ACCELERATING GRAPH EDIT DISTANCE SEARCH BY CHEMICAL SPACE ENUMERATION

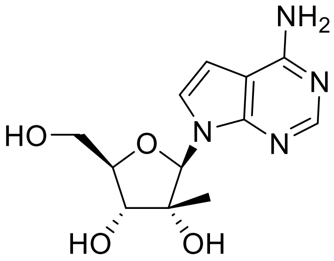
Roger Sayle, Richard Gowers and John Mayfield

NextMove Software, Cambridge, UK

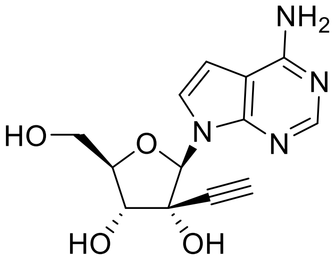
Small Organic Molecules Workshop, Oxford, UK, Tuesday 24th March 2020

MOTIVATION: CHEMICAL SIMILARITY

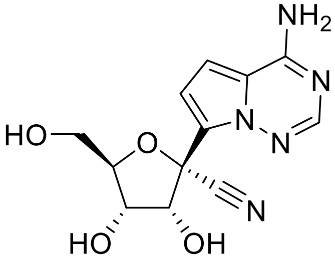
Evolution of SARS-CoV-2 (COVID-19) antivirals



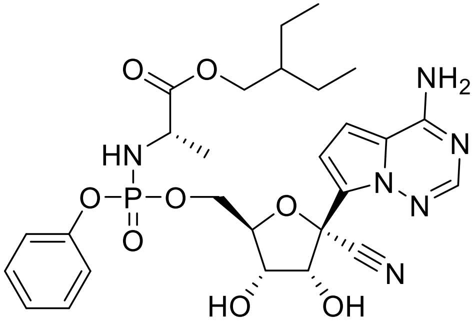
MK-608



NITD008



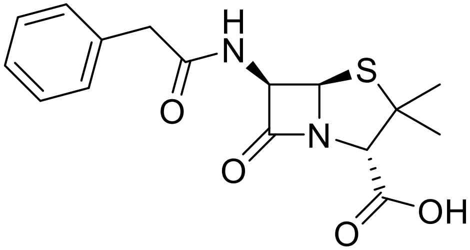
GS-441524



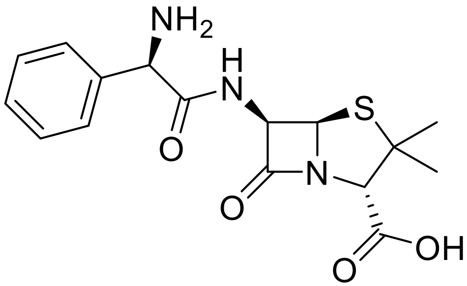
Remdesivir
(GS-5734)

US20160122356A1

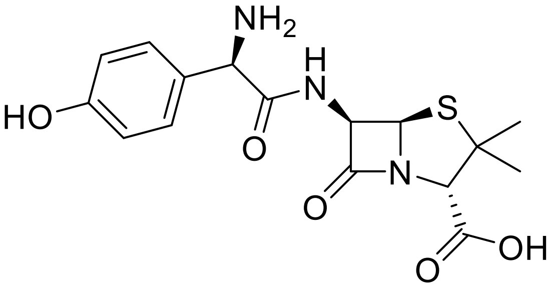
MOTIVATION: CHEMICAL SIMILARITY



Penicillin G (1942)



Ampicillin (1961)



Amoxicillin (1972)

EDIT DISTANCE

- **Edit Distance** is a measure of similarity (dissimilarity) between two discrete mathematical objects (formally a metric space).
  - String Edit Distance is a similarity metric between strings.
  - Tree Edit Distance is a similarity metric between trees.
  - **Graph Edit Distance** (GED) is a similarity metric between graphs.
- GED is the minimum number (or cost) of edit operations required to transform one graph into another.
- Edit operations consist of insertions, deletions and substitutions of nodes and edges (atoms and bonds).
- Unfortunately, computing GED is believed to be NP-Hard.

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.

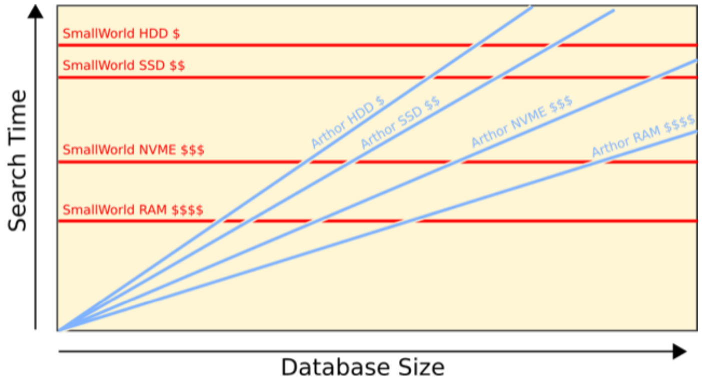
https://en.wikipedia.org/wiki/Graph_edit_distance

THE SMALLWORLD ALGORITHM

- SmallWorld is an algorithmic approach to accelerate graph edit distance searches on modern computer hardware.
- This approach makes heavy use of precomputation, increasing run-time performance at the expense of more storage space.
- A huge win from this trade-off is that database searches that used to scale linearly with increasing database size can now be performed in (near) constant time.
- "Fighting Big Data with bigger data".

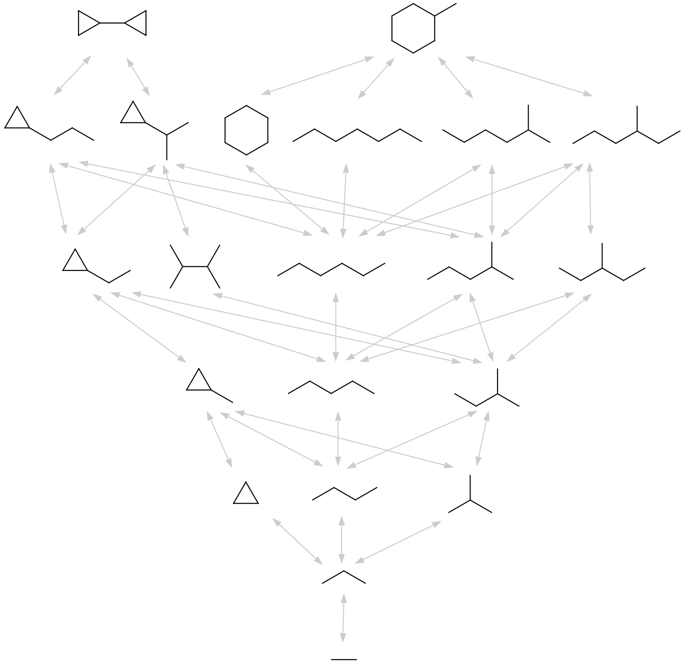
SmallWorld in Context

- Traditional binary fingerprint similarity and substructure searching of chemical databases scale (nearly) linearly with the size of the database.
  - At 2M compounds per second, searching ChEMBL takes 1s, searching PubChem takes 50s, and searching Enamine REAL over 10 minutes.
- Using SmallWorld, the top 100 search hits can be found in a few seconds independent of the size of a database.
  - UCSF's ZINC group regularly searches tens of billions of compounds.



A MAP OF CHEMICAL (GRAPH) SPACE

- The data structure underlying SmallWorld is a graph of graphs.
- Each vertex represents a molecule (with less than 99 bonds).
- Each edge represents an insertion or deletion edit operation.



- Currently contains 380 billion vertices and 2.8 trillion edges.

First of several trillion triples

```
**          *
***         **
**(*)*      ***
****        ***
*1**1       ***
**(*)(*)*   **(*)*
***(*)*     **(*)*
***(*)*     ****
*****       ****
**1**1      *1**1
**1**1      **(*)*
**1**1      ****
*1***1      ****
*1***1      *1**1
```

One representation of a SmallWorld graph index is as an tab-delimited ASCII text file, with two SMILES strings per line.

Such a file would contain 2,756,346,958,754 lines.

and be 268.59TB in length.

14TB when gzip compressed.

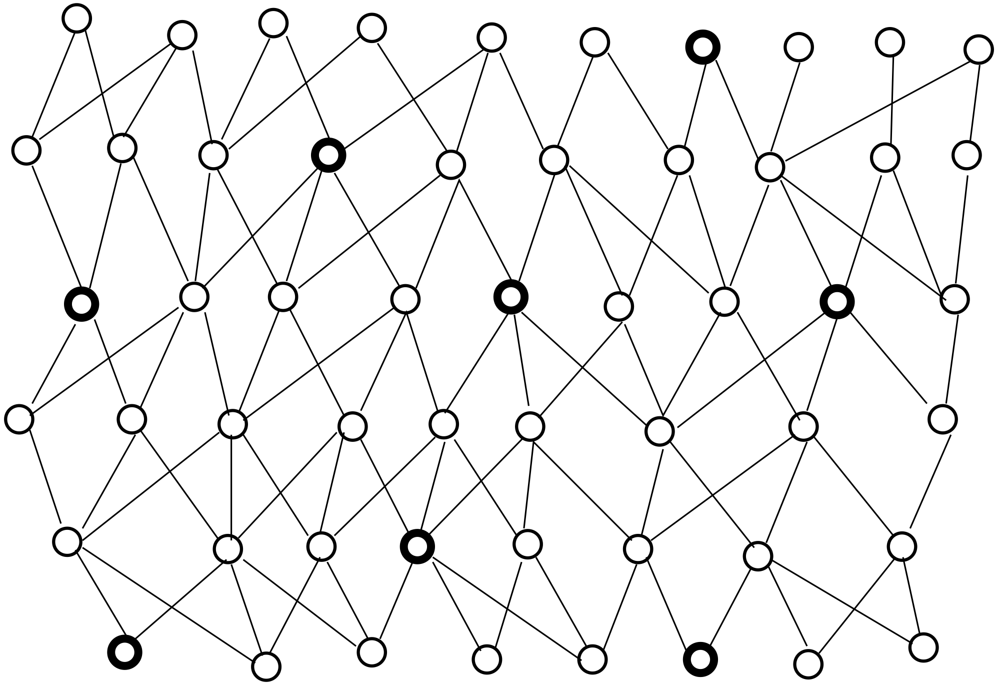
Caveats and Disclaimers

- Chemical (graph) space is infinite.

  "Space is big. You just won't believe how vastly, hugely, mind-bogglingly big it is. I mean, you may think it's a long way down the road to the drug store, but that's just peanuts to space." - Douglas Adams, HHGTTG.

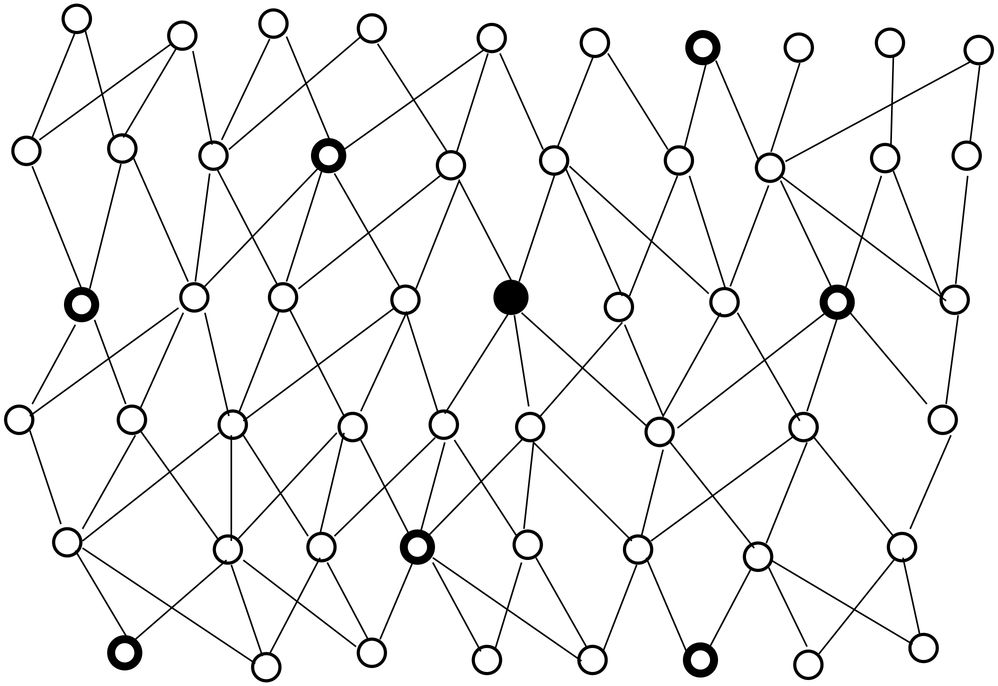
- Chemical graph space smaller than 100 (any fixed number of) bonds is finite, but impractical.
- Fortunately, we only care about subgraphs of molecules in our database, rather than all of theoretical graph space (GDB).
- Previously SmallWorld also had a maximum degree bound, no atoms with more than 4 neighbours, but this restriction has been lifted to support inorganics and groups such as $-SF_5$.
- Fortunately, Hasse networks are robust.

SMALLWORLD SEARCH



SmallWorld lattice: Circles represent virtual subgraphs, bold circles denote molecules mapped to subgraphs.

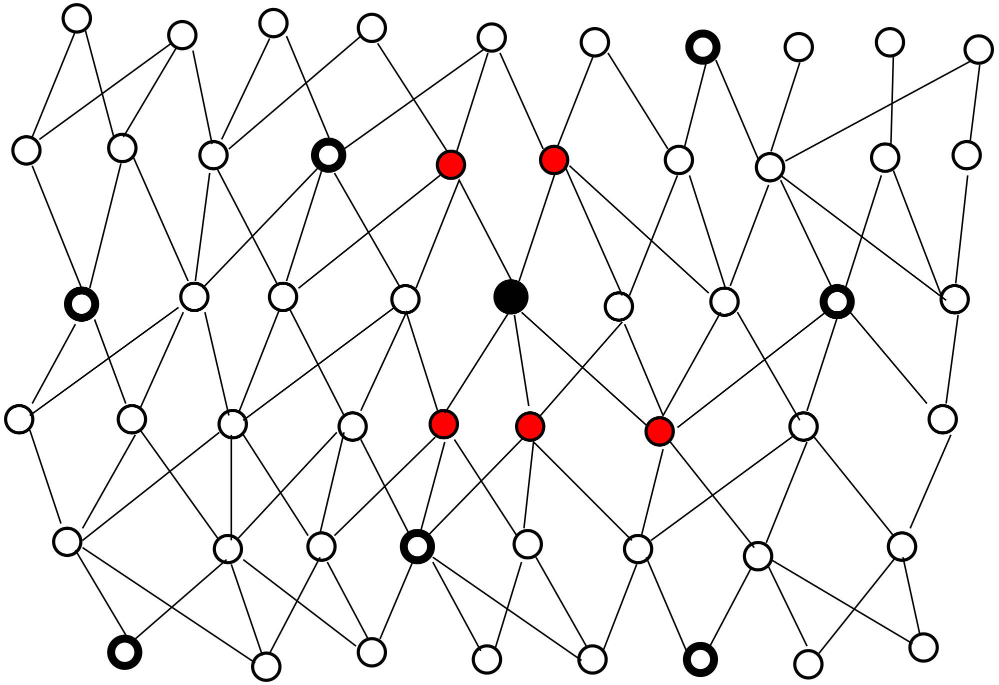
SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1

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

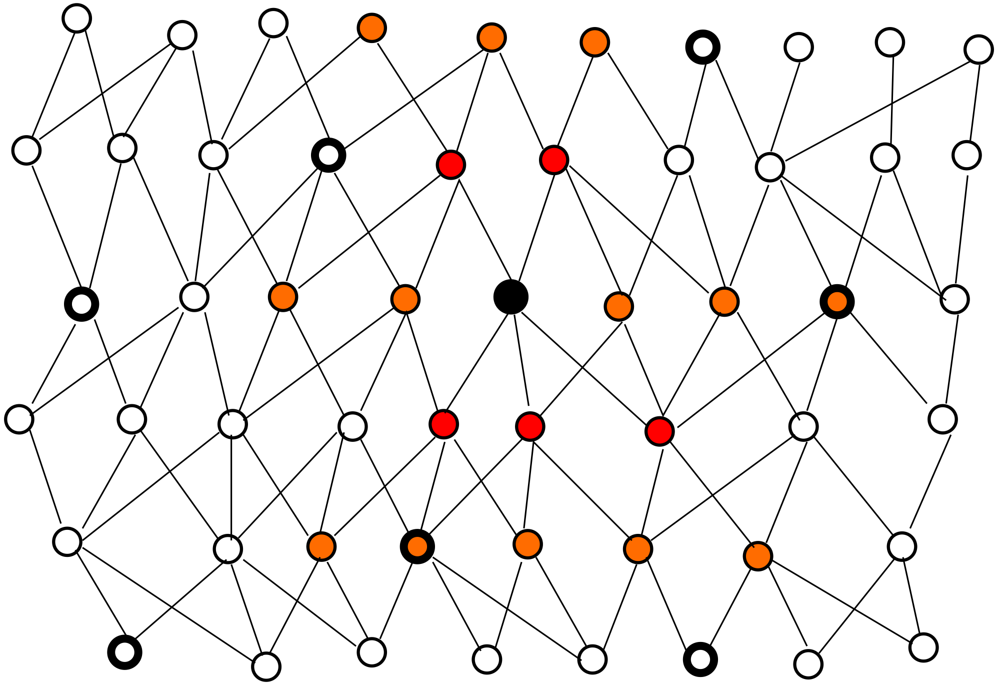
SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1
1	5	0	1

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

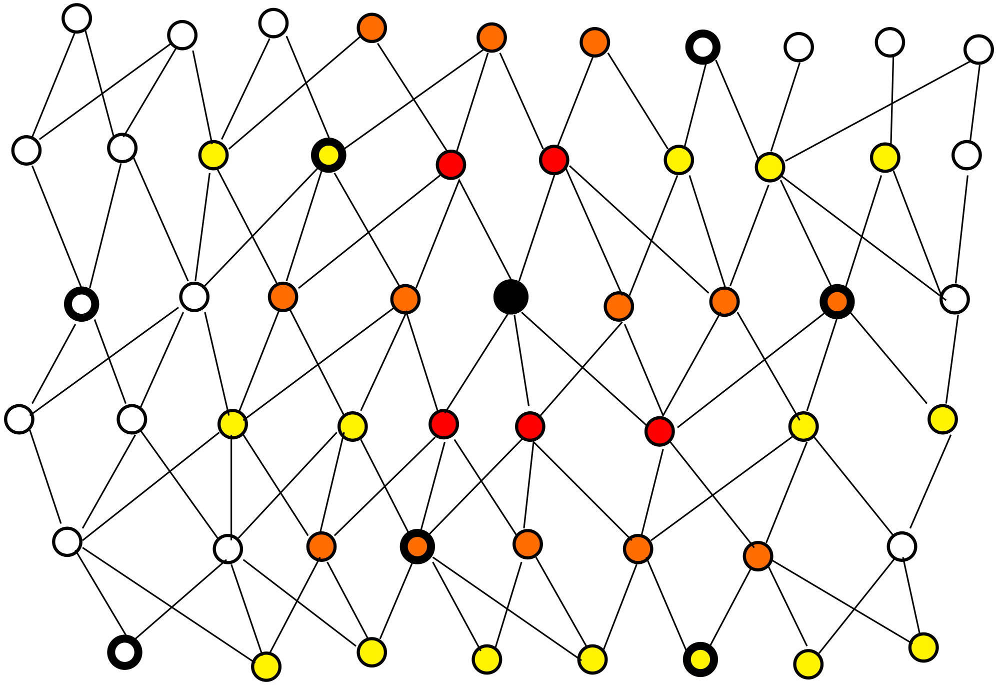
SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1
1	5	0	1
2	13	2	3

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

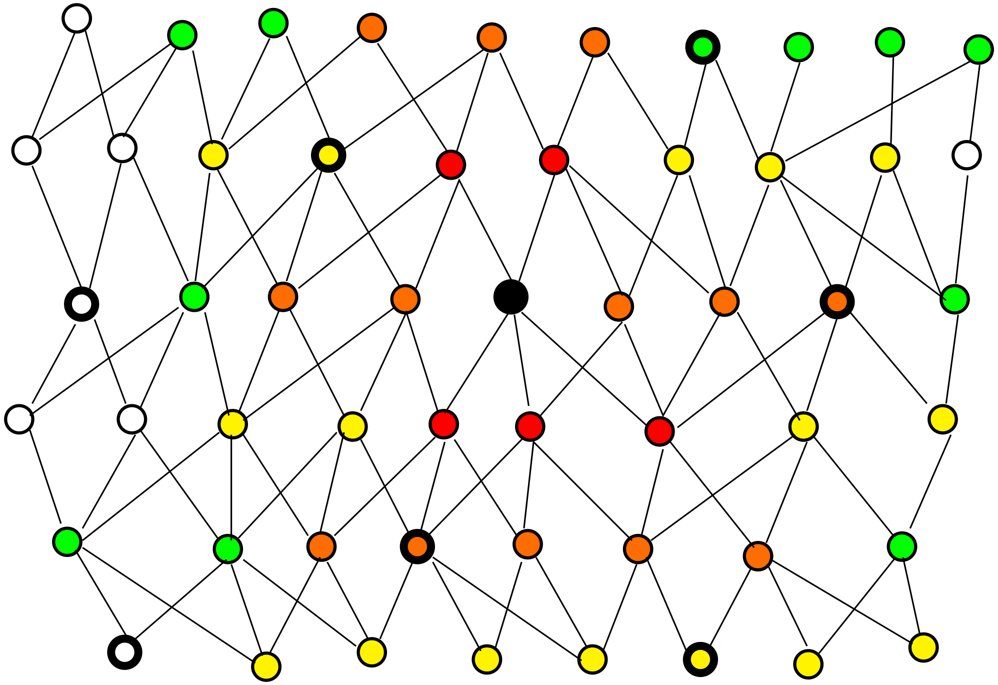
SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1
1	5	0	1
2	13	2	3
3	16	2	5

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

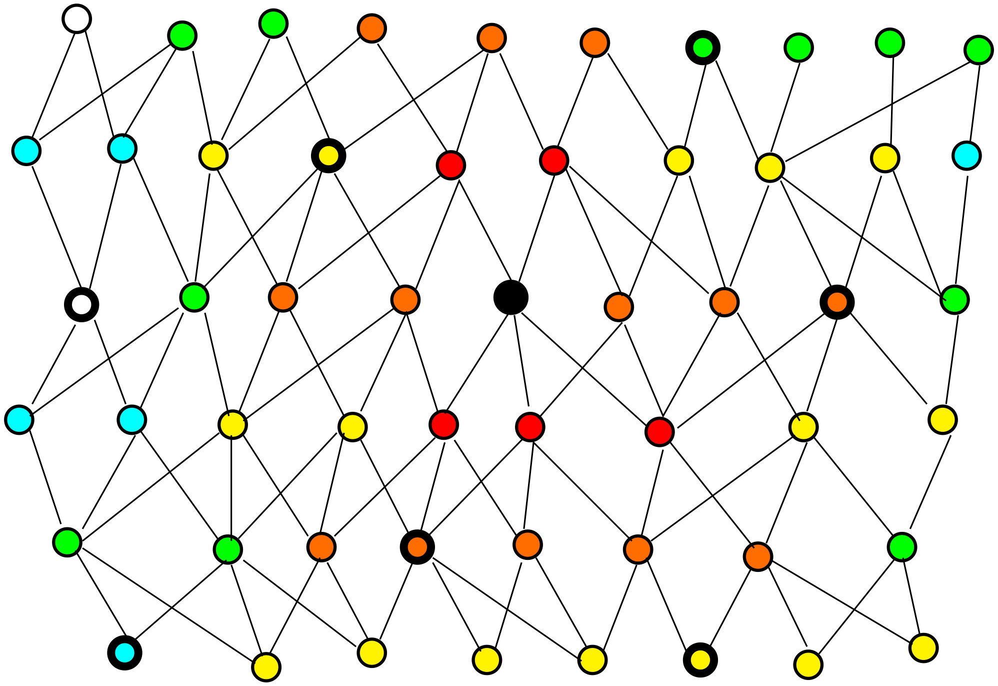
SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1
1	5	0	1
2	13	2	3
3	16	2	5
4	11	1	6

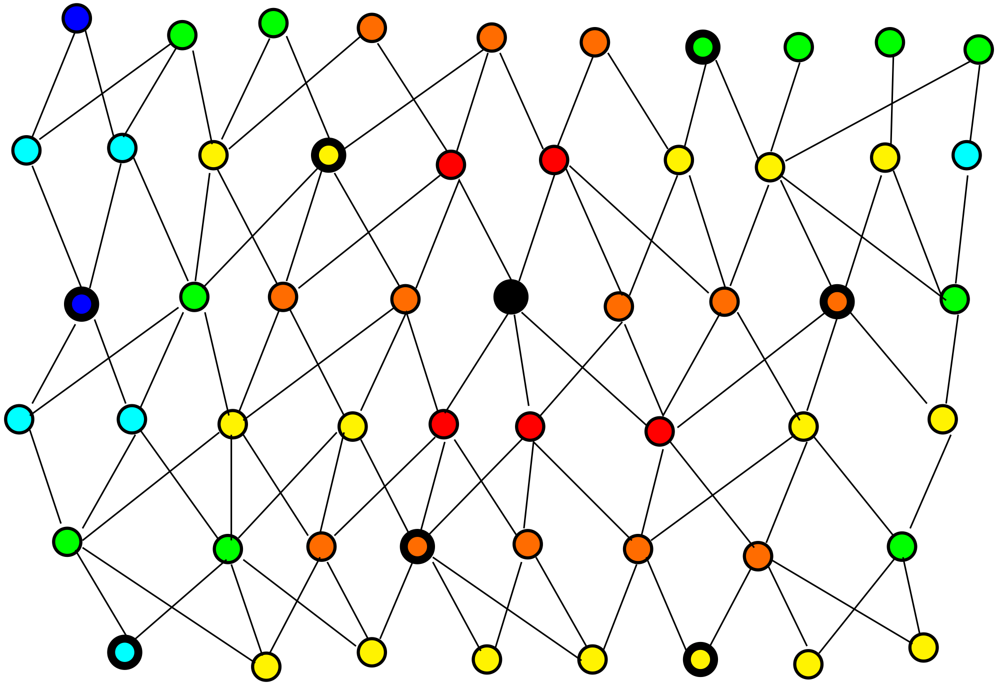
The search terminates once sufficient mapped neighbours have been found (or a suitable iteration limit is reached).

SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1
1	5	0	1
2	13	2	3
3	16	2	5
4	11	1	6
5	6	1	7

SMALLWORLD SEARCH



Dist	WF	New	Hits
0	1	1	1
1	5	0	1
2	13	2	3
3	16	2	5
4	11	1	6
5	6	1	7
6	2	1	8

The use of breadth-first (or best-first) search is similar to the Graph500 benchmark of supercomputers, measured in TEPS.

https://graph500.org/

MILEAGE CHART ANALOGY

- SmallWorld is a domain index (like GPS), unlike the instance indexes found in database systems.
- A mileage chart can lookup distance between chosen mapped cities, and approximate other distances.

Cambridge					
352.9	Edinburgh				
65.3	396.0	London			
171.0	217.5	204.4	Manchester		
84.0	365.7	56.1	160.7	Oxford	
132.5	430.2	79.9	225.2	66.3	Southampton

NEXTMOVE SOFTWARE'S IMPLEMENTATION

- The preceding theoretical description should be sufficient to implement a SmallWorld system for performing Graph Edit Distance searches.
- In theory, the 2.8 trillion rows/triples could be loaded in Neo4J or Oracle and queried with SPARQL.
- The rest of this presentation covers the many clever implementation details that when combined allow for very efficient chemical database searching.

GRAPH CANONICALIZATION

- The most important ingredient is canonical SMILES.
- Bounded degree (chemical) isomorphism is significantly easier than general case.
- The existence of canonical forms changes everything.

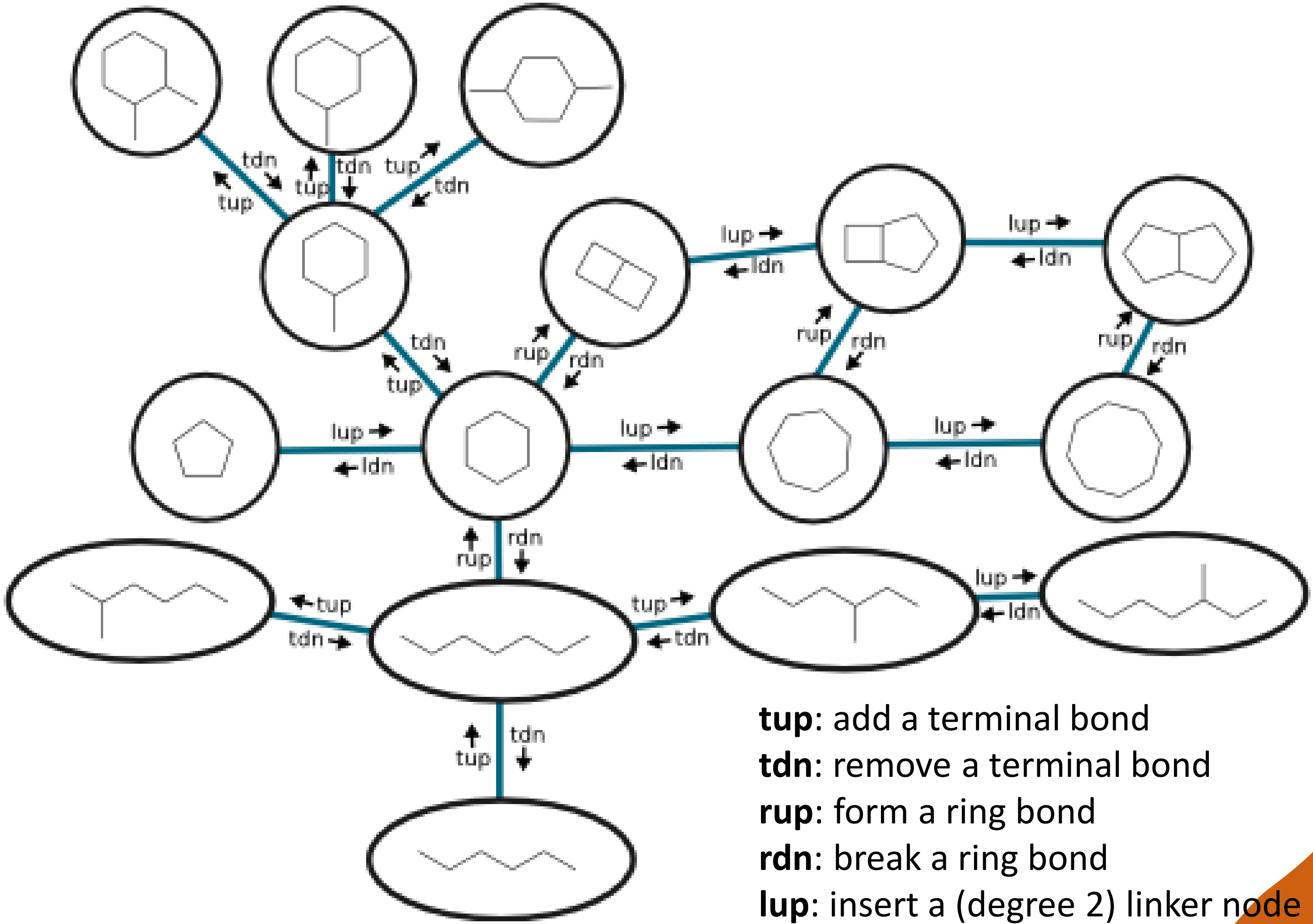
– RDKit 2019	6815 mol/s
– InChI (Open Babel)	7320 mol/s
– Open Babel	10.3 Kmol/s
– OpenEye OEChem	50 Kmol/s
– SWChem	113 Kmol/s

Schneider, Sayle and Landrum, “Get Your Atoms in Order-An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm”, J. Chem. Inf. Model. 55(10):2111-2120, 2015.

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.
- Assigning cyclic vs. acyclic bonds can be done efficiently in O(N) time, and this only needs to be recalculated after deleting a ring bond, as deleting terminal bonds doesn’t affect ring membership.
- An “UndeleteBond” function is also beneficial.

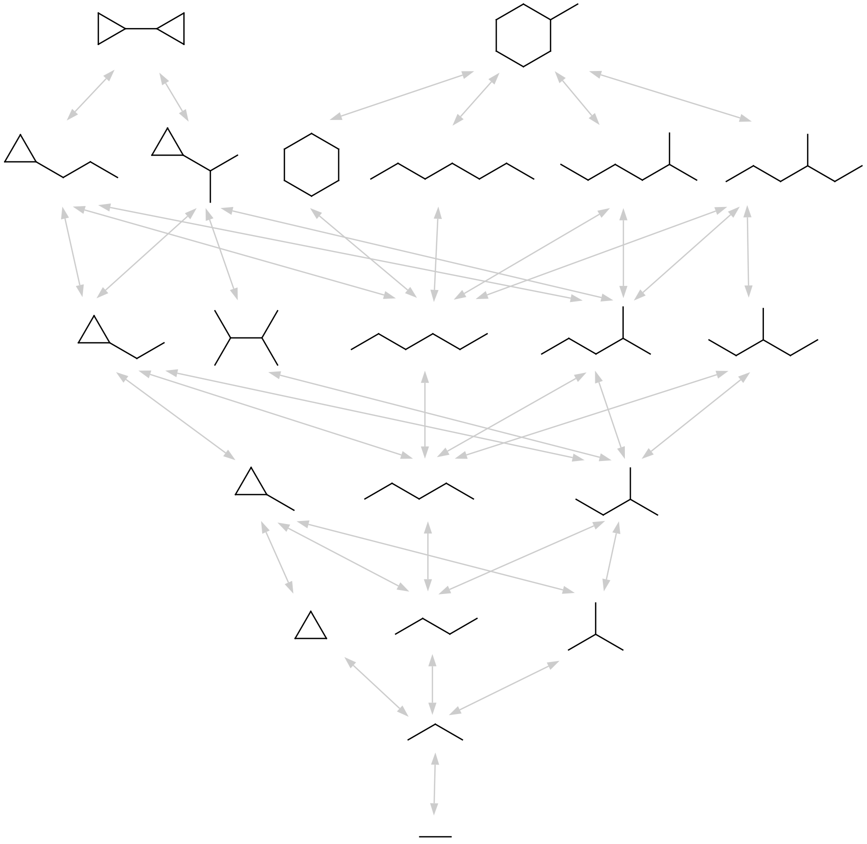
TOPOLOGICAL EDIT/EDGE TYPES



tup: add a terminal bond
tdn: remove a terminal bond
rup: form a ring bond
rdn: break a ring bond
lup: insert a (degree 2) linker node
ldn: remove a (degree 2) linker node

Partitioning Vertices

Chemical/Graph space (is bipartite and) may be partitioned by the number of bonds.



7 bonds

6 bonds

5 bonds

4 bonds

3 bonds

2 bonds

1 bond

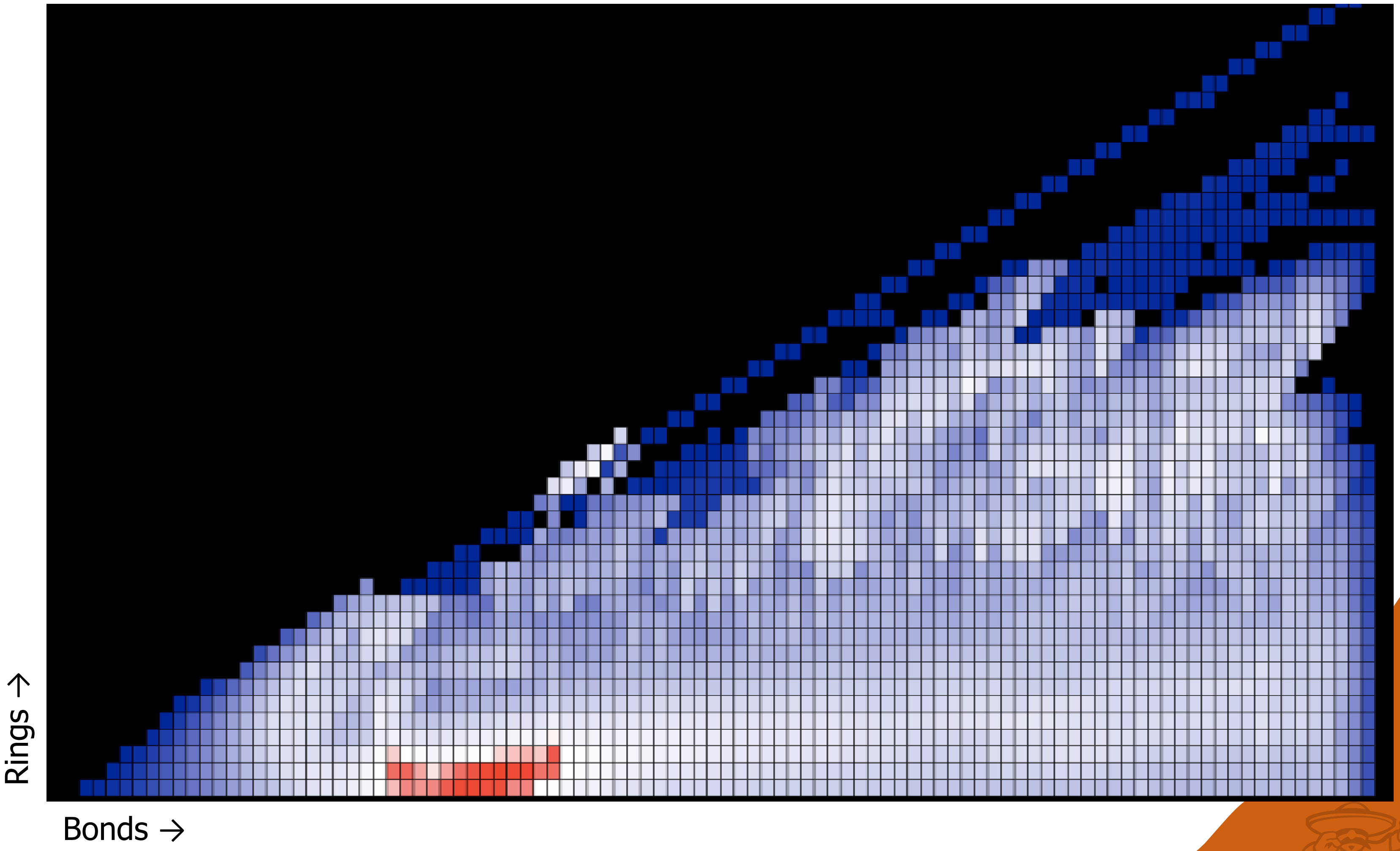
DATABASE PARTITIONING

- SmallWorld is actually partitioned by atoms, bonds, and rings [using the equation A=(B+1)-R].
- This results in 2842 partitions, named B_xR_y where x is the number of bonds, y is the number of rings.
- Each edge links vertices in neighbouring partitions.
  - A tdn edge from B_xR_y leads to $B_{x-1}R_y$, tup to $B_{x+1}R_y$.
  - A rdn edge from B_xR_y leads to $B_{x-1}R_{y-1}$, rup to $B_{x+1}R_{y+1}$.
  - A ldn edge from B_xR_y leads to $B_{x-1}R_y$, lup to $B_{x+1}R_y$.

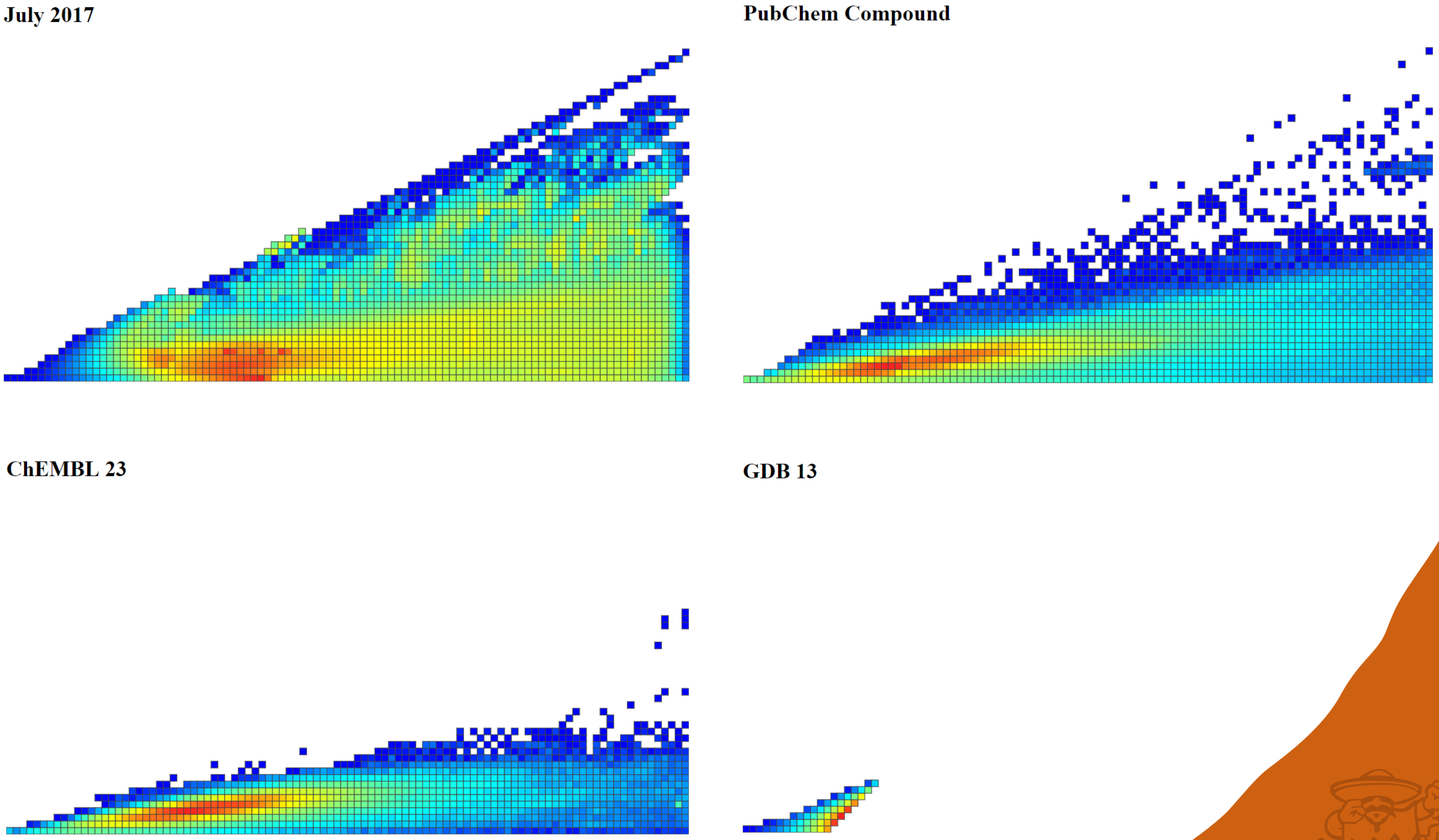
FIRST OF SEVERAL TRILLION TRIPLES

**	*	B1R0/tdn
***	**	B2R0/tdn
**(*)*	***	B3R0/tdn
****	***	B3R0/tdn
*1**1	***	B3R1/rdn
**(*)(*)*	**(*)*	B4R0/tdn
***(*)*	**(*)*	B4R0/tdn
***(*)*	****	B4R0/tdn
*****	****	B4R0/tdn
11	*1**1	B4R1/tdn
11	**(*)*	B4R1/rdn
11	****	B4R1/rdn
*1***1	****	B4R1/rdn
*1***1	*1**1	B4R1/ldn

MAP OF SMALLWORLD SPACE



SmallWorld Density Heatmaps



July 2017

PubChem Compound

ChEMBL 23

GDB 13

NUMBERING AND NAMING VERTICES

- The graphs in each partition (all having the same number of atoms, bonds and rings) are (arbitrarily) numbered sequentially from one.
- Hence any vertex may be referenced by ID: BxRy.Z

– Penicillin G	B25R3.284481020
– Ampicillin	B26R3.489483828
– Amoxicillin	B27R3.40995378

- Each edge can therefore be represented as a pair of integers, the src index and the dst index.

VERTEX NUMBERING IN B6R1

- Mapping from graphs to indices in B6R1 looks like:

1(*1)* 1

1*1* 2

11 3

****1**1 4

1**1 5

(*)*11 6

*1*****1 7

***1**1* 8

***1(**1)* 9

11(*)* 10

1(*1)* 11

1(1)(*)* 12

**1*(*1*)* 13

SPEEDING UP THE MAPPING PROCESS

- Determining the vertex ID for a given molecule is essentially a key-value pair (dictionary) lookup from canonical SMILES.
- Early versions of SmallWorld used binary search of a alphabetically sorted text file; faster than a RDBMS.
- The current implementation uses three refinements:
  - Custom multigram SMILES compression
    - https://www.daylight.com/meetings/mug01/Sayle/SmiZip/sld001.htm
  - Multiplicative Binary Search
    - https://en.wikipedia.org/wiki/Multiplicative_binary_search
  - Key-length partitioning for fixed length binary search.

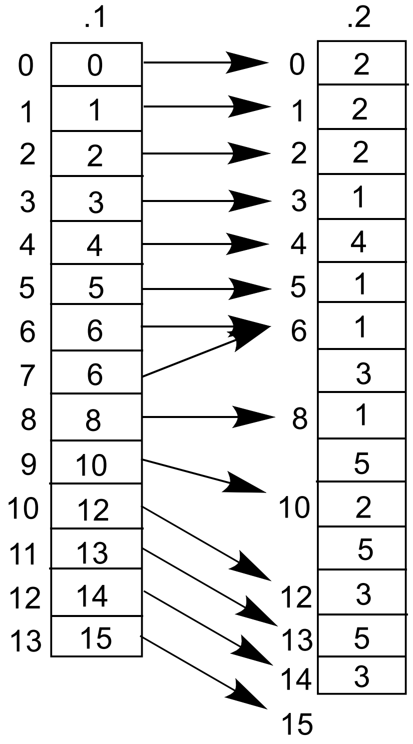
Storing Vertices

- Using the techniques on the previous slide, the graphs of all 380 billion vertices are stored in only 4.2TB, or around 12 bytes per graph/SMILES.

STORING EDGES: THE PRESENT

- Directional edges are stored in Compressed Sparse Row (CSR) format. https://en.wikipedia.org/wiki/Sparse_matrix

15 edges
1-2
2-2
3-2
4-1
5-4
6-1
8-1
8-3
9-1
9-5
10-2
10-5
11-3
12-5
13-3



	.1
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	6
8	8
9	10
10	12
11	13
12	14
13	15

	.2
0	2
1	2
2	2
3	1
4	4
5	1
6	1
	3
8	1
	5
10	2
	5
12	3
13	5
14	3
15	

Storing Edges: The Future

- Currently both tables use 5-byte (40 bit) pointers, as the max vertices/edges in a partition is >2^{32} and <2^{40}.
- The next iteration of SmallWorld will use packed integers, using custom widths for each table.
- As the typical fanout is low, the values in the pointer table are almost random, but first table is sorted, which allows for further compression.
  - By maintaining a directory of every N^{th} value, every (other) value may be represented as a (smaller) delta from that previous reference, allowing compression+random access.

CURRENT DATABASE STATISTICS

- As of March 2020, the SmallWorld index has
- 380,162,460,266 nodes (~380B or ~2^{38} nodes)
- 2,756,346,958,754 edges (~2.8T or ~2^{42} edges)
  - 1,472,058,112,318 ring edges.
  - 752,057,044,898 terminal edges
  - 532,231,801,538 linker edges.
- Average degree (fan-out) of node: ~14
- Runtime index requires 40TB of disk space.

Search Algorithms

- Chemical similarity may be implemented using either breadth-first (BFS) or best-first search.
- Searches that only follow tup and rup edges implement “substructure” search.
- Searches that only follow tdn and rdn edges implement “superstructure” search.
- Searches that only follow tdn and tup edges find hits with the same Bemis-Murcko scaffold.

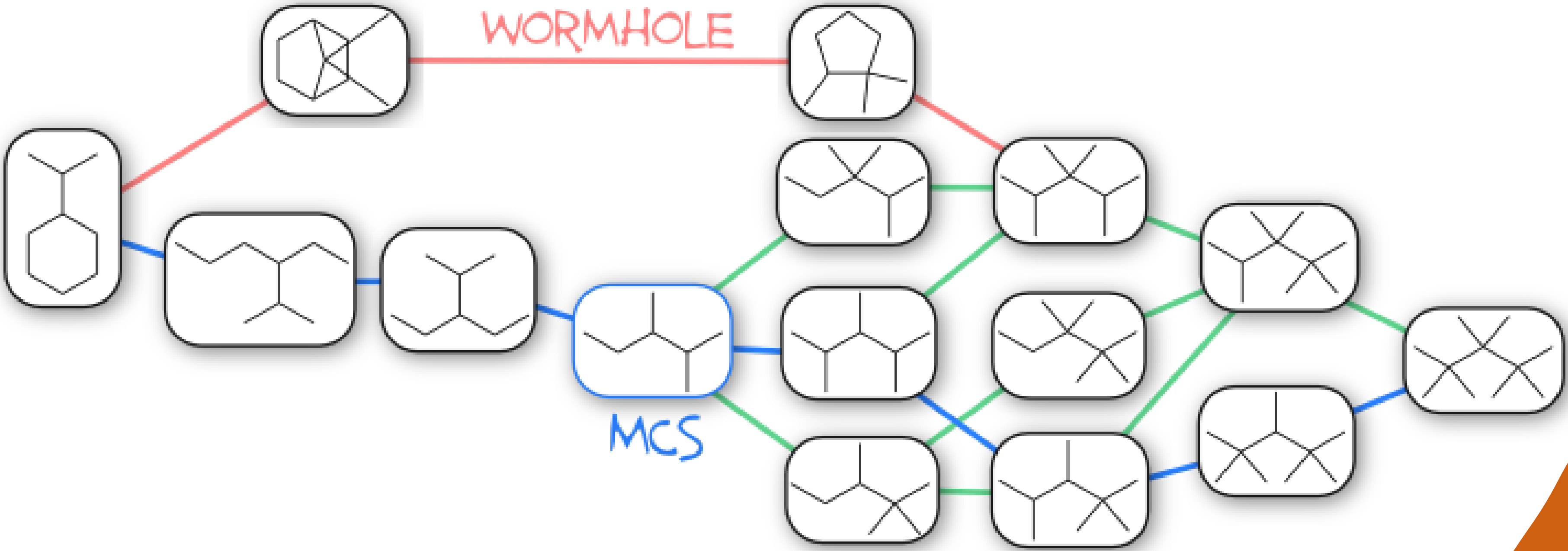
SHORTEST PATH ALGORITHMS

- Finding the graph edit distance between a specified pair of molecules reduces to finding the shortest path between them.
- A well known improvement in computer science is to use bidirectional search to reduce the search from $O(b^d)$ to $O(b^{d/2})$ where b is the branching factor.
  - https://en.wikipedia.org/wiki/Bidirectional_search
- Less widely known improvement is the variant of bidirectional search that at each iteration advances the smaller wavefront.

ALGEBRA OF GRAPH EDIT OPERATIONS

- Symmetries with a SmallWorld network mean that there are often multiple paths (of the same distance) between a pair of vertices, and this can be useful in improving search performance by pruning edges.
- Perform all down edges before any up edges.
- Such paths pass through the MCES (inflection point).
- Likewise perform rdn before tdn, and rup after tup.
  - Intuitively tdns don’t affect rdns, but rdn may create tdns.
- All paths look like [rdn*][tdn*][ldn*][lup*][tup*][rup*].

BEWARE OF DALKE WORMHOLES



Shorter paths (called wormholes) may exist going via a minimal common superstructure; counter-intuitive to chemical similarity these may have applications in synthesis.

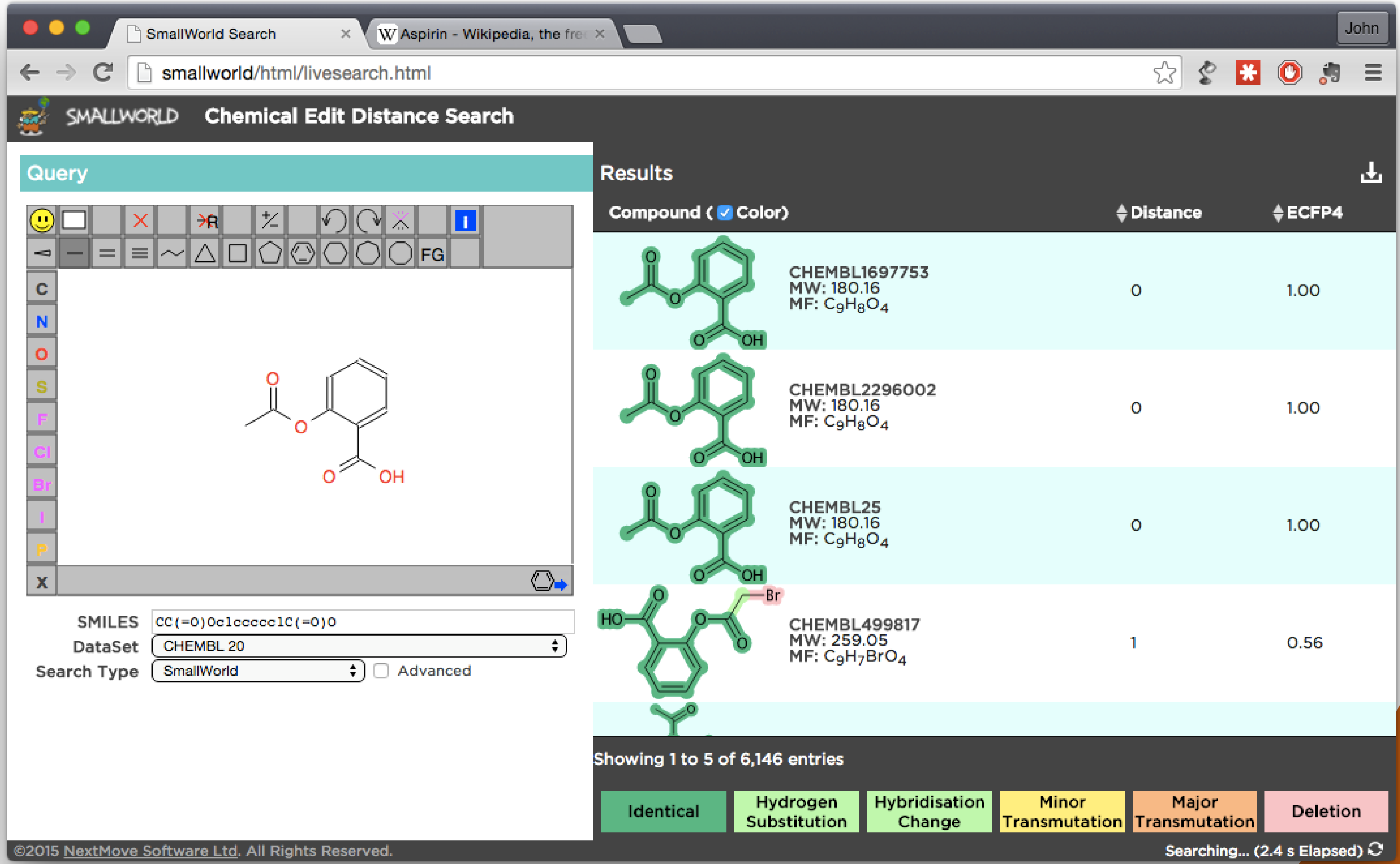
And finally… Bloom Filter Joins

- As the BFS advances, each visited vertex needs to check the database molecules mapped to it (hits).
- This is effectively a database join (intersection) between the vertices and the mapped molecules.
- Each lookup is efficient using the binary search techniques described previously, but a large fraction of lookups are unproductive (find no hits).
- To improve performance we use a bloom filter as a fast pre-screen, reducing the number of lookups.

SMALLWORLD

Chemical Edit Distance Search

NextMove Software Ltd
https://www.nextmovesoftware.com



SMILES CC(=O)Oc1ccccc1C(=O)O

DataSet CHEMBL 20

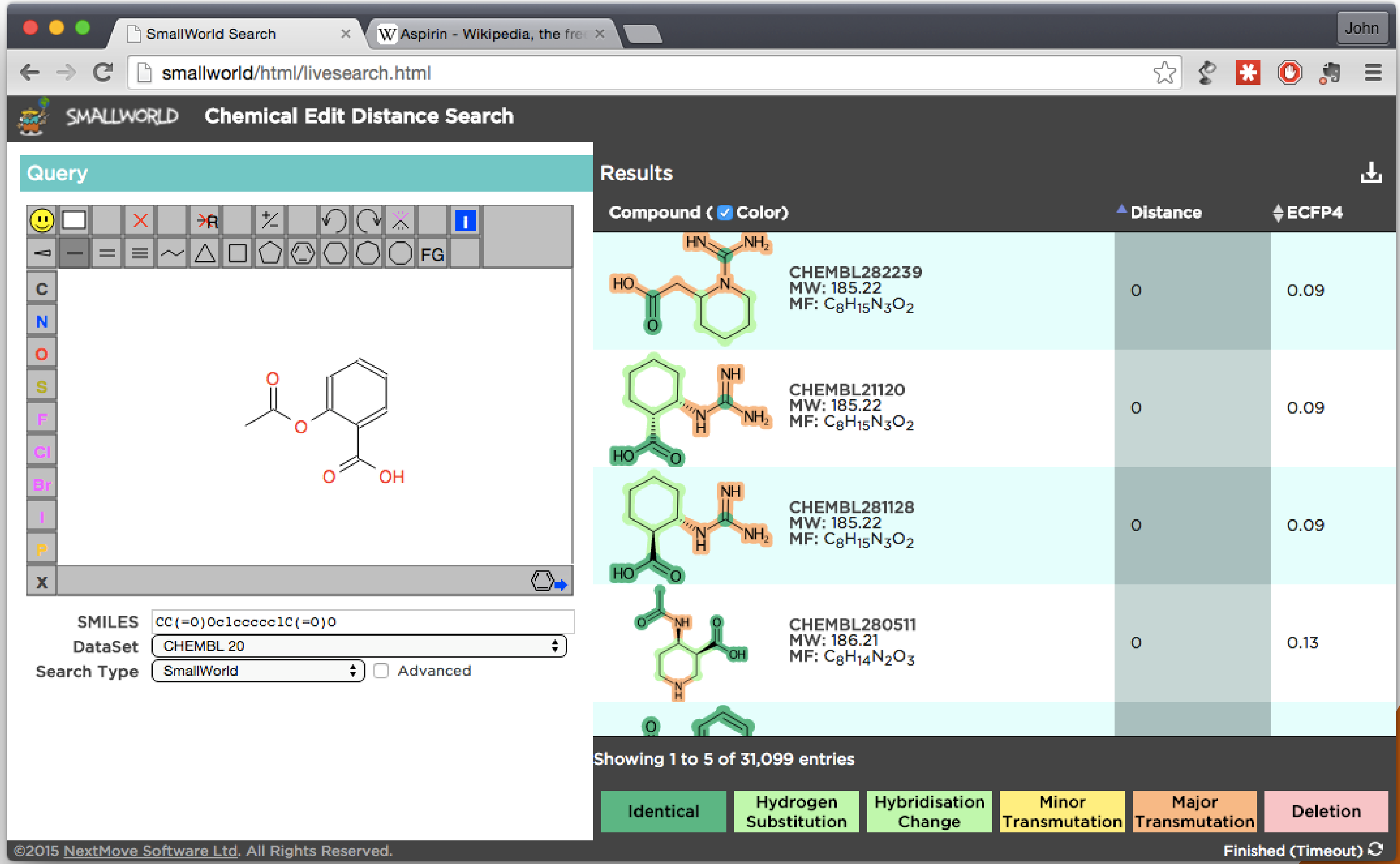
Search Type SmallWorld ☐ Advanced

Compound (☑ Color)	Distance	ECFP4
CHEMBL1697753 MW: 180.16 MF: $C_9H_8O_4$	0	1.00
CHEMBL2296002 MW: 180.16 MF: $C_9H_8O_4$	0	1.00
CHEMBL25 MW: 180.16 MF: $C_9H_8O_4$	0	1.00
CHEMBL499817 MW: 259.05 MF: $C_9H_7BrO_4$	1	0.56

Showing 1 to 5 of 6,146 entries

©2015 NextMove Software Ltd. All Rights Reserved.

Searching... (2.4 s Elapsed)



Chemical Edit Distance Search

Query

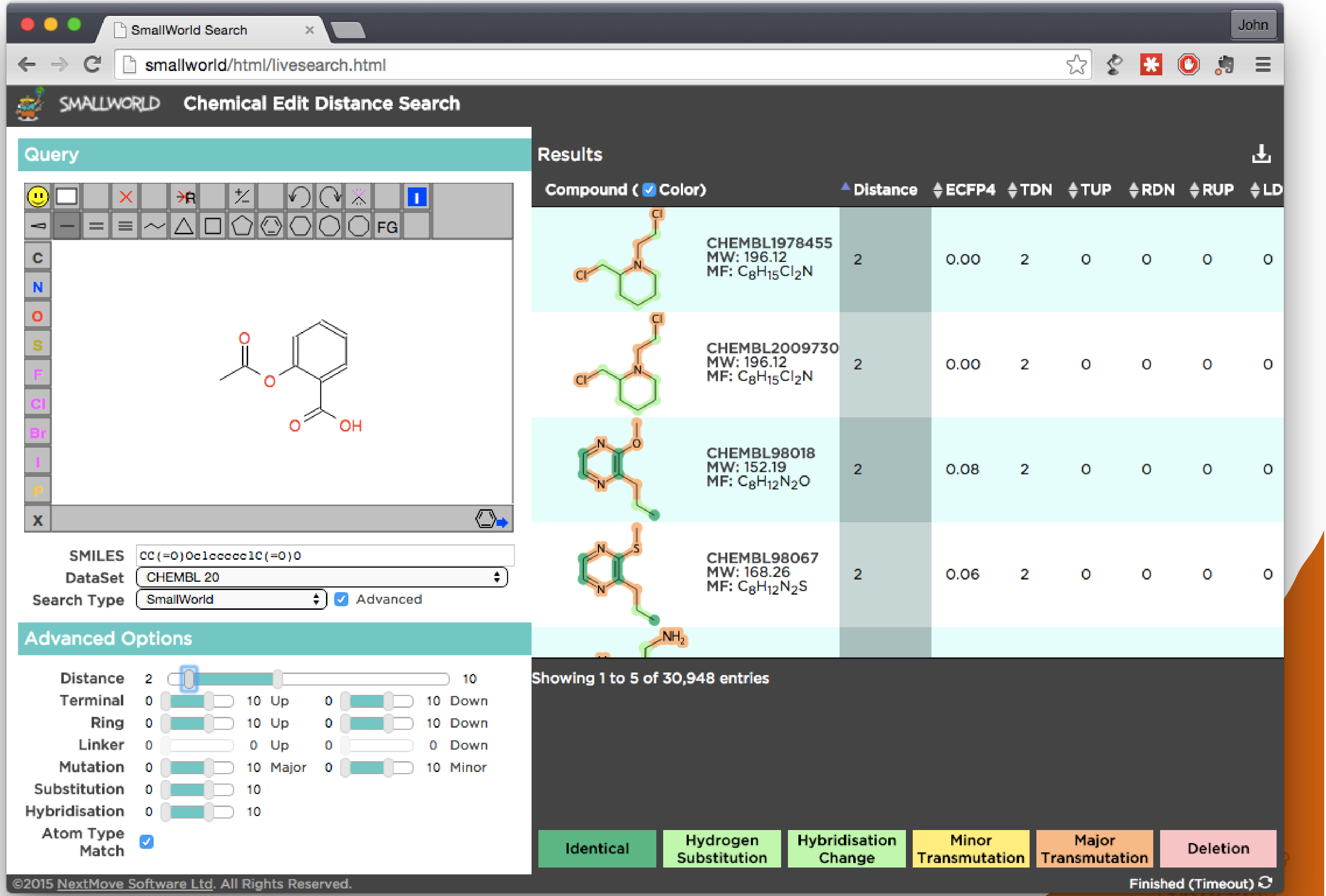
SMILES: CC(=O)Oc1ccccc1C(=O)O
DataSet: CHEMBL 20
Search Type: SmallWorld
☐ Advanced

Results

Compound (☑ Color)	▲Distance	ECFP4
CHEMBL282239 MW: 185.22 MF: $C_8H_{15}N_3O_2$	0	0.09
CHEMBL21120 MW: 185.22 MF: $C_8H_{15}N_3O_2$	0	0.09
CHEMBL281128 MW: 185.22 MF: $C_8H_{15}N_3O_2$	0	0.09
CHEMBL280511 MW: 186.21 MF: $C_8H_{14}N_2O_3$	0	0.13

Showing 1 to 5 of 31,099 entries

Finished (Timeout)



Chemical Edit Distance Search

Query

SMILES CC(=O)Oc1ccccc1C(=O)O
DataSet CHEMBL 20
Search Type SmallWorld ☑ Advanced

Advanced Options

Distance	2		10				
Terminal	0	10	Up	0	10	Down	
Ring	0	10	Up	0	10	Down	
Linker	0	0	Up	0	0	Down	
Mutation	0	10	Major	0	10	Minor	
Substitution	0	10					
Hybridisation	0	10					
Atom Type Match	☑						

Results

Compound (☑ Color)	Distance	ECFP4	TDN	TUP	RDN	RUP	LD
CHEMBL1978455 MW: 196.12 MF: $C_8H_{15}Cl_2N$	2	0.00	2	0	0	0	0
CHEMBL2009730 MW: 196.12 MF: $C_8H_{15}Cl_2N$	2	0.00	2	0	0	0	0
CHEMBL98018 MW: 152.19 MF: $C_8H_{12}N_2O$	2	0.08	2	0	0	0	0
CHEMBL98067 MW: 168.26 MF: $C_8H_{12}N_2S$	2	0.06	2	0	0	0	0

Showing 1 to 5 of 30,948 entries

Identical | Hydrogen Substitution | Hybridisation Change | Minor Transmutation | Major Transmutation | Deletion

©2015 NextMove Software Ltd. All Rights Reserved.

Finished (Timeout)

CONCLUSIONS

- The growth in storage capacity of modern hardware allows enumeration of graph space to accelerate graph edit distance search in ways that were impossible just a few years ago.
- The resulting sublinear (constant time) searches avoid the pending apocalypse caused by the growth of virtual on-demand databases.

ACKNOWLEDGEMENTS

- In memoriam Andy Grant, thank you for everything.
- AstraZeneca R&D, Alderley Park, U.K.
- GlaxoSmithKline, Stevenage, U.K.
- Relay Therapeutics, Boston, U.S.A.
- Eli Lilly, Indianapolis, U.S.A.
- Hoffmann-La Roche, Basel, Switzerland.
- John Irwin, ZINC group, UCSF, San Francisco, U.S.A.
- Catherine Wong, Deane Group, University of Oxford, U.K.
- Jose Batista, OpenEye Scientific Software, Germany.
- Jameed Hussain, Dotmatics Limited, U.K.

- Thank you for your time, Any questions?

J. ANDREW GRANT (1963-2012)

Me and Andy at OpenEye EuroCUP 2008

COUNTING MOLECULAR SUBGRAPHS

Name	Atoms	MW	Subgraphs
Benzene	6	78	7
Cubane	8	104	64
Ferrocene	11	186	3,154
Aspirin	13	180	127
Dodecahedrane	20	260	440,473
Ranitidine	21	314	436
Clopidrogel	21	322	10,071
Morphine	21	285	176,541
Amlodipine	28	409	58,139
Lisinopril	29	405	24,619
Gefitinib	31	447	190,901
Atorvastatin	41	559	3,638,523

≤ Bond Count	%PubChem
≤ 20 bonds	14%
≤ 25 bonds	30%
≤ 30 bonds	55%
≤ 35 bonds	77%
≤ 40 bonds	89%
≤ 45 bonds	93%
≤ 50 bonds	95%
≤ 55 bonds	97%
≤ 60 bonds	98%
≤ 65 bonds	98%
≤ 70 bonds	99%