

# Comparison of large chemical libraries using unsupervised classification techniques

Alexander Böcker

Evotec AG, Schnackenburgallee 114, 22525 Hamburg, Germany

## Introduction

Evotec is a global service provider along the drug discovery value chain progressing new chemical entities identified during high-throughput screening (HTS) to clinical candidate nomination. A carefully selected maximum diverse library of more than 250,000 drug-like compounds is readily available for screening. Beside commercially available compounds, the Evotec Lead Discovery Library comprises a significant fraction of Evotec exclusive compounds (about 30%). The library also includes several target-focused libraries.

A recurring question is how the Evotec Lead Discovery Library compares to other screening collections with respect to physicochemical properties and chemical novelty. To address this question the Evotec Lead Discovery Library was compared to the MDDR library (N = 190K)<sup>1</sup> using factor analysis, principle component analysis, hierarchical clustering<sup>2</sup> and self-organising maps<sup>3</sup> in combination with MOE physicochemical descriptors,<sup>4</sup> EvoCATS pharmacophore descriptors and chemical fingerprints.<sup>5</sup>

## Comparison based on key physicochemical properties

Key physicochemical properties were calculated in MOE for the molecules in the MDDR library and the Evotec Lead Discovery Library. Extreme values obtained for large molecules in the MDDR library were discarded ahead of generating the visualisations.

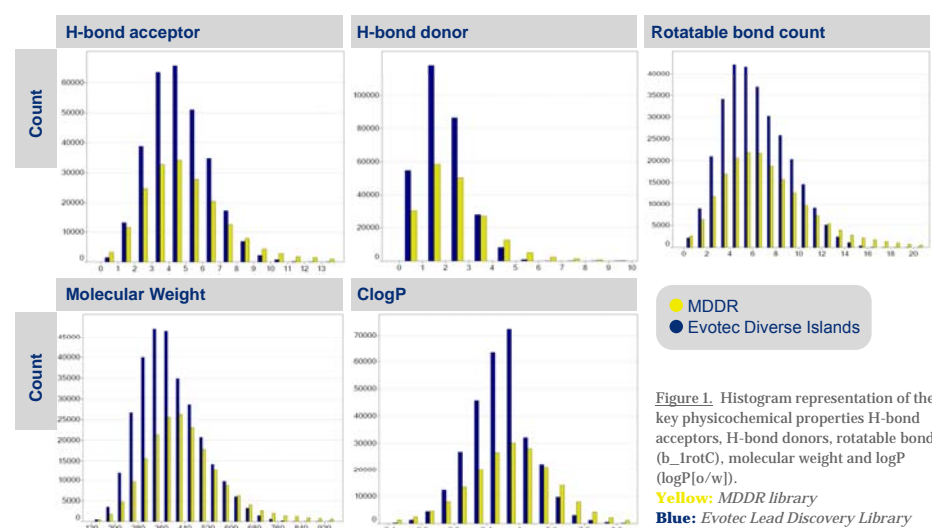


Figure 1. Histogram representation of the key physicochemical properties H-bond acceptors, H-bond donors, rotatable bonds (b\_rotC), molecular weight and logP (logP<sub>o/w</sub>). Yellow: MDDR library Blue: Evotec Lead Discovery Library

Box plot comparisons were performed in Spotfire. The statistical significance of the difference between both libraries was determined using the comparison circle algorithm.<sup>4</sup>

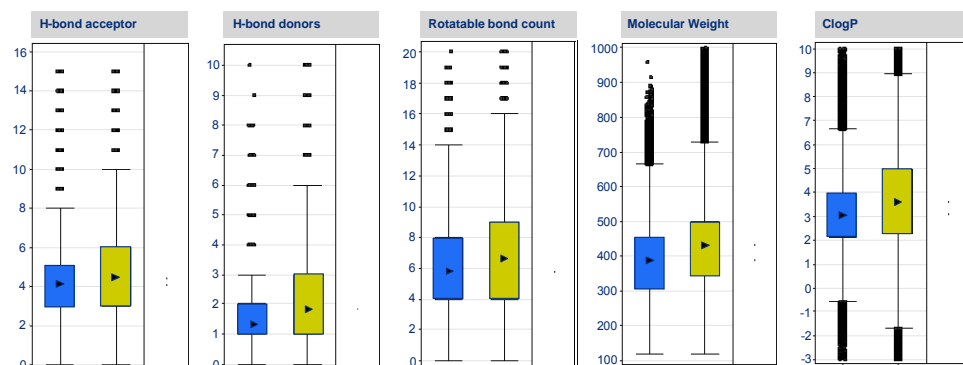


Figure 2. Box plot comparison of the key physicochemical properties H-bond acceptors, H-bond donors, rotatable bonds (b\_rotC), molecular weight and logP (logP<sub>o/w</sub>). Yellow: MDDR library Blue: Evotec Lead Discovery Library. Comparison circles are depicted on the right side of each plot.

- Evotec Lead Discovery Library and MDDR with similar physicochemical property profile
- Physicochemical properties significantly lower in the mean for the Evotec Lead Discovery Library

## PCA comparison using physicochemical and pharmacophore descriptors

MOE2D physicochemical descriptors (N=156) and EvoCATS descriptors (N=210) were calculated for the Evotec Lead Discovery Library and the MDDR library. Descriptors with low variance ( $\sigma < 0.005$ ) and highly correlated descriptors ( $R^2 > 0.99$ ) were discarded by unsupervised forward selection.<sup>6</sup> 201 EvoCATS descriptors and 76 physicochemical descriptors remained. Descriptors were centered to the mean and scaled to unit-variance. Principle component analyses and visualisations were performed in Spotfire. A focus was laid on the first 2 principle components explaining 72% of the variance for the EvoCATS descriptors and 74% for the physicochemical descriptors.

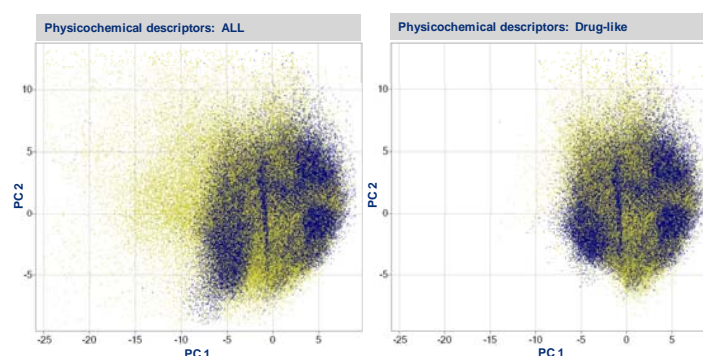


Figure 3. Scatter plot representation of the first 2 principle components obtained for the MOE2D physicochemical descriptors. Yellow: MDDR library Blue: Evotec lead discovery library

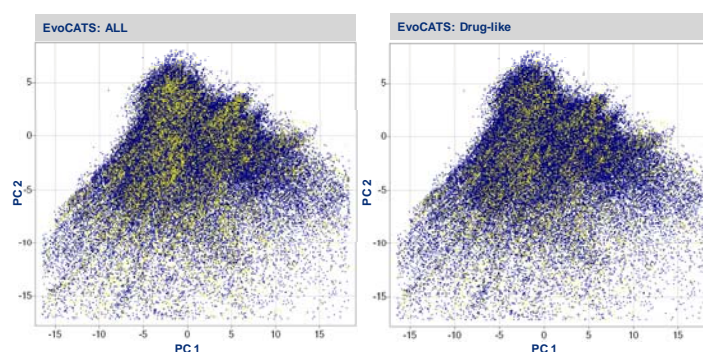


Figure 4. Scatter plot representation of the first 2 principle components obtained for the EvoCATS descriptors. Yellow: MDDR library Blue: Evotec Lead Discovery Library

- Evotec Lead Discovery Library provides high coverage of drug-like bioactive physicochemical space
- Evotec Lead Discovery Library provides high coverage of bioactive pharmacophore space

## SOM comparison using physicochemical descriptors

A self-organising map was calculated in Spotfire with default conditions using 76 physicochemical descriptors (*vide supra*).

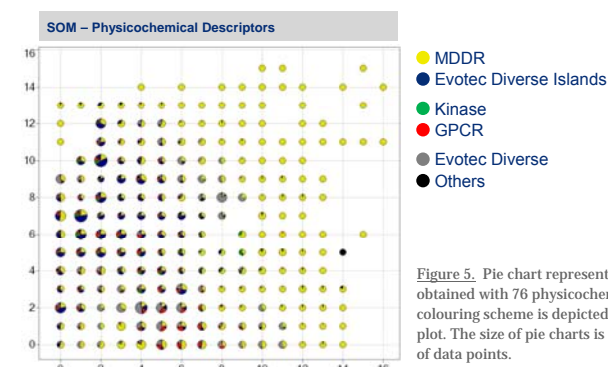


Figure 5. Pie chart representation of the SOM obtained with 76 physicochemical descriptors. The colouring scheme is depicted on the right side of the plot. The size of pie charts is defined by the number of data points.

- Evotec Lead Discovery Library and MDDR provide overlapping and distinct regions in physicochemical property space

## Clustering-based comparison using physicochemical descriptors and fingerprints

The hierarchical k-means clustering program was downloaded from <http://gecco.org.chemie.uni-frankfurt.de/hkmeans/index.html>. Clustering was performed using 76 physicochemical descriptors (*vide supra*) and chemical fingerprints from ChemAxon (default conditions). Stop thresholds were determined as described in reference 2. As "class" file, a flag for the Evotec Lead Discovery library and the MDDR library was incorporated. Shannon entropy values and terminal clusters were extracted, processed and further analysed in Excel and Spotfire, respectively.

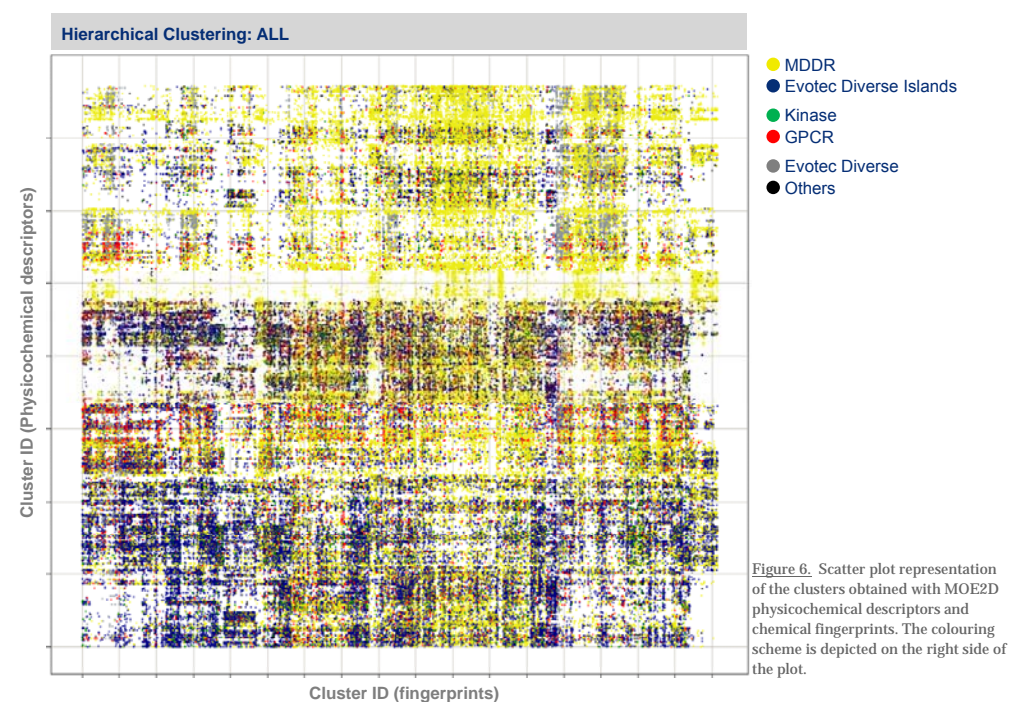


Figure 6. Scatter plot representation of the clusters obtained with MOE2D physicochemical descriptors and chemical fingerprints. The colouring scheme is depicted on the right side of the plot.

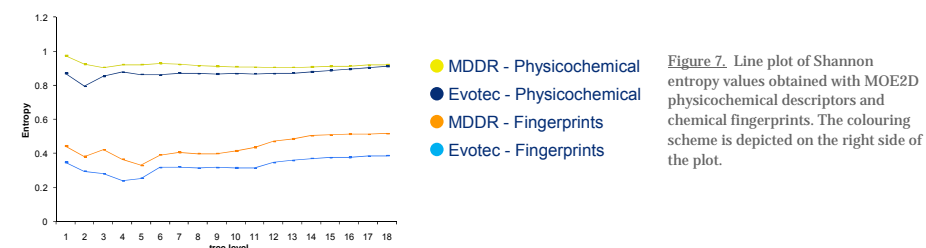


Figure 7. Line plot of Shannon entropy values obtained with MOE2D physicochemical descriptors and chemical fingerprints. The colouring scheme is depicted on the right side of the plot.

- Chemical fingerprints better suited to separate the Evotec from the MDDR library
- Evotec lead discovery library and MDDR provide overlapping and distinct regions in physicochemical property space

## Conclusion

Our conclusion is that coarse-grained analyses such as PCA are suitable to provide a rough overview of the chemical space and to triage for compounds covering a desired physicochemical and/or pharmacophore property space. Fine grained cluster analyses should then be applied for the selection of novel chemical subspaces with respect to a given reference library. At Evotec workflows have been implemented to automatically guide such a selection which will provide added value for future client drug discovery programs.

## References

- (1) MDL Drug Data Report, Version December 2009; Symyx Technologies Inc.: Santa Clara, CA, 2009.
- (2) A. Boecker *et al.*, J. Chem. Inf. Mod. (2005) 45, 807-815
- (3) Molecular Operating Environment (MOE), Version 2007.09; Chemical Computing Group Inc.: Montreal, Canada, 2007.
- (4) Spotfire DecisionSite, Version 9.1; TIBCO Software Inc.: Palo Alto, CA, 2008.
- (5) JChem, Version 3.2.8; ChemAxon Ltd.: Budapest, Hungary, 2006.
- (6) Whitley *et al.*, J. Chem. Inf. Comput. Sci. (2000) 40, 1160-1168