

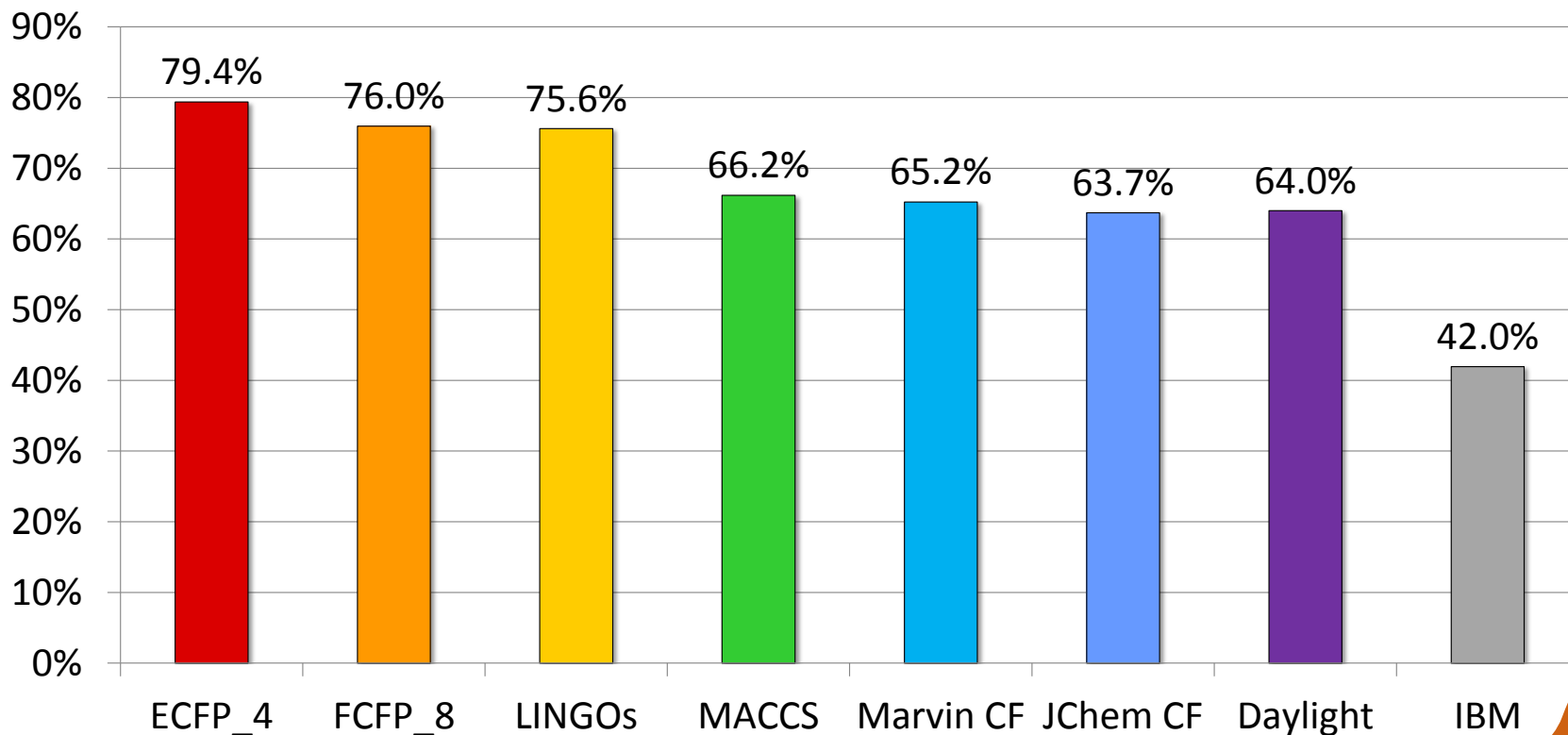


# SmallWorld: Efficient Maximum Common Subgraph Searching of Large Chemical Databases

Roger Sayle, Jose Batista and Andrew Grant  
NextMove Software, Cambridge, UK  
AstraZeneca R&D, Alderley Park, UK



# 2D CHEMICAL SIMILARITY



Briem & Lessel bioactivity benchmark [2000]

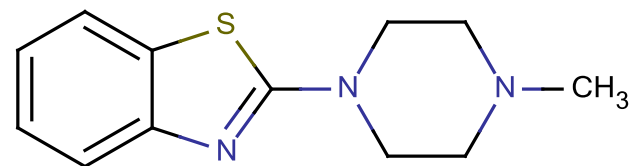
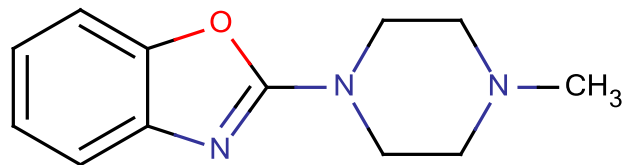


# THE MYTH OF FINGERPRINTS

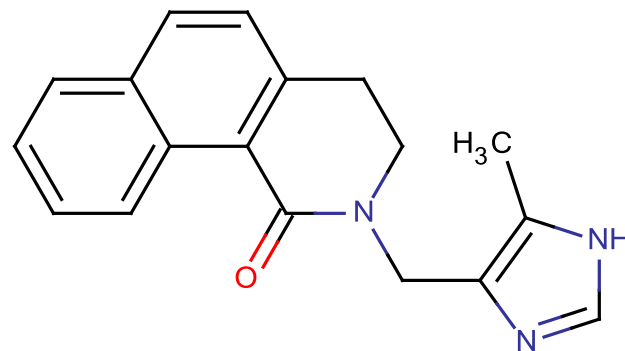
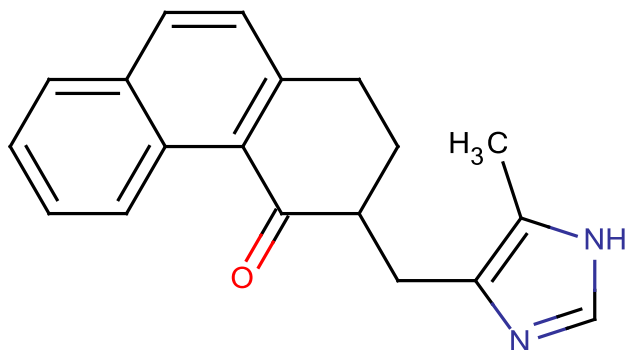
- The most popular chemical similarity approaches reduce the graph representation of a molecule to a vector of local features, and compare these.
- “All the right notes, just not in the right order”.
- Binary fingerprints (MACCS and Daylight) suffer from feature saturation where all long alkanes, proteins and nucleic acids have the same fingerprint.



# COUNTER INTUITION



Daylight Tanimoto: 0.44



Daylight Tanimoto: 0.39



# STRING EDIT DISTANCE

- In 1965, Vladimir Levenshtein introduced string edit distance as similarity measure between strings.
- The minimum number of insertions, deletions and substitutions to transform one string to another.
- Thanks to efficient dynamic programming algorithms of Needleman-Wunsch and Smith-Waterman, sequence alignment is at the core of bioinformatics.
- Prior to string edit distance, similarity between sequences was performed by heuristic methods locally comparing short words, or  $k$ -tuples.



# GRAPH EDIT DISTANCE

- Graph Edit Distance (GED) is the extension of this concept to graphs, as the minimum number of edit operations required to transform one graph into another.
  - Alberto Sanfeliu and K.S. Fu, “A Distance Measure between Attributed Relational Graphs for Pattern Recognition”, IEEE Transactions of Systems, Man and Cybernetics (SMC), Vol. 13, No. 3, pp. 353-362, 1983.
- Edit operations consist of insertions, deletions and substitutions of nodes and edges (atoms and bonds).



# GED AND MCES

- Unfortunately, calculating GED has been shown to be a generalization of calculating Maximum Common Subgraph (MCS), a favorite of chemists but known to be computationally very expensive (NP-Hard).
  - Horst Bunke, “On a Relation between Graph Edit Distance and Maximum Common Subgraph”, Pattern Recognition Letters, Vol. 18, No. 8, pp. 689-694, 1997.
  - Horst Bunke and Kim Shearer, “A Graph Distance Metric Based on the Maximal Common Subgraph”, Pattern Recognition Letters, Vol. 19, pp. 255-259, 1998.



# ADVANCES IN COMPUTER SCIENCE

- Fortunately, recent developments in theoretical computer science and advances in processing power and disk storage finesse the problem.
  - B.T. Messmer and H. Bunke, “Subgraph Isomorphism in Polynomial Time”, Tech. Report, Univ. Berlin, 1995.
  - B.T. Messmer and H. Bunke, “Subgraph Isomorphism Detection in Polynomial Time on Preprocessed Graphs”, In Proc. Asian Conf. on Computer Vision, pp. 151-155, 1995.
  - Michel Neuhaus and Horst Bunke, “Bridging the Gap Between Graph Edit Distance and Kernel Machines”, World Scientific Press, 2007.

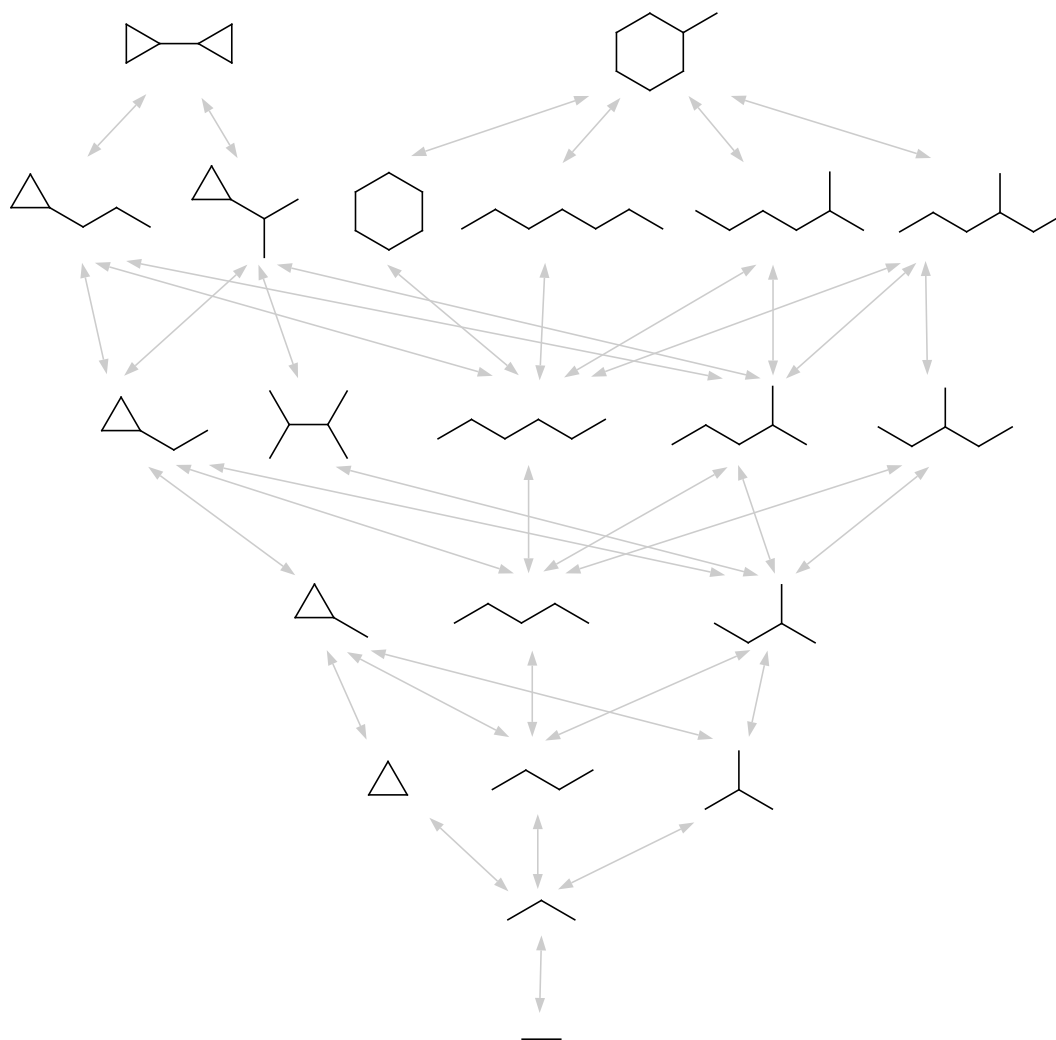


# THE SECRET... PRE-PROCESSING

- Imagine in advance enumerating, canonicalizing and storing all subgraphs for a given molecule, sorted by the number of bonds.
- The task of finding the largest common subgraph then becomes an almost trivial ordered intersection.
- Modern state-of-the-art cheminformatics systems can easily handle databases several orders of magnitude larger than the current largest non-virtual databases (i.e. many billions of subgraphs).



# SMALLWORLD CHEMICAL SPACE



# A GRAND UNIFIED THEORY

- The three fundamental of forces of cheminformatics:
  - Molecular Identity
  - Substructure Search
  - Chemical Similarity
- Bemis and Murcko Scaffolds
- Shuffenhauer Scaffold Trees
- Feature, Path and Radial Fingerprints
- Matched Molecular Pairs



# GRAPH VS. SUBGRAPH ENUMERATION

- All traditional MCS algorithms start from one bond and grow; SmallWorld starts from N bonds and shrinks.
- In typical databases, highly similar molecules which cause problems for MCS, are found quickly, typically terminating the search.
- The work of Jean-Louis Raymond *et al.* on GDB-15, using Brendan McKay's GENG to enumerate all possible graphs, significantly over estimates the number of (sub)graphs encountered in practice.



# EFFICIENT SUBGRAPH ENUMERATION

- A connected Maximum Common Edge Subgraph (MCES) with one less bond is formed by either (i) deleting a bond to a terminal atom, or (ii) deleting a ring (cyclic) bond.
- Partitioning cyclic from acyclic bonds can be done efficiently in  $O(N)$  time, and even this only needs to be recalculated after deleting a ring bond, as deleting terminal bonds doesn't affect ring membership.



# SUBGRAPH COUNTS OF MOLECULES

Name	Atoms	MW	Anon SGs	Elem SGs
Benzene	6	78	7	7
Cubane	8	104	64	64
Ferrocene	11	186	3,154	3,219
Aspirin	13	180	127	332
Dodecahedrane	20	260	440,473	440,473
Ranitidine	21	314	436	1,207
Clopidrogel	21	322	10,071	22,170
Morphine	21	285	176,541	496,467
Amlodipine	28	409	58,139	147,128
Lisinopril	29	405	24,619	34,496
Gefitinib	31	447	190,901	337,174
Atorvastatin	41	559	3,638,523	6,019,427

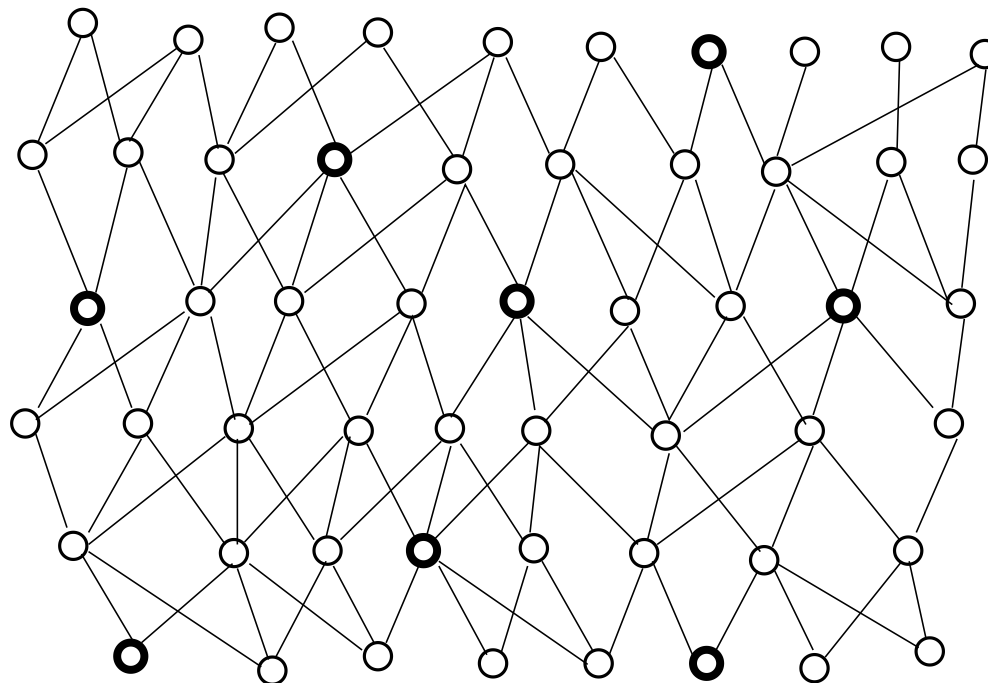


# MOLECULE DB SIZE DISTRIBUTION

$\leq$ Bond Count	% MDDR 2011.2	% NCBI PubChem
$\leq$ 20 bonds	6%	14%
$\leq$ 25 bonds	18%	30%
$\leq$ 30 bonds	36%	55%
$\leq$ 35 bonds	56%	77%
$\leq$ 40 bonds	73%	89%
$\leq$ 45 bonds	83%	93%
$\leq$ 50 bonds	89%	95%
$\leq$ 55 bonds	92%	97%
$\leq$ 60 bonds	93%	98%
$\leq$ 65 bonds	94%	98%
$\leq$ 70 bonds	95%	99%



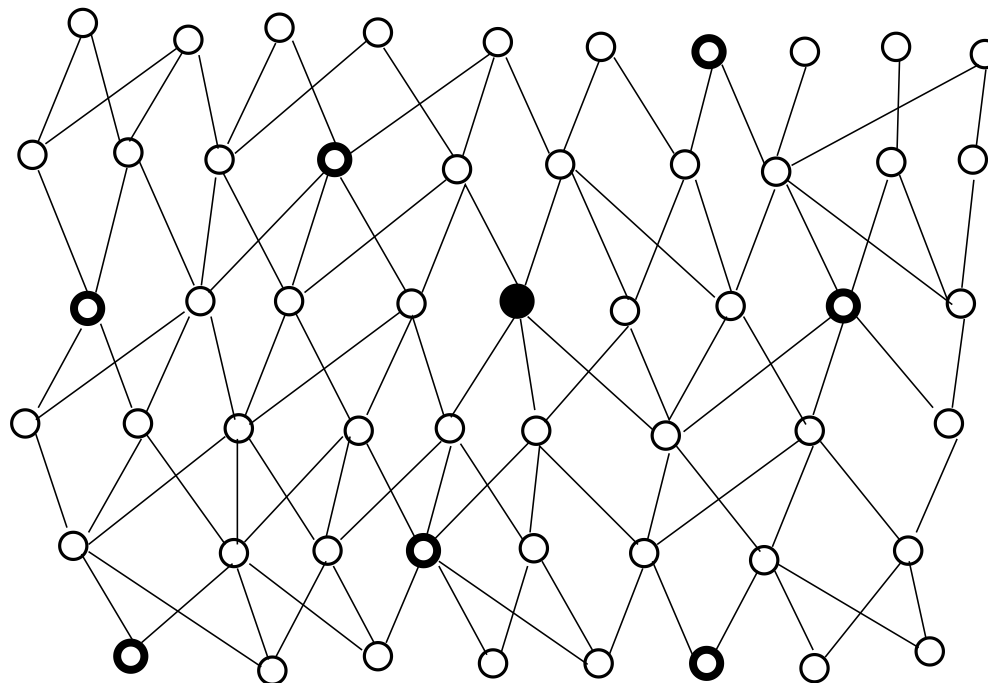
# SMALLWORLD SEARCH



SmallWorld lattice: Bold circles denote indexed molecules, thin circles represent virtual subgraphs.



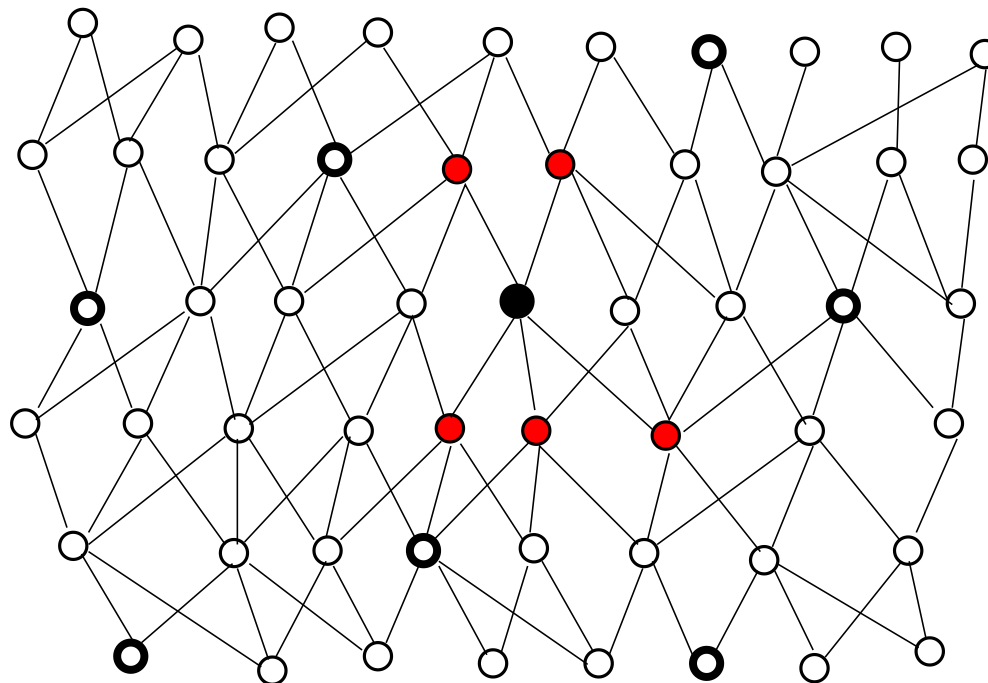
# SMALLWORLD SEARCH



The solid circle denotes a query structure which may be either an indexed molecule or a virtual subgraph.



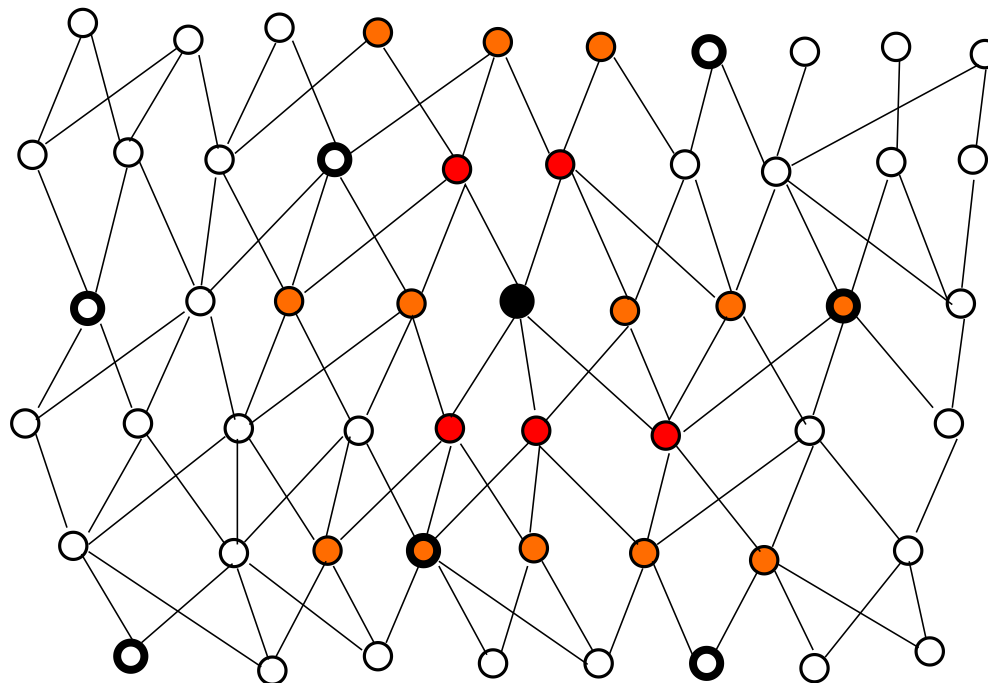
# SMALLWORLD SEARCH



The first iteration of the search adds the neighbors of the query to the “search wavefront”.



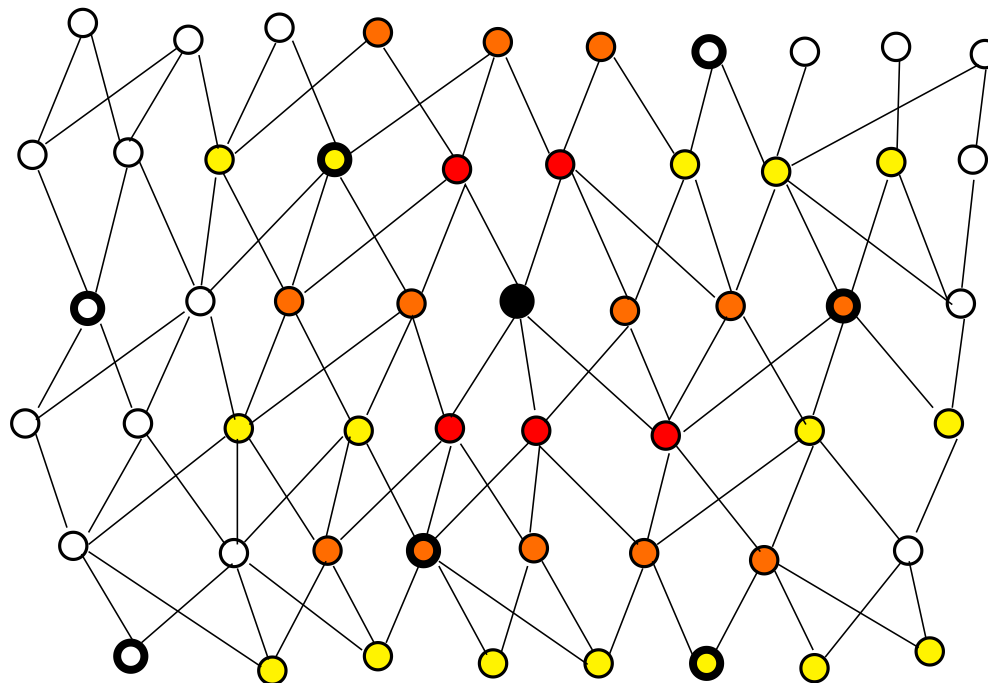
# SMALLWORLD SEARCH



Each subsequent iteration propagates the wavefront by considering the unvisited neighbors of the wavefront.



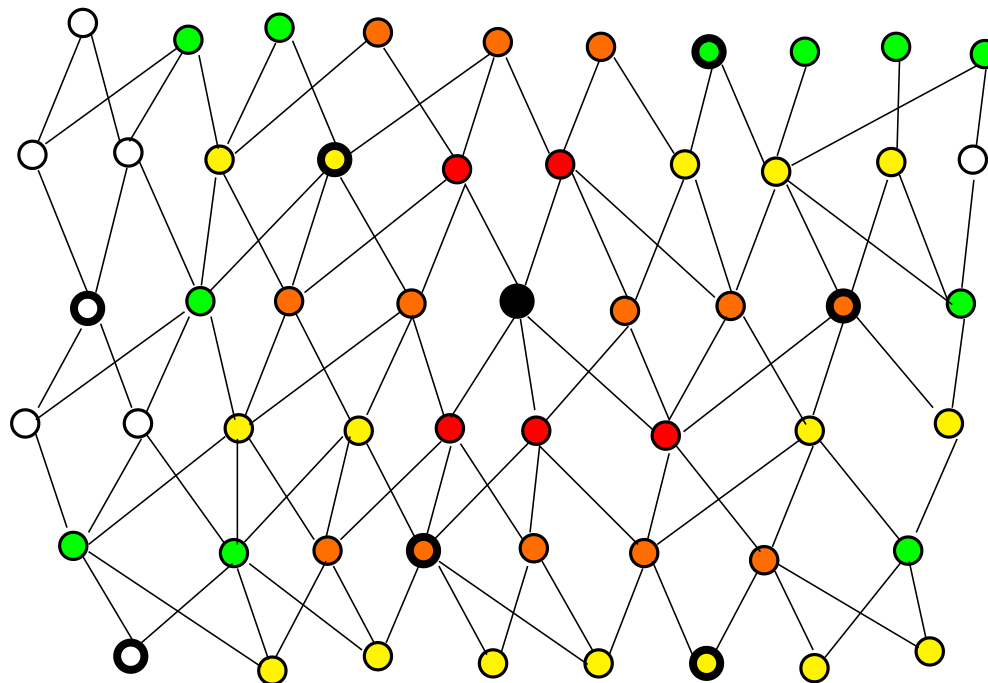
# SMALLWORLD SEARCH



At each iteration, “hits” are reported as the set of indexed molecules that are members of the wavefront.



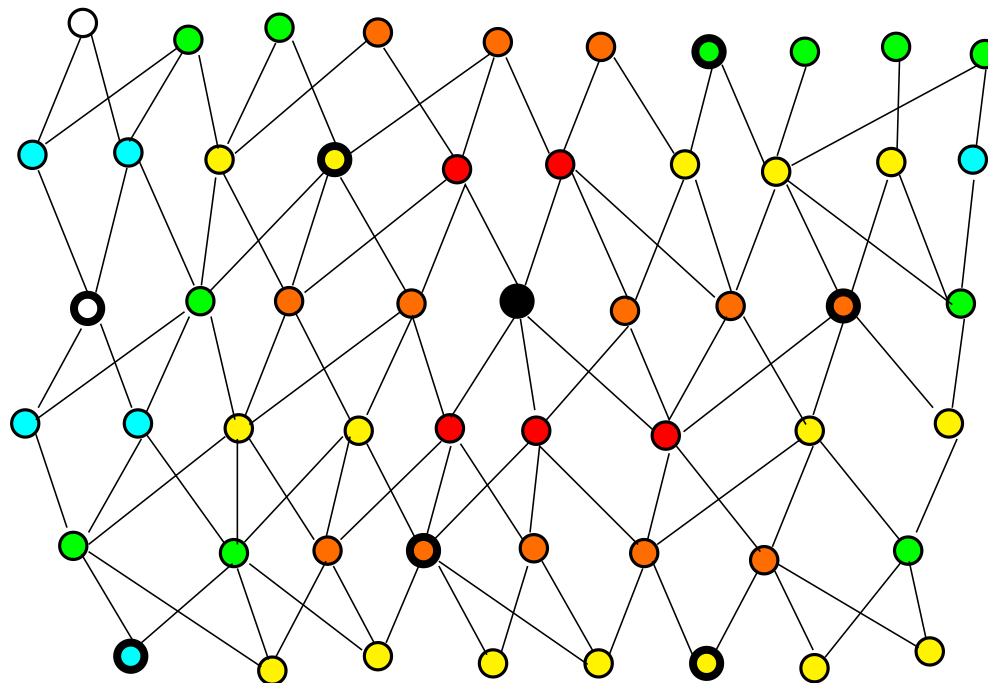
# SMALLWORLD SEARCH



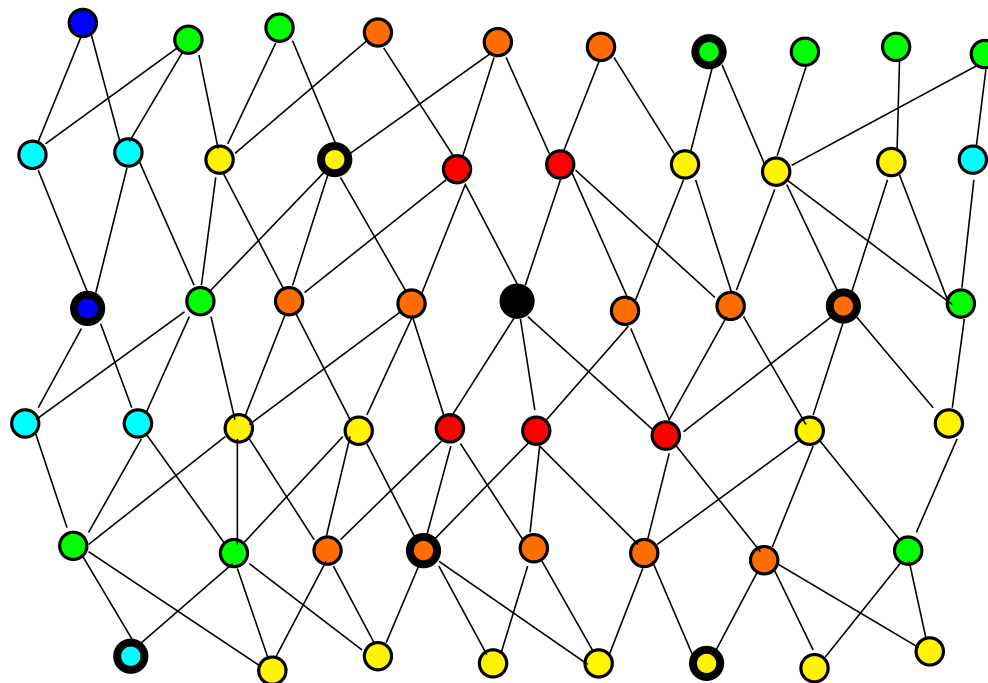
The search terminates once sufficient indexed neighbors have been found (or a suitable iteration limit is reached).



# SMALLWORLD SEARCH



# SMALLWORLD SEARCH



# ALGORITHM REFERENCES

- Scott Beamer *et al.* “Seaching for a Parent Instead of Fighting Over Children: A Fast Breadth-First Search Implementation for Graph500”, UCB Tech. Rep. 2011.
- Ricardo A. Baeza-Yates, “A Fast Set Intersection Algorithm for Sorted Sequences”, In. Proc. 15<sup>th</sup> Symposium on Combinatorial Pattern Matching, LNCS Vol. 3109, pp. 400-408, 2004.



# PROOF-OF-CONCEPT DB STATISTICS

- To evaluate design ideas a small prototype proof-of-concept “anonymous” database was implemented.
- 202,169,109 nodes ( $\sim 2^{28}$  nodes)
- 1,763,328,320 edges ( $\sim 2^{31}$  edges)
- Storage requirements:  $\sim 75$ Gbytes.
- Average degree (fan-out) of node: 17.44
- 62.5M acyclic nodes, 78.3M have single ring.
- 1,030,565,730 edges were terminal edges and 732,762,590 edges were ring deletion edges.



# EXAMPLE SEARCH STATISTICS

- A Briem & Lessel benchmark run, searching for the (upto) 10 nearest neighbors of each of 380 drug-sized query molecules takes 8m42s (less than 1.4 seconds per query) on single thread i7-2600/16GB.
- The median distance for each search was 8 edits, with the shortest finishing after 2 edits and the furthest reaching 24 edits.
- Limiting to 10 edits reduced search time to 7 mins.
- Worst case “wavefront” size was 6,624,624 nodes which occurred at distance/iteration nine.

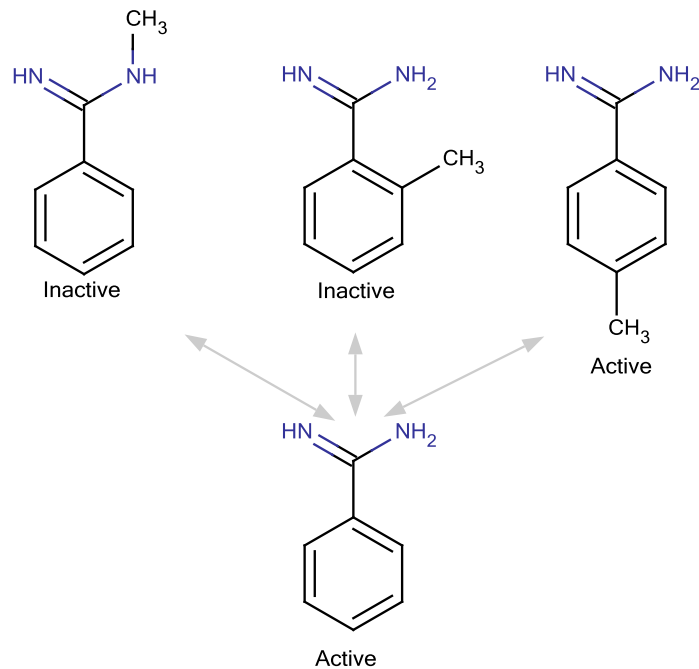


# CONNECTED VS. DISCONNECTED MCS

- A frequent question which maximum common subgraph methods is whether to perform connected or disconnected MCS.
- One solution to this dilemma with SmallWorld is to allow an additional “edit operation” that can eliminate a vertex of degree two.
- This captures the similarity between two molecules that differ in the length of a linker or by the size of a ring [classic challenges with MCS similarity].



# REPRESENTING ACTIVITY CLIFFS



A useful feature of graph edit distance is intuitive interpretation of  $GED(A,B) < GED(A,C)$ .



# CONCLUSIONS

- The sub-linear behaviour of SmallWorld's nearest neighbor calculation makes it faster than fingerprint-based similarity methods for sufficiently large data sets.
- This is thanks to the “blessing of dimensionality”.
- With continual advances in computer hardware, SmallWorld is likely to become the basis of most chemical similarity calculations within a decade.



# ACKNOWLEDGEMENTS

- AstraZeneca R&D for their support.
- Thank you for your time.
- Any questions?

