RECENT IMPROVEMENTS TO THE RDKIT

ROGER SAYLE

NEXTMOVE SOFTWARE, CAMBRIDGE, UK

6th RDKit UGM, Berlin, Germany, Thursday 21st September 2017

MOTIVATION: COMPOUND ACQUISITION

- Given a existing screening collection of *X* compounds, and with *Y* vendor compounds available for purchase, how should I select the next *Z* diverse compounds to buy.

- Typically, *X* is about 2M and *Y* is about 100M.

RDKIT'S MAXMINPICKER

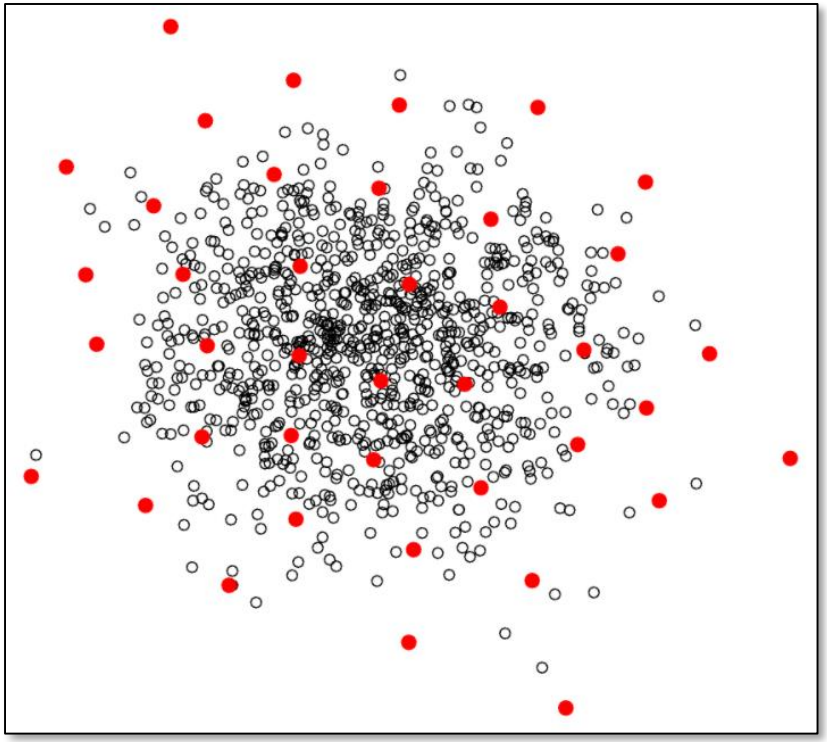
- Picking diverse compounds from large sets, 2014/08
- Optimizing Diversity Picking in the RDKit, 2014/08

- M. Ashton, J. Barnard, P. Willett et al., "Identification of Diverse Database Subsets using Property-based and Fragment-based Molecular Descriptors", Quant. Struct.-Act. Relat., Vol. 21, pp. 598-604, 2002.
- R. Kennard and L. Stone, "Computer aided design of experiments", Technometrics, Vol. 11, No. 1, pp. 137-148, 1969.

CONCEPTUAL ALGORITHM

- If no compounds have been picked so far, choose the first picked compound at random.
- Repeatedly select the compound furthest from it's nearest picked compound [hence the name maximum-minimum distance].
- Continue until the desired number of picked compounds has been selected (or the pool of available compounds has been exhausted).

SELECTION VISUALIZATION



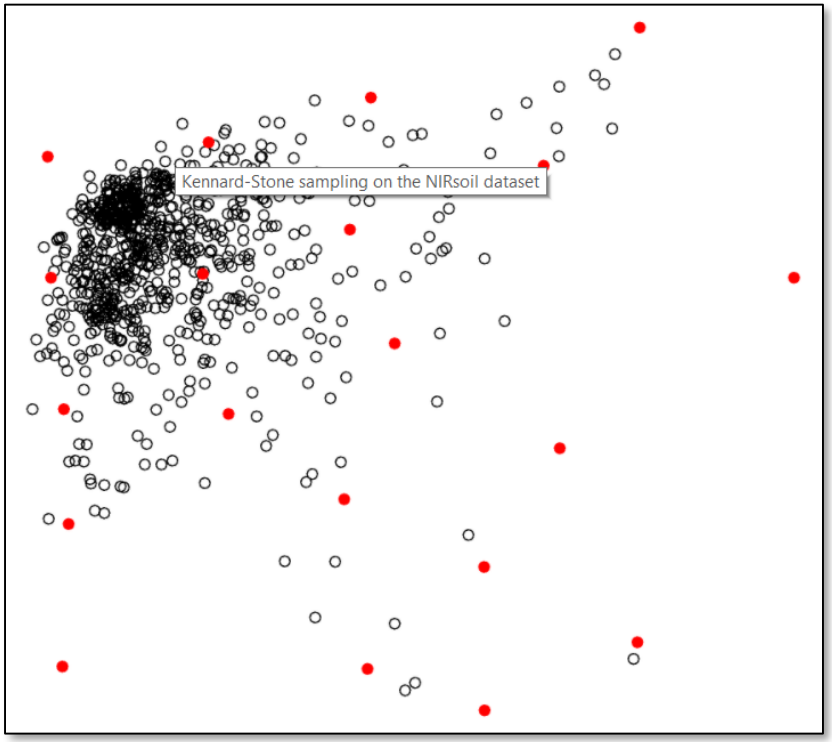


Image Credits: Antoine Stevens, the ProspectR package on github

PROS VS. CONS

- Optimal picking is (NP-)Hard.
- Density vs. Diversity
  - With biased data sets, random sampling follows density, where MaxMin optimizes coverage.
  - Picking != Clustering.
- Worst-case NN-1 bounds
  - It's possible to (estimate a) bound on the worst case distance to nearest neighbor.
- Fraction of Data Set to be Sampled
- Scaling Performance of Algorithm

DISTANCE MATRICES ARE A BAD IDEA

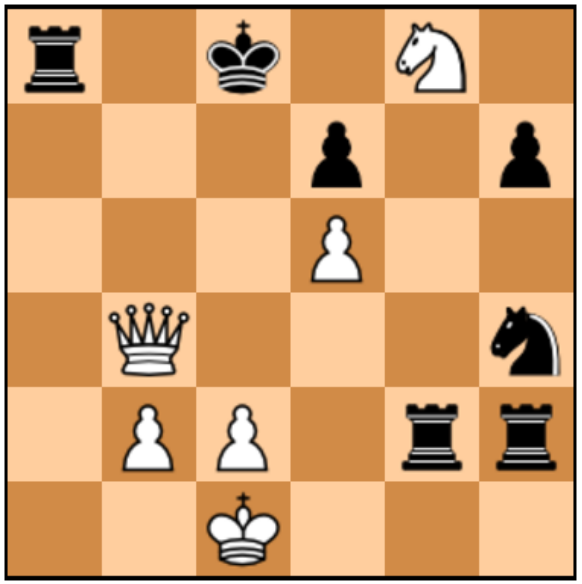
- Naïvely, one approach is to provide a picking algorithm with a full distance matrix.
- If X compounds are picked already, and Y is the initial pool to pick, from this requires (X+Y-1)*(X+Y)/2 time and space.

- One goal is to keep memory requirements O(X+Y).
- An intelligent algorithm should be able to avoid calculating any given distance more than once.

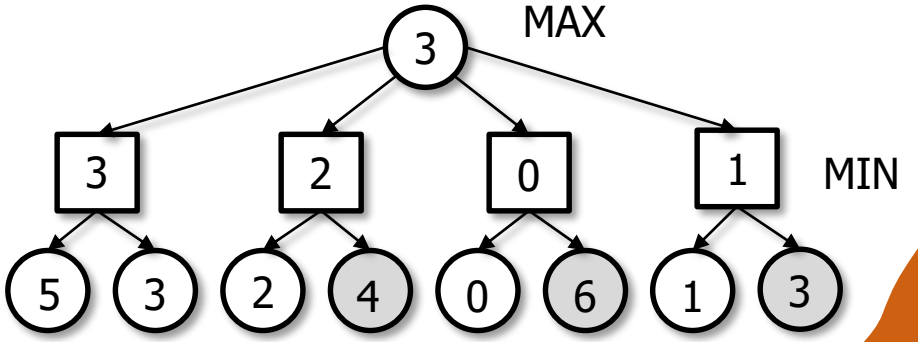
TAYLOR-BUTINA VS. MAXMIN PICKING

- At a comparable distance threshold, MaxMin picks are also a Leader (Tabu) clustering.
- As MaxMin picking doesn't require a distance matrix (NN list) it is significantly cheaper than Taylor-Butina.
- The distance bound for a given coverage is discovered rather than specified (by trial and error).
  - Taylor, JCICS, 1995, 35, pp. 59-67.
  - Butina, JCICS, 1999, 39, pp. 747-750.

ARTIFICIAL INTELLIGENCE (MINI-MAX)



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



Alpha cut-offs allow us to prune the search tree.

MAX-MIN PICKING

		Candidate Pool									
		0	1	2	3	4	5	6	7	8	9
Picks	0	3	1	4	1	5	9	2	6	5	3
	1	5	8	9	7	9	3	2	3	8	4
	2	6	2	6	4	3	3	8	3	2	7
	3	9	5	0	2	8	8	4	1	9	7
Minimums		3	1	0	1	3	3	2	1	2	3
Maximum		3									

MAX-MIN PICKING

		Candidate Pool									
		0	1	2	3	4	5	6	7	8	9
Picks	0	**3**	**1**	**4**	**1**	**5**	**9**	**2**	**6**	**5**	**3**
	1	**5**	8	**9**	7	**9**	**3**	2	**3**	**8**	4
	2	**6**	2	**6**	4	**3**	3	8	3	**2**	7
	3	**9**	5	**0**	2	8	8	4	1	9	7
< Bounds		3	1	0	1	3	3	2	3	2	3
Maximum		3									

RDKIT IMPLEMENTATION

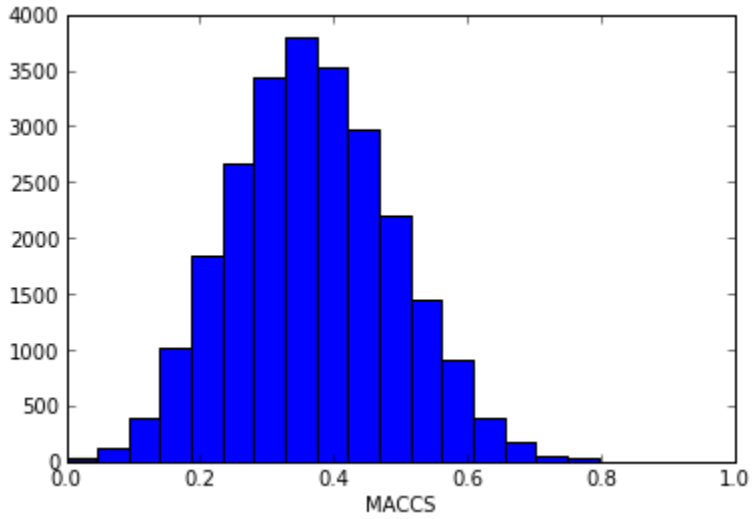
- For each candidate we track its bound, the number of picks used to calculate this bound, and a pointer to the next pool candidate (a singly linked list).
- Memory usage is O(poolsize), less than INT_LIST.
- No need for a distance matrix or distmat cache.
- Linked list preserves tie splitting/skips picked items.
- The same data structure is (ab)used as a visit array in an initial pass to remove firstPicks from the pool.

PERFORMANCE IMPROVEMENT

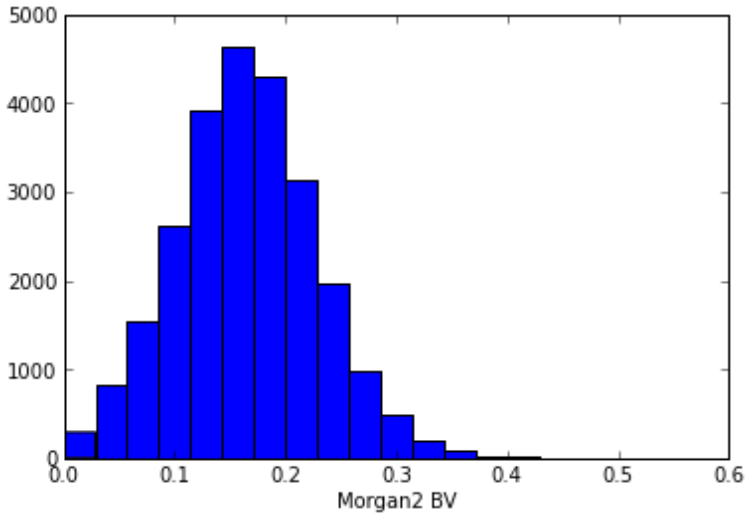
- Using Andrew Dalke's data set of 12386 benzodiazepines, using the first one thousand as the existing screening library, select the next 18 most diverse molecules from the remaining set (of 11386).
  - Original RDKit Implementation:
    - 224,688,273 FP comparisons 96.30s
  - Pruning using alpha cut-off:
    - 16,069,573 FP comparisons 6.79s (14x)
  - Preserving bounds across picks:
    - 1,047,982 FP comparisons 0.46s (209x)
  - Timings on Dell laptop, using 2048 bit Morgan radius 2 FPs.

SIGNIFICANT SIMILARITY

- Different similarity measures require different significance thresholds to distinguish random hits.



90%:0.528 95%:0.574 99%:0.656



90%:0.245 95%:0.271 99%:0.323

Thresholds for “random” in fingerprints the RDKit supports. 2013/10

In this talk I’ll ignore the influence of query and database composition on score significance.

RESULTS: DATA SET SAMPLING

- ChEMBL 23 (1727053 compounds)
  - 90%: 15107 compounds [5.45B FP cmps]
  - 95%: 24430 compounds [9.82B FP cmps]
  - 99%: 51463 compounds [23.9B FP cmps]
- eMolecules170601 (14328534 compounds)
  - 90%: 19049 compounds [45.7B FP cmps]
  - 95%: 32780 compounds [86.5B FP cmps]
  - 99%: 80135 compounds [247B FP cmps]

ASTRAZENECA VS. BAYER

AstraZeneca

Similarity of Bayer and AstraZeneca collection Origin of Libraries

Bayer Collection

2.7M structures (3 M)

- SCHERING part of the collection (875.000 structures) was cleaned and expanded library with mainly purchasable external compounds (2003-2005)
- Huge investments at BAYER between 2000 and 2007 to expand library based on proprietary building blocks
- Compound design based on favorable PhysChem properties and undesirable groups filtering
- Realization through *external* collaborations and internal combinatorial chemistry
- 1/3 classical medchem structures from optimization projects
- 2/3 combichem compounds

AZ Collection

1.4M structures (1.9M)

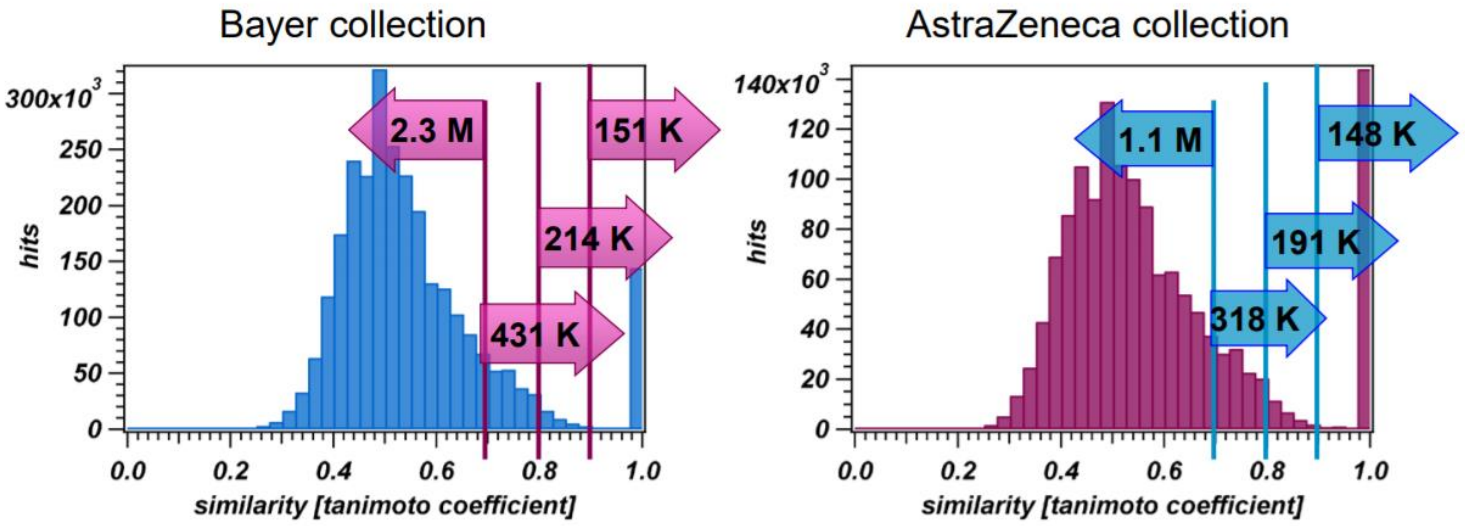
- AstraZeneca screeing collection underwent major review (structural and sample quality) in 2001/2002
- Strict classification on Phys-Chem and structural features
- Major acquisition campaign in 2002
- Three consecutive *Compound Collection Enhancement* programmes (2003-2005, 2006-2008, and 2008-2011)
- Internal design from Lead Generation chemistry, outsourced production of small libraries, no combichem
- >80% proprietary compounds

- Different portfolio history might lead to different biological activity space
- Proprietary compound based collections
- "Real" medicinal chemistry project compounds (drug-like/lead-like compounds)

ASTRAZENECA VS. BAYER

AstraZeneca

Similarity of Bayer and AstraZeneca collection
Distribution of Nearest Neighbour



"The AZ business case":
2.3 M Compounds are "new" to Astra

"The Bayer business case":
1.1 M Compounds are "new" to Bayer

SCREENING LIBRARY ENHANCEMENT

- Selecting 1K compounds for purchase from eMolecules (14M) to enhance ChEMBL (1.7M).
  - Reading eMolecules: 4780s
  - Reading ChEMBL: 821s
  - Generating FPs: 1456s
  - MaxMinPicker: 42773s [80B FP cmps]
- Fazit: Large scale diversity selection can be run overnight on a single CPU core.
- Selecting the first 18 compounds takes only 399s [715M FP cmps].

ADDITIONAL CONSIDERATIONS #1

- Rule 1: It's better to filter compounds for desirability, physical properties, Lipinski, price before picking.
  - Filtering is O(N), Picking isn't.
- With library screening, it's often preferable to have hits with neighbors rather than singletons, to confirm true positives, or provide initial QSAR; either
  - Picking should be performed on the pool of compounds that have neighbors
  - Each pick confirmed to have a neighbour as it is chosen.

ADDITIONAL CONSIDERATIONS #2

- If required, it is possible to sample some areas of chemical space (kinase inhibitors or fragments) with higher density than others.
- For those that want to do fingerprint comparison faster than RDKit, Andrew Dalke's ChemFP is worth a look.
- One possible refinement is to de-duplicate identical fingerprints (efficiently done using a hash table).
- Finally, a better similarity measure should produce better results (SmallWorld sales pitch goes here)

IMPLEMENTING FULL KENNARD-STONE

- The original Kennard-Stone algorithm requires the first two picks to be the furthest apart in a data set.
- Traditionally, this has required $O(N^2)$ time.
- A break-through in theoretical computer science since 2013 has changed this assumption:
  - Pilu Crescenzi, Roberto Grossi, Michel Habib, Leonardo Lanzi, Andrea Marino, "On computing the diameter of real-world undirected graphs", Theoretical Computer Science, Vol. 514, pp. 84-95, 2013.
  - Michele Borassi, Pierluigi Crescenzi, Michel Habib, Walter Kosters, Andrea Marino and Frank Takes, "On the Solvability of the Six Degrees of Kevin Bacon Game: A Faster Graph Diameter and Radius Computation Method", 2014.
  - Michele Borassi, Pierluigi Crescenzi, Michel Habib, Walter A. Kosters, Andrea Marino and Frank W. Takes, "Fast diameter and radius BFS-based computation in (weakly connected) real-world graphs with an application to the six degrees of separation games", Theoretical Computer Science, Vol. 586, pp. 59-80, 2015.

BOUNDED VERTEX ECCENTRICITY

- The eccentricity of a vertex, v, is the greatest distance to any other vertex.
- The radius of a graph is the minimum eccentricity.
- The diameter of a graph is a maximum eccentricity.

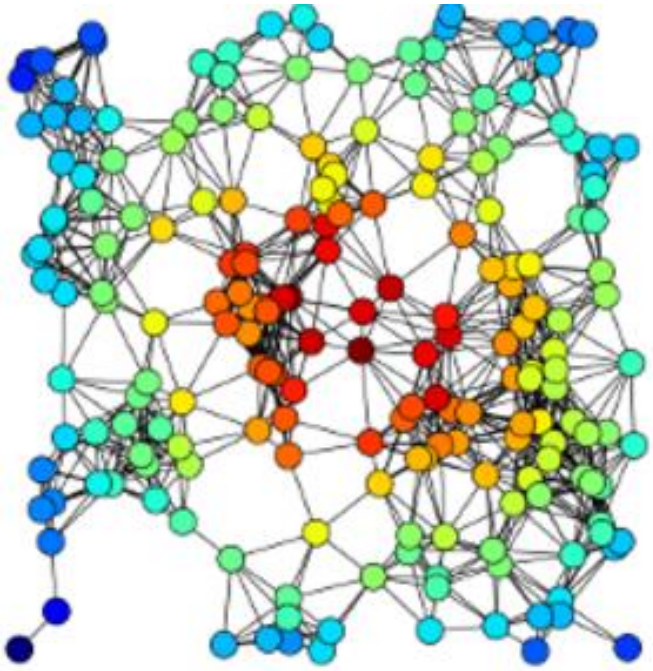


Image Credit: Tapiocozzo, Wikipedia page on “Centrality”.

SUMSWEEP RESULTS

- Finding furthest two atoms of 327 heavy atoms in the protein crambin (1CRN)
  - Brute Force: 53301 comparisons
  - SUMSWEEP: 6636 comparisons [k=21]
- Finding furthest apart (using Hamming distance on 2K MFP2 FPs) of 250251 compounds in NCI August 2000 data set.
  - Brute Force: 31,312,656,375 comparisons
  - SUMSWEEP: 43,278,372 comparisons [k=173]

CONCLUSIONS

- A little computer science can make compound purchasing decisions a whole lot faster.

ACKNOWLEDGEMENTS

- The RDKit crew
  - Greg Landrum and Brian Kelley
- The NextMove Software crew
  - John Mayfield and Noel O'Boyle
- Industrial Inspiration
  - Darren Green, GSK
  - Roman Affentranger, Roche
  - Pat Walters, Relay Therapeutics
  - Andrew Dalke, Dalke Scientific.