# IMPROVED COMPOUND LIBRARY ENHANCEMENT USING ARTIFICIAL INTELLIGENCE ALGORITHMS FROM COMPUTER CHESS

**Roger Sayle[1], John Mayfield[1], Noel O'Boyle[1] and Nicolas Zorn[2]**

[1] NextMove Software Ltd, Cambridge, UK.

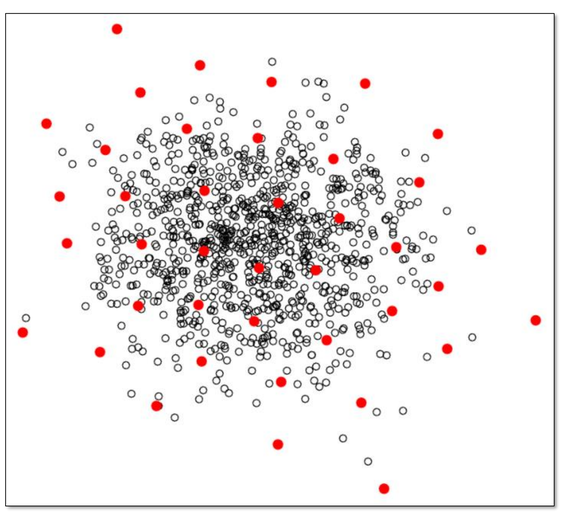
[2] Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland.

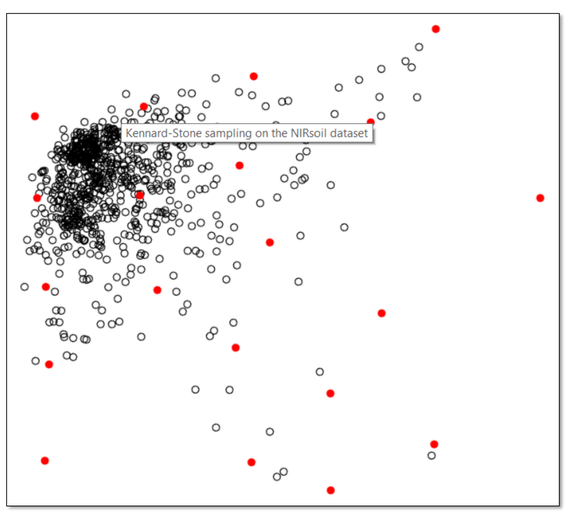
## 1. Abstract

As the number of small molecules available for purchase increases dramatically, the task of maintaining a screening collection of diverse/representative compounds grows ever more challenging. Currently (in 2019), there are around a billion small organic compounds available from chemical vendors, requiring efficient algorithms for performing diversity selection and compound profiling. In real-world applications, diversity selection is further complicated by the existence of a current/previous screening collection of several million compounds (which get depleted over time and may no longer be available or optimally desirable) and the desire to sample different regions of chemical space (such as kinase inhibitors or peptides) with different densities.

## 2. Diversity Selection

The MaxMin algorithm[1,2] is frequently used in cheminformatics to pick a diverse set of compounds. By repeatedly selecting compounds that are furthest from the closest reference compound, this approach uniformly samples chemical space. Picking differs from clustering in that it ignores biases due to varying density.

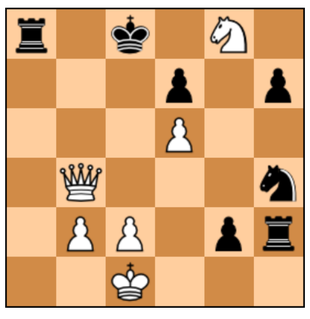





As samples selected by MaxMin are guaranteed to be a minimum distance from each other, this method is related to Taylor-Butina (Sphere exclusion) clustering.

## 3. Computer Chess

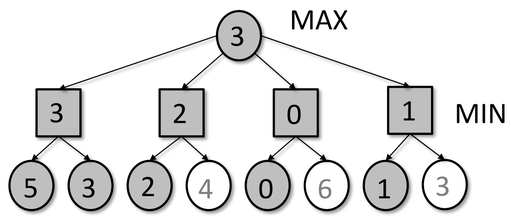
A significant improvement to traditional MaxMin diversity selection is to apply an AI technique from Chess and Go programs to prune the number of positions considered in Minimax game trees, called alpha-beta pruning.



Los Alamos Chess (6x6 board)
White has 17 possible moves.
The 11 that don't check, lose.
Five checks, lose the queen.

Not all of black's moves need be considered to identify white's best option.

Likewise, in the game tree on the right only the grey nodes need to be visited to determine the Maximin of the tree root.




Conceptually, the insight is that given a list of list of numbers, calculating the min of the min, or the max of the max requires inspecting every value, but calculating the min of the max or the max of the min can use upper/lower bounds to avoid inspecting every element, sometimes reducing the work significantly.

## 4. Performance Improvement

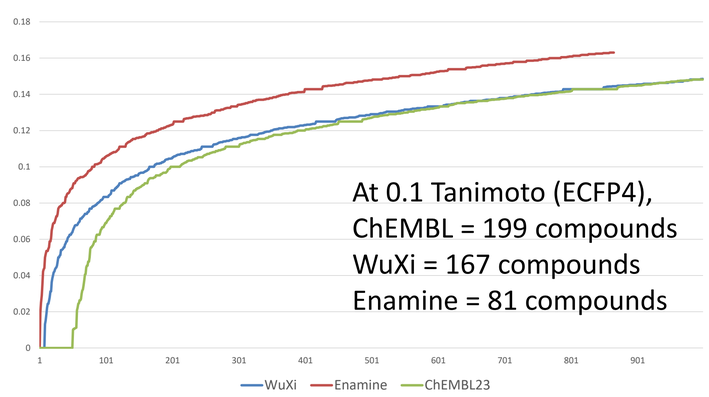
Thanks to alpha-beta pruning, diversity selection no longer requires the full $N^2$ distance matrix to be calculated. In practice, the performance savings are dramatic, for example selecting the single compound from the 171M compounds in Enamine REAL 2017 to add to ChEMBL v23 (1.7M compounds) requires only 181 billion binary fingerprint comparisons, or 1/82750 the effort previously required, and selecting the 50 most diverse compounds to add requires only 469 billion FP comparisons.
For comparison, a traditional distance matrix would require 60 petabytes of storage and 15 quadrillion (1.4E16) FP comparisons, which would be prohibitive on conventional hardware.

The improved MaxMin Diversity Picker was contributed to RDKit in September 2017, and has been part of the standard RDKit distribution since release 2017_09.

## 5. MaxMin Progress as a Measure of Data Set Diversity

As the MaxMin diversity selection algorithm progresses, additional samples are selected closer and closer to previous samples. The evolution of this threshold reveals interesting properties of the distribution of points in the underlying data. Consider three data sets: ChEMBL v23 with 1.7M compounds, a WuXi screening collection with 215M compounds and Enamine REAL2018 with 680M compounds.

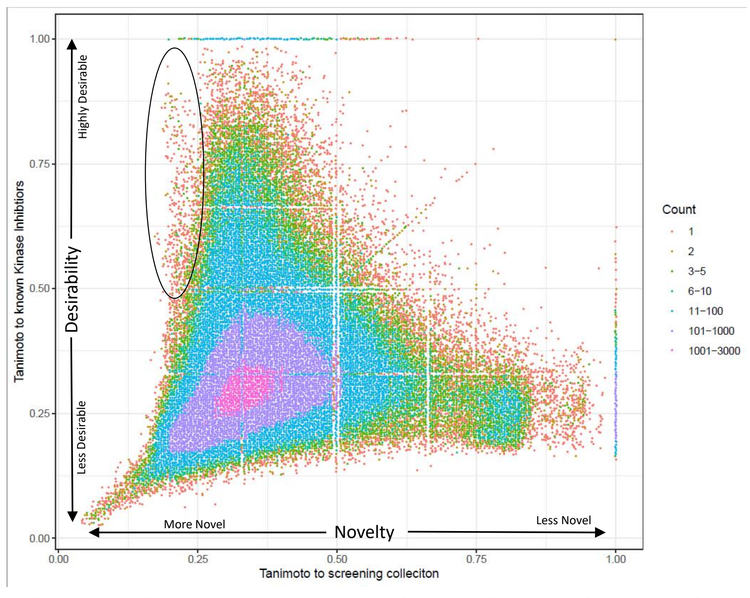
| Count | WuXi | Enamine | ChEMBL23 |
|---|---|---|---|
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0.019608 | 0 |
| 4 | 0 | 0.025 | 0 |
| 5 | 0 | 0.028986 | 0 |
| 6 | 0 | 0.035294 | 0 |
| 7 | 0 | 0.042254 | 0 |
| 8 | 0 | 0.043478 | 0 |
| 9 | 0 | 0.044118 | 0 |
| 10 | 0.013158 | 0.05 | 0 |
| 11 | 0.015152 | 0.051724 | 0 |
| 12 | 0.018868 | 0.053571 | 0 |
| 13 | 0.022222 | 0.053571 | 0 |
| 14 | 0.02439 | 0.053571 | 0 |
| 15 | 0.02439 | 0.054795 | 0 |
| 16 | 0.025974 | 0.055556 | 0 |
| 17 | 0.027778 | 0.058824 | 0 |
| 18 | 0.03125 | 0.058824 | 0 |
| 19 | 0.03125 | 0.060606 | 0 |
| 20 | 0.036364 | 0.064935 | 0 |
| 21 | 0.038462 | 0.066667 | 0 |
| 22 | 0.04 | 0.068493 | 0 |
| 23 | 0.041667 | 0.068966 | 0 |
| 24 | 0.042254 | 0.068966 | 0 |
| 25 | 0.043478 | 0.070175 | 0 |




The above plot shows that 25 compounds can be selected from ChEMBL, each with zero Tanimoto to the others, 9 compounds can be selected from WuXi with this property, but only 2 from Enamine. Likewise, to cover ChEMBL's chemical space at an ECFP4 Tanimoto of 0.1 requires 199 compounds, WuXi's space requires 167 compounds and Enamine's only 81 compounds. These numbers reveal that larger data sets don't automatically provide greater diversity.

## 6. Multi-objective Optimization

A further insight is that in practice diversity selection is actually a multi-objective optimization; without additional constraints diversity picking tends to initially select "wacky" molecules, choosing obscure chemical functionality and enriching for errors in molecular representation ("broken" molecules). Hence we formulate diversity selection as an operation over three compound sets; selecting from available compounds those that are (most) dissimilar to an existing in-house screening collection, but are maximally similar to a reference set of desirable compounds (recent FDA approvals or known kinase inhibitors). The goal is therefore not to select the most diverse compounds in the entirety of chemical space, but to uniformly sample from within a constrained "drug-like" space.




This approach can be visualized as a scatter plot with novelty (similarity to the current collection) on the X-axis, and desirability (similarity to prototype ideal compounds) on the Y-axis. That alpha-beta pruning can be applied to one axis and not the other, leads to an interesting asymmetry, but enables the efficient identification of compounds on (or near) the Pareto frontier (i.e. those to be considered for purchasing) in a fraction of the computational effort previously required.

## 7. Bibliography

1. M. Ashton, J. Barnard, P. Willett *et al.*, "**Identification of Diverse Database Subsets using Property-based and Fragment-based Molecular Descriptors**", *Quantitative Structure-Activity Relationships*, Vol. 21, pp. 598-604, 2003.
2. R. Kennard and L. Stone, "**Computer aided design of experiments**", *Technometrics*, Vol. 11, No. 1, pp. 137-148, 1969.
3. G. Landrum, "**Picking diverse compounds from large sets**" and "**Optimizing Diversity Picking in the RDKit**", RDKit blog, August 2014.

NextMove Software

www.nextmovesoftware.co.uk
www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
England, UK CB4 0EY