

# Compound set optimization and sequential screening using Emerging Chemical Patterns

J. Auer<sup>1</sup>

J. Bajorath<sup>1</sup>

<sup>1</sup> *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, University of Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany.*

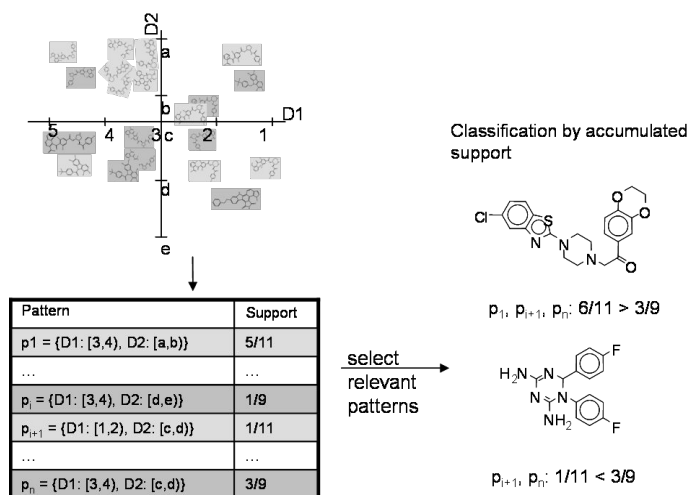
A method called “Emerging Chemical Patterns” (ECP) has recently been introduced as a novel approach to binary molecular classification<sup>1</sup>. The underlying pattern recognition algorithm was first introduced in computer science and then adopted for applications in medicinal chemistry and compound screening. The methodology makes it possible to extract key molecular features from very few known active compounds and classify molecules according to different potency levels. The approach was developed in light of the situation often faced during the early stages of lead optimization efforts: too few active reference molecules are available to build computational models for the prediction of potent compounds. The ECP method generates high-resolution signatures of active compounds by selecting class-specific combinations of 2D descriptor value ranges. These signatures can then be used to build highly accurate classifiers (Figure 1).

A special feature of ECP is its ability to accurately classify molecules on the basis of very small training sets containing only a few compounds. This feature is highly relevant for virtual compound screening when only very few experimental hits are available as templates. We designed an experiment based on four classes from literature sources (benzodiazepines, dihydrofolate reductase inhibitors, glycogen synthase kinase-3 inhibitors and HIV protease inhibitors), comparing ECP to a decision tree approach and a binary QSAR implementation. The analysis showed that ECP produced predictive models on the basis of training sets consisting of only three compounds.<sup>1</sup>

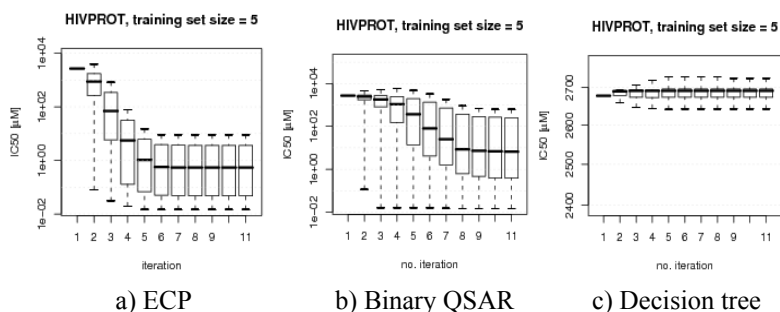
In addition to individual compound predictions, an iterative ECP scheme has been designed which optimizes a compound set's potency in a sequential manner. In each iteration, small compound sets are selected as training sets and used to remove weakly potent compounds. We could show that this iterative ECP classification produced compound selection sets with increases in average potency of up to 3 orders of magnitude (Figure 2).

The ability of ECP to produce highly accurate classifiers based on small training sets can also be used to reduce the experimental effort in high-throughput screening campaigns by combining experimental screening and ECP classification in a sequential screening methodology<sup>2</sup>. We simulated sequential screening using an experimental high-throughput screening (HTS) data set containing inhibitors of dihydrofolate reductase. We focused on minimizing the number of database compounds that need to be evaluated in order to identify a substantial fraction of available hits. Iterative ECP calculations recovered on average between 19% and 39% of available hits in the data set while dramatically reducing the number of compounds that need to be tested to 0.002% - 9% of the screening database.<sup>2</sup>

1. Auer, J. and Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection *J. Chem. Inf. Model.* **2006**, 46, 6, 2502 – 2514.
2. Auer, J. and Bajorath, J. Simulation of Sequential Screening Experiments Using Emerging Chemical Patterns, *Medicinal Chemistry* **2008**, 4, 1, 80 – 90.



**Figure 1: Classification of compounds based in ECPs.** Class-specific descriptor combinations (patterns) are computed from a training set. For classification, the supports (fraction of training compounds that match a pattern) of all patterns matching the test compounds are accumulated and the class with the highest accumulated support is the predicted class.



**Figure 2: Simulated lead optimization.** ECP, binary QSAR, and a decision tree are used to select highly potent sets of active compounds in an iterative procedure. During each step, compounds are classified as highly active ( $\leq 1\mu\text{M}$ ) or weakly active ( $> 1\mu\text{M}$ ) and weakly active compounds are iteratively removed.