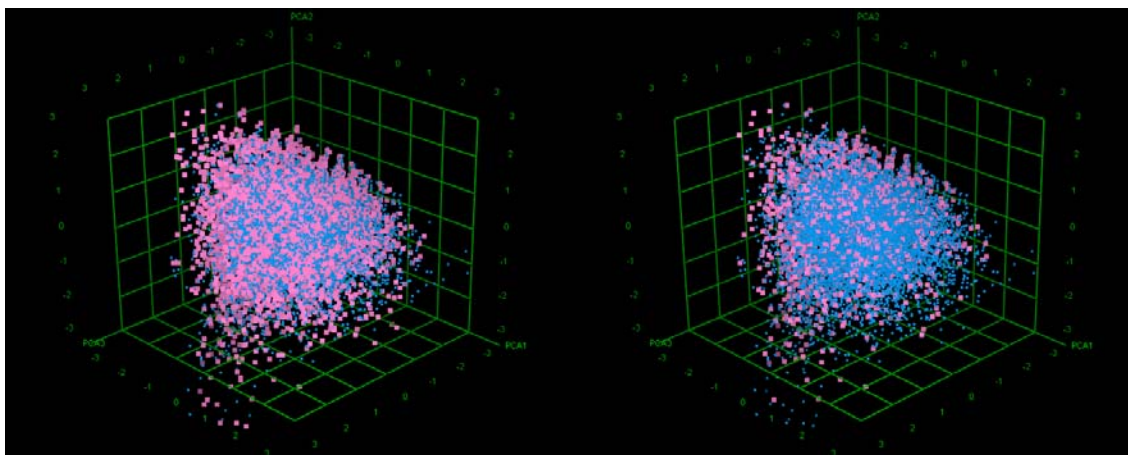
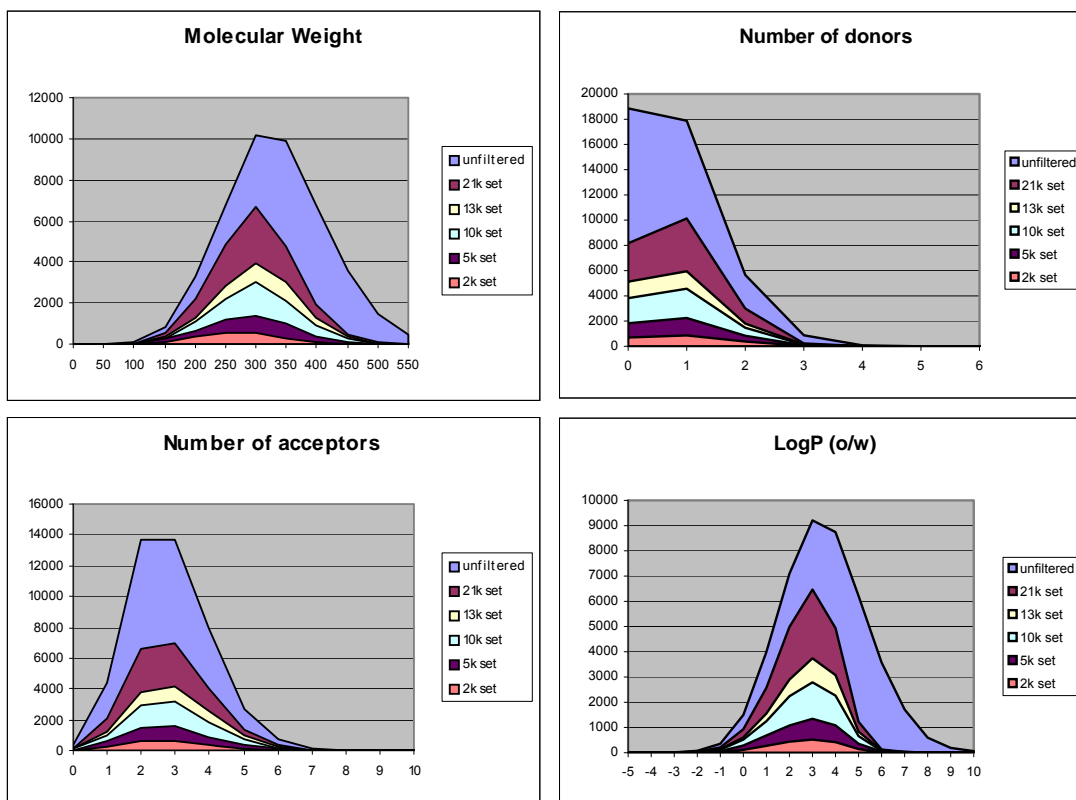


Figure 2. The 13k set (top left), with the 5k (bottom left) and 2 k (bottom right) highlighted.

In essence:

- The first stage removed undesirable substructures
- The second stage selected for a maximally diverse set of physicochemical properties
- The final stage selected for a maximally diverse set of chemical substructures (fingerprints).

Analysis



References

Hou, T.J., Xia, K., Zhang, W., Xu, X.J.; ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach; *J. Chem. Inf. Comput. Sci.* **44**, 266-275 (2004).

Oprea, Tudor I. Property Distribution of Drug-Related Chemical Databases. *J. Comp. Aid. Mol. Des.* **14**, 251-264 (2000).

Xu J & Stevenson J. Drug-like index: a new approach to measure drug-like compounds and their diversity *J Chem Inf Comput Sci.* **40**:1177-1187 (2000)

P. Labute. Quasar-Cluster: A Different View Of Molecular Clustering,
http://www.chemcomp.com/Journal_of_CCG/Articles/cluster.htm