ADVANCES IN SEARCHING ULTRA-LARGE (100+ BILLION) COMPOUND CHEMICAL DATABASES: ARTHOR & SMALLWORLD

Roger Sayle, Richard Gowers and John Mayfield

NextMove Software, Cambridge, UK

ZINC20 PAPER IN JCIM

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

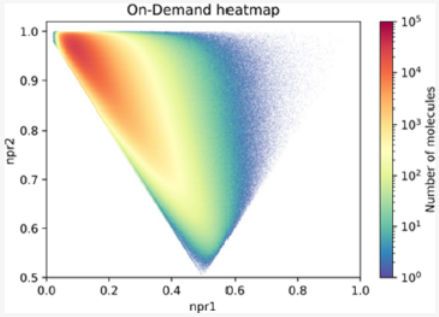
ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery

John J. Irwin,* Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle

Cite This: https://dx.doi.org/10.1021/acs.jcim.0c00675 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Identifying and purchasing new small molecules to test in biological assays are enabling for ligand discovery, but as purchasable chemical space continues to grow into the tens of billions based on inexpensive make-on-demand compounds, simply searching this space becomes a major challenge. We have therefore developed ZINC20, a new version of ZINC with two major new features: billions of new molecules and new methods to search them. As a fully enumerated database, ZINC can be searched precisely using explicit atomic-level graph-based methods, such as SmallWorld for similarity and Arthor for pattern and substructure search, as well as 3D methods such as docking. Analysis of the new make-on-demand compound sets by these and related tools reveals startling features. For instance, over 97% of the core Bemis−Murcko scaffolds in make-on-demand libraries are unavailable from "in-stock" collections. Correspondingly, the number of new Bemis−Murcko scaffolds is rising almost as a linear fraction of the elaborated molecules. Thus, an 88-fold increase in the number of molecules in the make-on-demand versus the in-stock sets is built upon a 16-fold increase in the number of Bemis−Murcko scaffolds. The make-on-demand library is also more structurally diverse than physical libraries, with a massive increase in disc- and sphere-like shaped molecules. The new system is freely available at zinc20.docking.org.



https://pubs.acs.org/doi/10.1021/acs.jcim.0c00675

ZINC15 BACK IN 2018...

Timed Out

Your query took too long.

We are working on a better solution! For now, Here are some things you may try (after ? in URL)

Ask for fewer (e.g. `count=1000`)
Turn off (possibly default) sorting (e.g. `sort=no`)
Turn off (possibly default) duplicate removal (e.g. `distinct=no`)

EVOLUTION VS. REVOLUTION

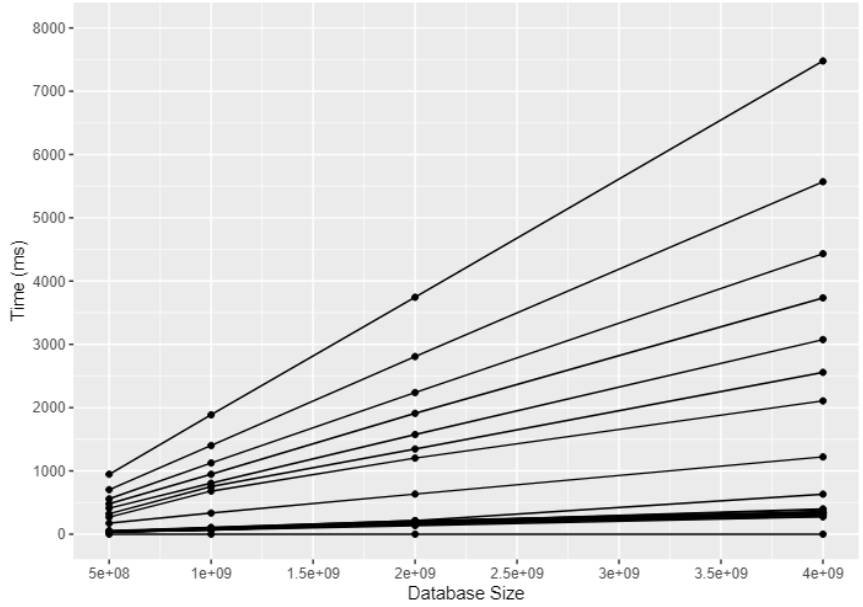
- Approaches to tackle the dramatic growth of chemical databases fall into two main strategies…

1. Push existing linear scaling technologies, such as binary fingerprint Tanimoto and SMARTS substructure search as fast as they'll go.
2. Develop innovative (significantly) sublinear scaling chemical search methods.

- Both these strategies take advantage of technical improvements in computer hardware, particularly storage technology.

WHAT'S THE ISSUE WITH O(N)?

- A few years ago, good performance ~4Mmol/s.
  - ChEMBL with 2 million compounds in half a second.
  - PubChem with 100 million, searched in 25s.
  - Enamine REAL/ZINC with 1.4 billion, takes nearly 6 mins.
- Today ~400 million mols per second, but still linear!



Timings of different SMARTS queries (Arthor running on kryten desktop) shows showing linear scaling with database size.

Worst case: 4B mols in under 8s.

COMPUTER HARDWARE 2020

	Size	Total Size	USD Price	GB/s	MB/s/$
GPU	48GB	x4=192GB	$22400	672	30
RAM	2TB		$21800	200	9.2
NVME	2TB	x3=6TB	$1281	32	25
SSD	7.68TB	x8=61TB	$10200	3	0.3
HD	18TB	x8=144TB	$4600	0.3	0.066
Network				0.12	

- For similarity searching, 1 billion molecules requires 34GB as 256 bit FPs, 68GB as 512 bit FPs, 136GB as 1024 bit FPs and 252GB as 2048 bit FPs.
- For substructure searching, 1 billion molecules, Arthor requires 219GB (~219 bytes per molecule) for optimal performance, or minimally 92GB (92 bytes per molecule).
- GPUs have ~3x bandwidth, but <10% capacity, at ~ same price.
- Hard disks have 72x capacity, at 20% the price.

BANDWIDTH IS KING

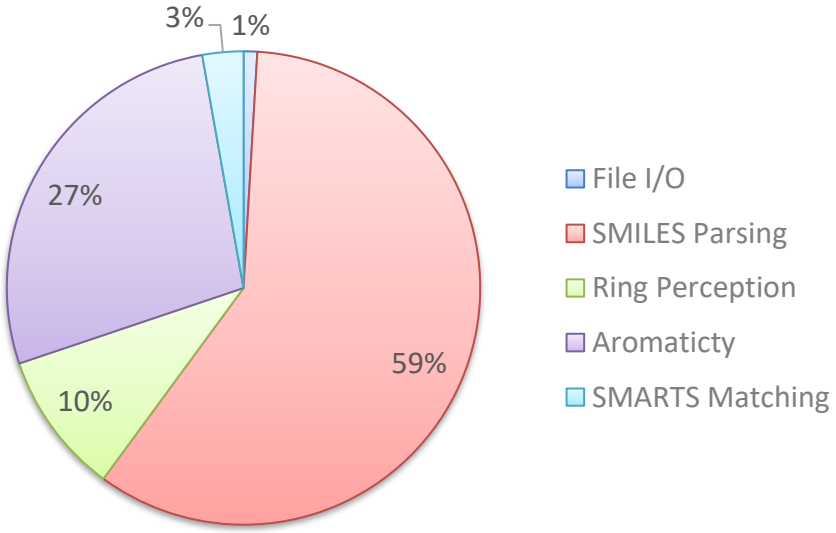
- For non-trivial data sets, that don't fit in cache, bandwidth becomes the performance bottleneck.
- Modern multi-core architectures assume a high computation density.
- On Intel/AMD chips, the number of memory channels and speed of the memory is sometimes more important than the number of cores.
- Bandwidth limits the value of cores on CPUs/GPUs.
- Bandwidth limits the value of nodes on a cloud.

NOT JUST A GOOD IDEA, IT'S THE LAW

- Kryten's DRAM is rated 2933 MHz, so on each clock light travels ~10 centimeters.

BANDWIDTH IN SMARTS SEARCHING 1

- Profiling reveals that majority of time spend in traditional state-of-the-art chemical database engines is not actually in the SMARTS matcher, but in representation conversion.

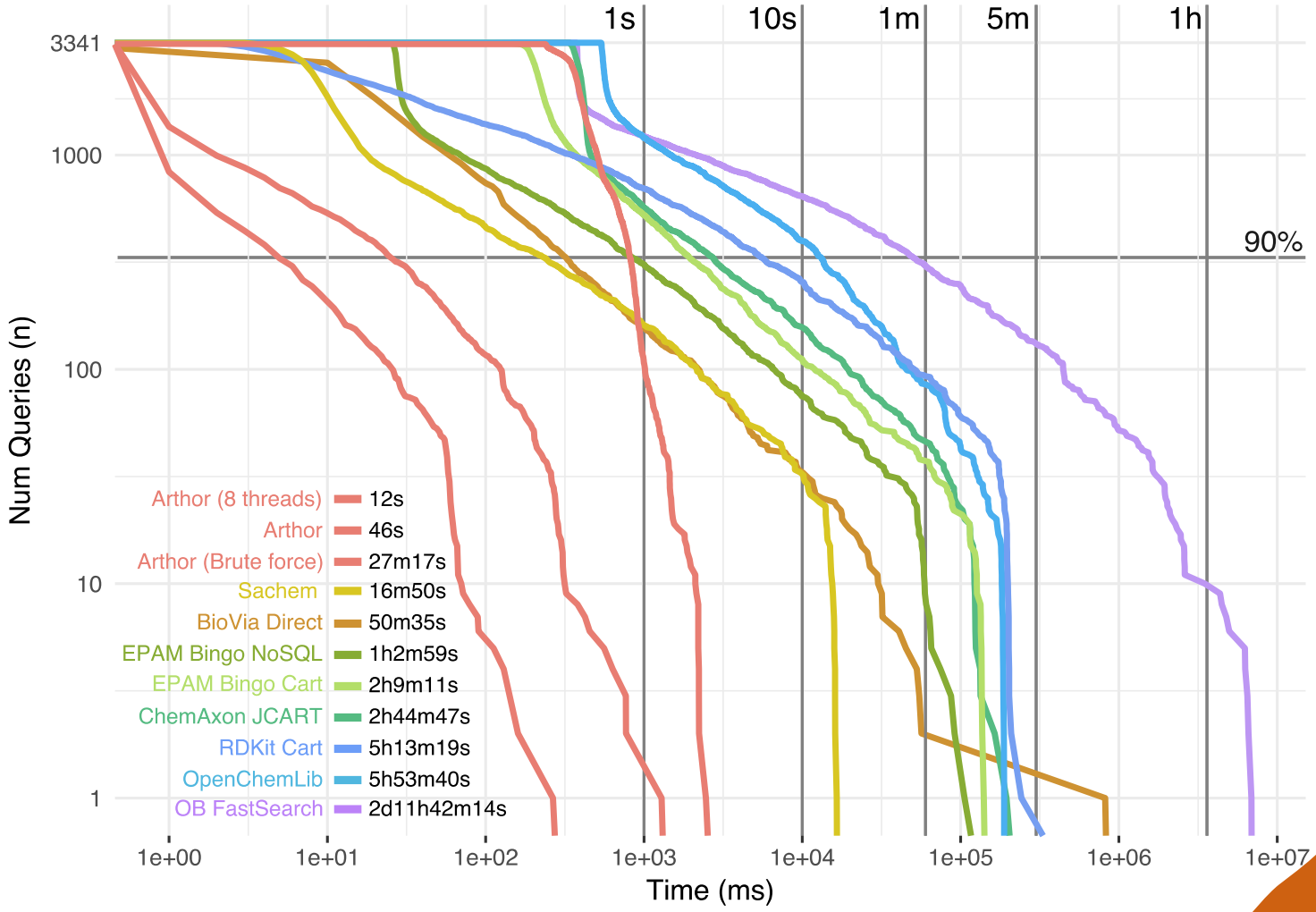


- The primary insight of NextMove Software's Arthor is to use a compact binary representation that is ready to search.
- The same bytes on disk as when processed in memory.

BANDWIDTH IN SMARTS SEARCHING 2

- Another major factor is efficient encoding (compression) of connection tables in a form that can be processed.
- Arthor requires 92 bytes on average per molecule.
  - The majority of molecules require only 2 bytes per atom and 2 bytes per bond.

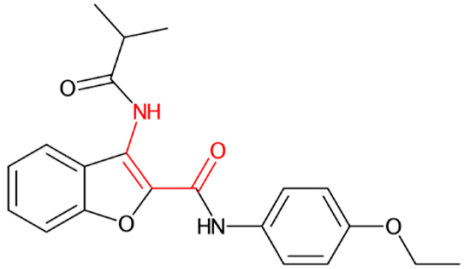
SQC SUBSTRUCTURE BENCHMARKS



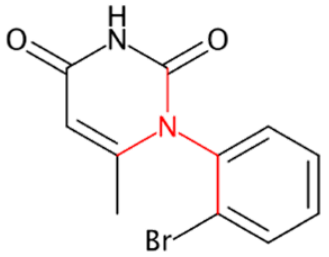
https://www.slideshare.net/NextMoveSoftware/substructure-search-faceoff

ARTHOR MATCHING OPTIMIZATIONS

- Efforts focused on addressing “worst case” pathological behaviour rather then FP prescreens.
- Advanced symmetry perception and pruning.
- SMARTS (& sketch) query optimization.
  - **Intramolecular H Bonds**
    - Input: O=[C,N]aa[N,O;!H0]
    - Internal: [O&i]=[C,N;D>1v>2i]-[aD3]:[aD3]-[N,O;!H0!D0]

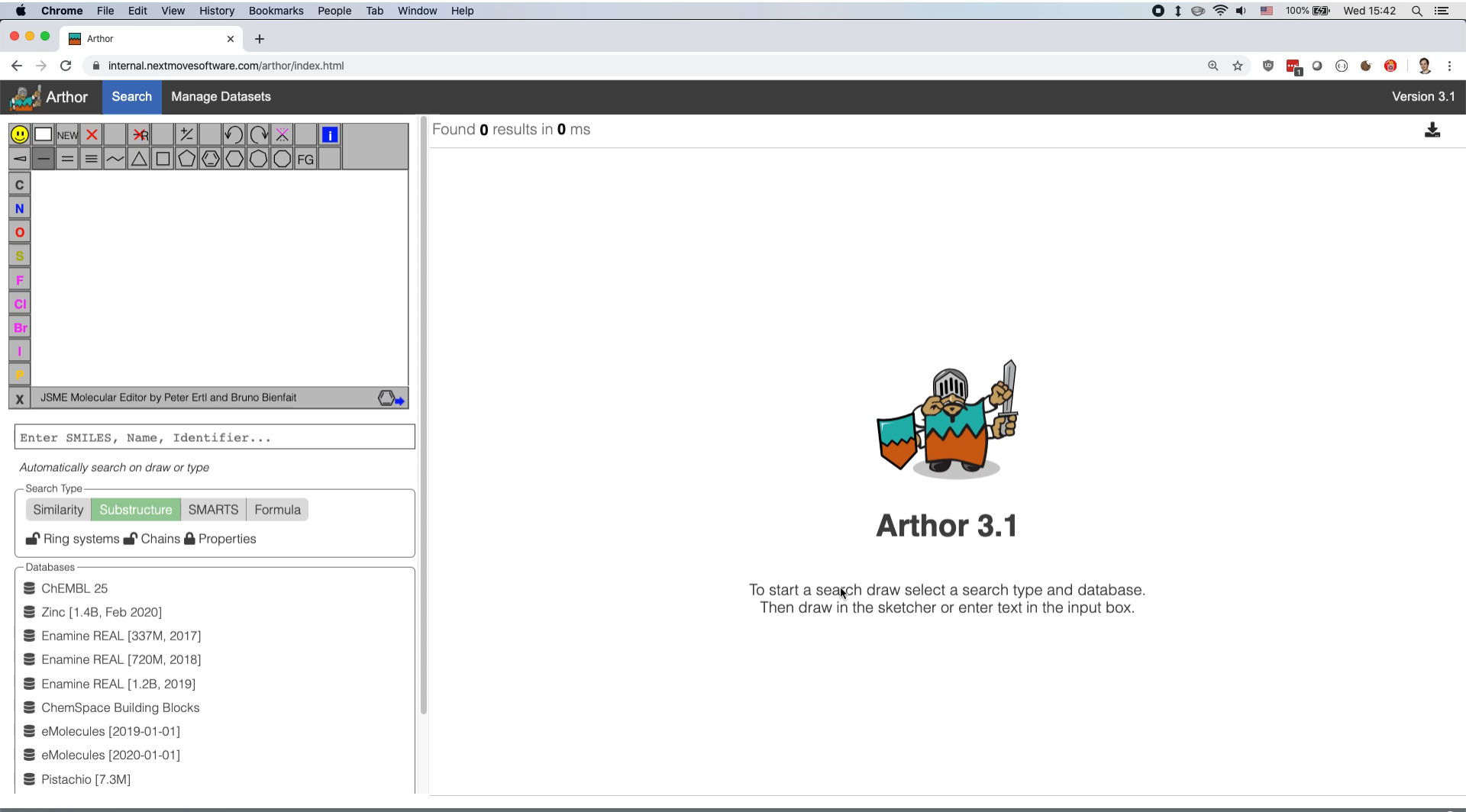


  - **Biaryl atropisomers**
    - Input: [aD3]a!@-a([aD3])[aD3]
    - Internal: [aD3]:[aD3x2]!@-[aD3x2](:[aD3]):[aD3]



  - **Cubane**
    - Input: C12C3C4C1C5C4C3C25
    - Internal: [CZ4Z6Z8x>2]12-@[CZ4Z6Z8x>2]3-@[CZ4Z6Z8x>2]4-@[CZ4Z6Z8x>2]-@1-@[CZ4Z6Z8x>2]5-@[CZ4Z6Z8x>2]-@4-@[CZ4Z6Z8x>2]-@3-@[CZ4Z6Z8x>2]-@2-@5

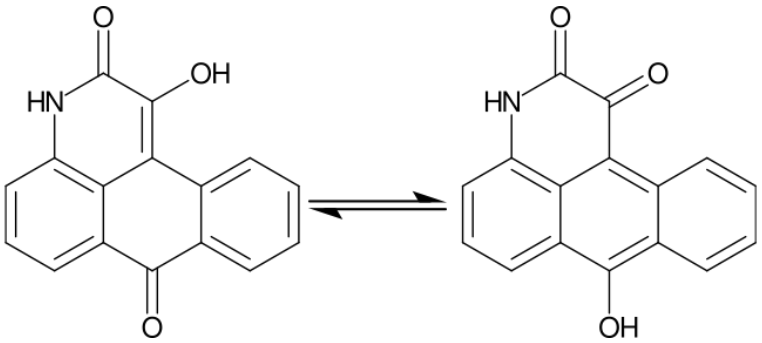
SUBSTRUCTURE VIDEO



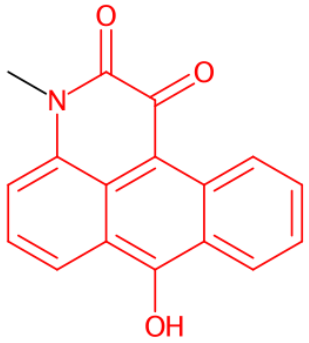
https://www.youtube.com/watch?v=NmmES_mNF9w

THE PROBLEM WITH TAUTOMERS

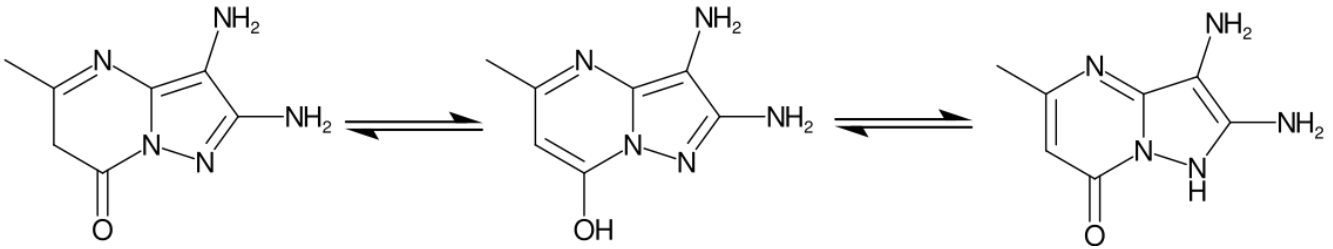
- Tautomerism causes problems for SMARTS matching.



quinone_C filter



CID 5409668

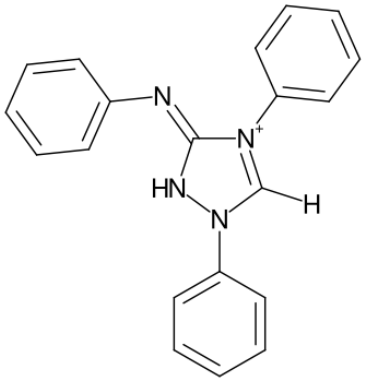


het_65_G filter

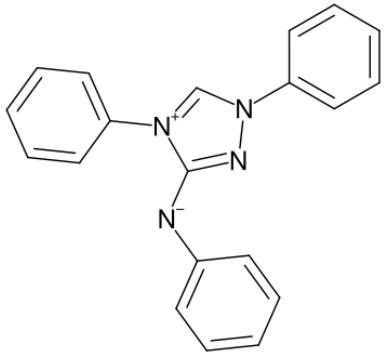
CID 4060544

AND WITH RESONANCE FORMS

- As do resonance forms (and nitro representation &.)

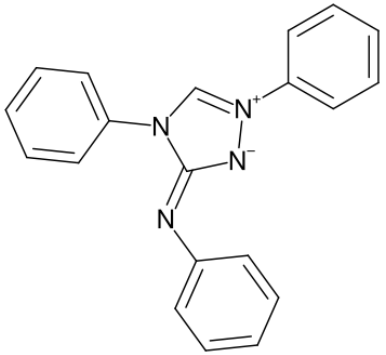


het_5_inium filter



CID 720071

(Oct 2018)

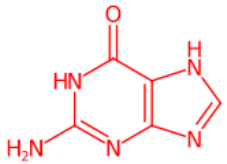
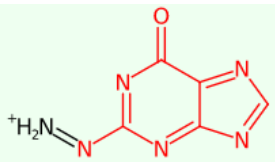
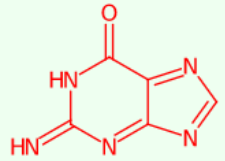
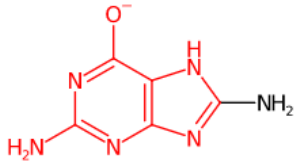
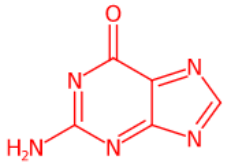
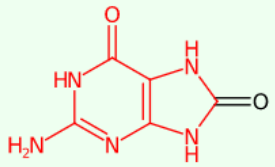
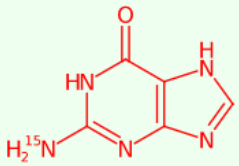
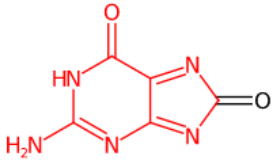
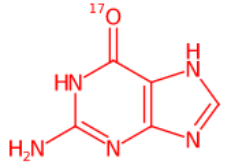
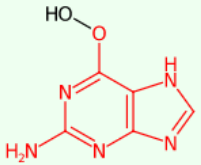


CID 720071

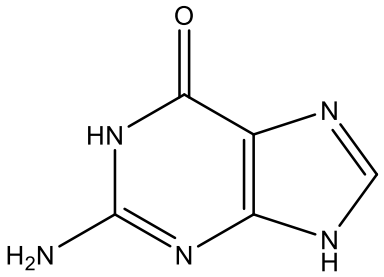
(Feb 2019)

GUANINE TAUTOMERS RESULTS

- Without post-processing, the raw results look like:

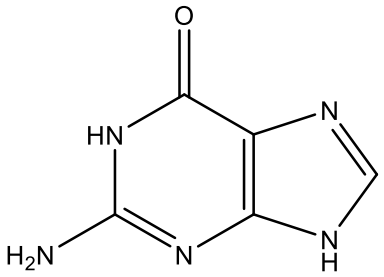
1		**135398634** $C_5H_5N_5O$ 151.126	54		**101930028** C_5HN_6O 161.101
2		**140355140** $C_5H_3N_5O$ 149.110	55		**102341194** $C_5H_5N_6O$ 165.133
3		**504258** $C_5H_3N_5O$ 149.110	56		**135420630** $C_5H_5N_5O_2$ 167.125
4		**135537618** $C_5H_3N_5O$ 150.104	57		**135498233** $C_5H_3N_5O_2$ 165.109
5		**136689991** $C_5H_5N_5O$ 152.126	58		**141357953** $C_5H_5N_5O_2$ 167.125

GUANINE TAUTOMERS SMARTS



- Carbons are constrained to be sp2, heteroatoms to standard valences.
- (Non-triple) bonds become single, double or aromatic.
- [#7]1-,=,:[#6]-,=,:[#7]-,=,:[#6]2-,=.:[#6]-,=,:1-,=,:[#7]-,=,:[#6](-,=,:[#7]) -,=,:[#7]-,=,:[#6] -,=,:2-,=,:[#8]
- Carbons: [#6X3v4i+0,#6X3v3-1,#6X3v3+1]
- Nitrogens: [#7X2v3+0,#7X3v3+0,#7X3v4+1,#7X4v4+1]
- Oxygens: [#8X1v2+0,#8X1v1-1,#8X2v2+0,#8X2v3+1,#8X3v3+1]

GUANINE TAUTOMERS SMARTS



- [#7X2v3+0,#7X3v3+0,#7X3v4+1,#7X4v4+1]1-,=,: [#6X3v4i+0]-,=,: [#7X2v3+0,#7X3v3+0,#7X3v4+1,#7X4v4+1] -,=,: [#6X3v4i+0]2-,=,:[#6X3v4i+0]-,=,:1-,=,: [#7X2v3+0,#7X3v3+0,#7X3v4+1,#7X4v4+1]-,=,: [#6X3v4i+0](-,=,: [#7X2v3+0,#7X3v3+0,#7X3v4+1,#7X4v4+1]) -,=,: [#7X2v3+0,#7X3v3+0,#7X3v4+1,#7X4v4+1]-,=,: [#6X3v4i+0] -,=,:2-,=,: [#8X1v2+0,#8X1v1-1,#8X2v2+0,#8X2v3+1,#8X3v3+1]

NETQ & TOTX: CONSERVATION OF X

- Identifying/filtering relevant tautomers

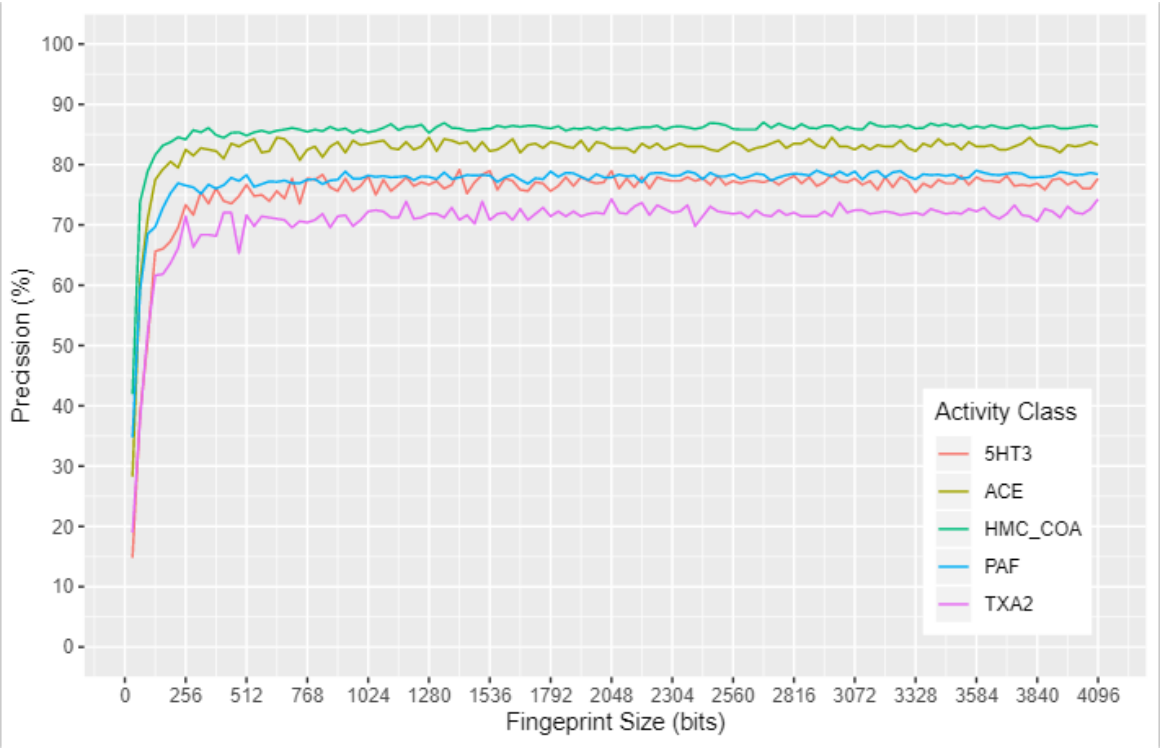
```
int netq = 0;
unsigned int totx = 0;
for (unsigned int idx : match) {
  Atom *atm = mol->getAtomWithIdx(idx);
  netq += atm->getFormalCharge();
  totx += atm->getTotalDegree();
}
// Gaunine tautomers have totx of 29, and 0 netq.
return totx+netq >= 29;
```

- Under reasonable valence constraints $\Sigma(Xi + qi)$ is a conserved quantity of substituted tautomers/protomers.

BANDWIDTH IN FP SIMILARITY

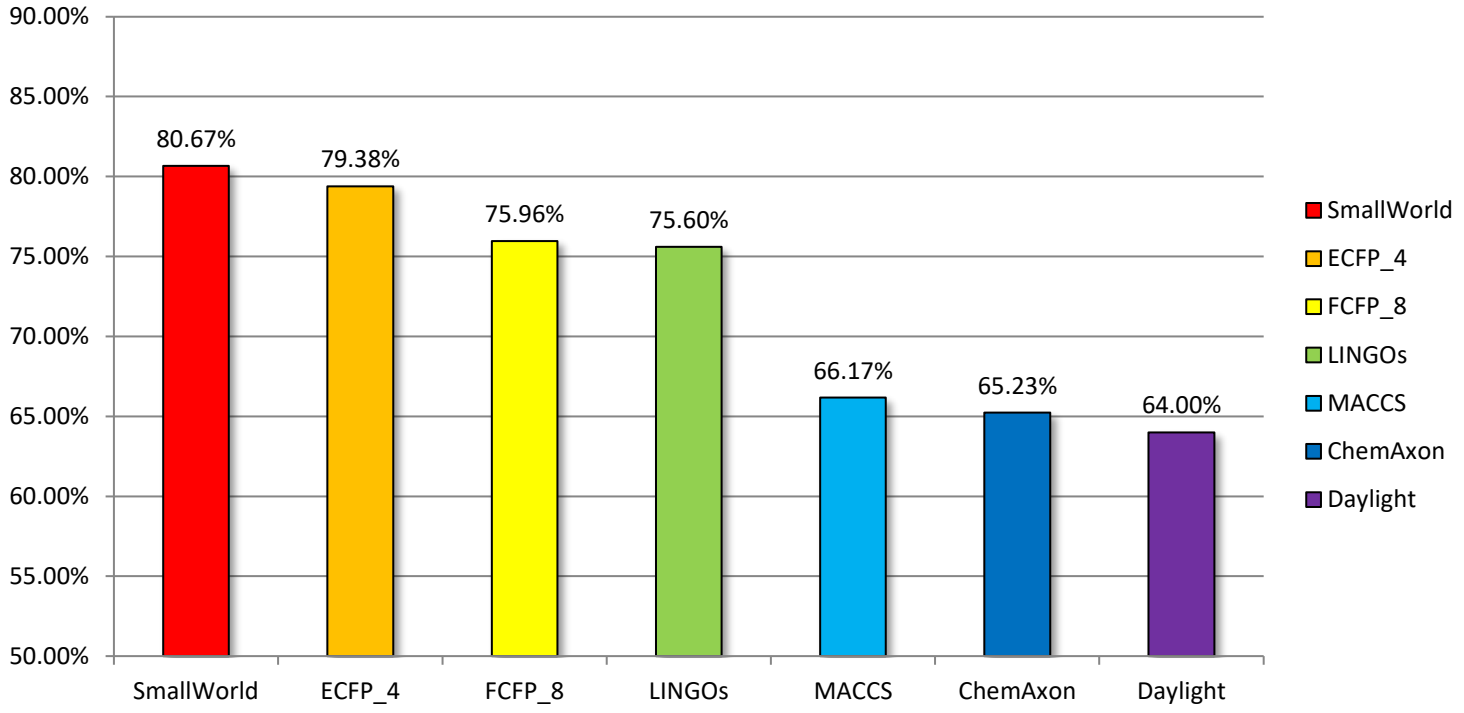
- Bandwidth limitations also apply computing similarity using Tanimoto of binary fingerprints.
- Fingerprints can be stored without storing the popcount of each fingerprint explicitly.
- The biggest (tuneable) trade-off is the length of each fingerprint.

INFLUENCE OF ECFP4 FINGERPRINT LEN



- For ultra-large databases 256 bits are sufficient, for regular (ChEMBL/PubChem) 1024 is reasonable.

ECFP4 FINGERPRINTS IN CONTEXT



- ECFP4 fingerprints (even at 256 bits) outperform the path-based fingerprints in prev-gen search engines.

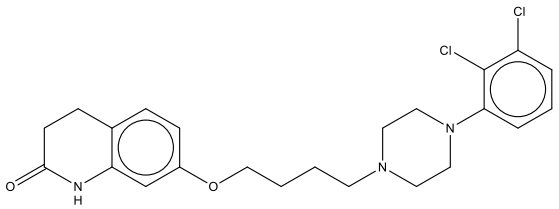
THE IMPACT OF FINGERPRINT CHOICE

- A side-effect of progress is that ECFP4 has a significantly lower bit density than path-based FPs.
- One observable is that Tanimoto values above 0.345 are now biologically significant where path-based fingerprints (traditionally) have a 0.7 threshold.
- Less obvious is that Baldi-bounds based optimizations are less effective.
- Fortunately, simpler multi-threading and other benefits more than make up for bounds pruning.
- Search that looks at over 50% of a database is O(N).

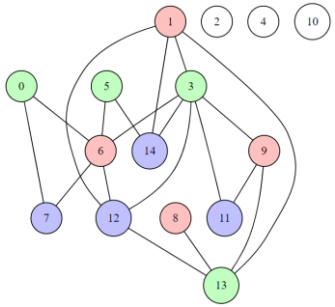
BREAKING THE SPEED BARRIER

- The application of Just-In-Time (JIT) compilation, the process of turning similarity queries into executable machine code enables a significant bandwidth breakthrough.
- Not every word of a fingerprint is necessary for calculating a Tanimoto, by reading only relevant words memory bandwidth is used more effectively.

JIT CODE GENERATION EXAMPLE



0: 4000400000101110
1: 0002000000082000
2: 0000000001000000
3: 0010001010046000
4: 0000000000001000
5: 9000002001000080
6: 1000000810000940
7: 0000000004800900
8: 0000100400000008
9: 0040804000040400
10: 0000000010000000
11: 0020000000240020
12: 00000000500a0000
13: 0000800400080002
14: 0000000001002004




```
c = __popcll((target[0] & 0x4000400000101110) +
             (target[3] & 0x0010001010046000) +
             (target[5] & 0x9000002001000080) +
            (target[13]& 0x0000800400080002))
  + __popcll((target[1] & 0x0002000000082000) +
             (target[6] & 0x1000000810000940) +
             (target[8] & 0x0000100400000008) +
             (target[9] & 0x0040804000040400))
  + __popcll((target[7] & 0x0000000004800900) +
             (target[11]& 0x0020000000240020) +
             (target[12]& 0x00000000500a0000) +
             (target[14]& 0x0000000001002004))
  + ((target[2]>>24)&1)
  + ((target[4]>>12)&1)
  + ((target[10]>>28)&1);
```

ARTHOR ATFP JIT BACKENDS

- NVidia PTX Assembly Language

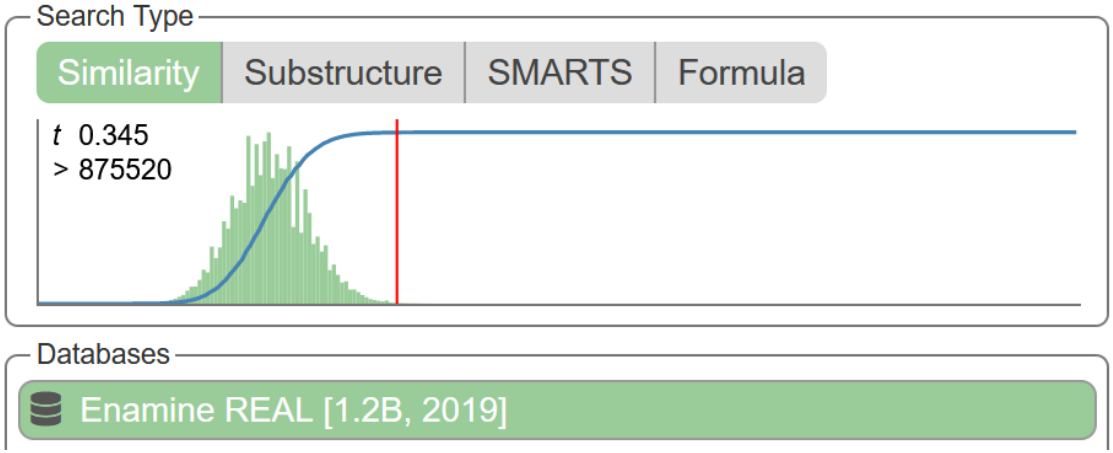
```
ld.global.u64   %rd27, [%rd10+56];
and.b64         %rd28, %rd27, 4947953319952;
popc.b64        %r19, %rd28;
add.s32         %r20, %r18, %r19;
ld.global.u32   %r21, [%rd10+68];
shr.u32         %r22, %r21, 3;
and.b32         %r23, %r22, 1;
add.s32         %r24, %r20, %r23;
```

- ARM v6 Assembly Language

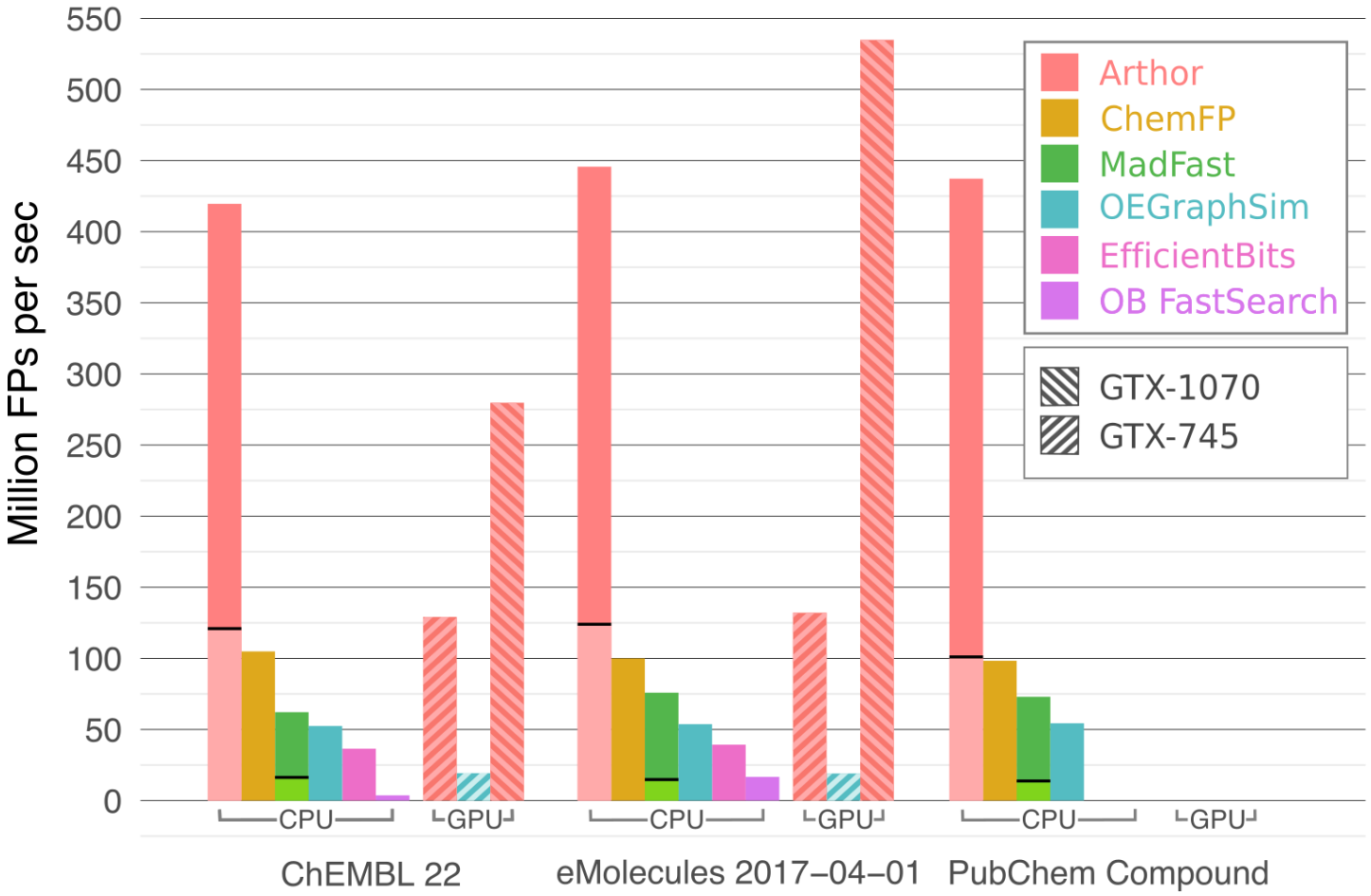
```
ldrd    r0, [fp]
lsl     r0, r0, #11
lsr     r7, r0, #31
lsl     r0, r0, #8
add     r7, r7, r0, lsr #31
lsl     r0, r0, #4
add     r7, r7, r0, lsr #31
lsl     r1, r1, #1
add     r7, r7, r1, lsr #31
lsl     r1, r1, #16
add     r7, r7, r1, lsr #31
```

TWO PASS COUNTING SORT

- Analysis of prev-gen FP search systems revealed that typically sorting, not searching, is the bottleneck.
- The search phase is O(*N*), but sorting the results is typically O(*N*.log*N*) for non-trivial numbers of hits.
- Arthor uses an efficient O(*N*) two-pass counting sort.



COMPARISON TO PREVIOUS WORK



Split bars indicate single thread vs. 16 threads for CPU.

FP SIMILARITY VIDEO

Chrome File Edit View History Bookmarks People Tab Window Help — 100% — Wed 15:43

Arthor

internal.nextmovesoftware.com/arthor/index.html

Arthor | Search | Manage Datasets | Version 3.1

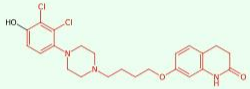
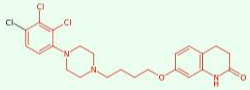
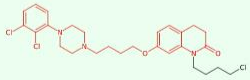
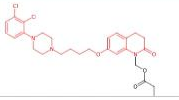
abilify

Automatically search on draw or type

Search Type: Similarity | Substructure | SMARTS | Formula

Databases:
- ChEMBL 25
- Zinc [1.4B, Feb 2020]
- Enamine REAL [337M, 2017]
- Enamine REAL [720M, 2018]
- Enamine REAL [1.2B, 2019]
- ChemSpace Building Blocks
- eMolecules [2019-01-01]

Matched **10** results in **60** ms

1		**ZINC000001851149** $C_{23}H_{27}Cl_2N_3O_2$ 448.384
2		**ZINC000095669780** $C_{23}H_{27}Cl_2N_3O_3$ 464.383
3		**ZINC000027722180** $C_{23}H_{26}Cl_3N_3O_2$ 482.829
4		**ZINC000027731699** $C_{23}H_{26}Cl_3N_3O_2$ 482.829
5		**ZINC000115098832** $C_{24}H_{29}Cl_2N_3O_3$ 478.410
6		**ZINC000100222367** $C_{27}H_{34}Cl_3N_3O_2$ 538.935
7		**ZINC000095564798** $C_{30}H_{39}Cl_2N_3O_4$ 576.553

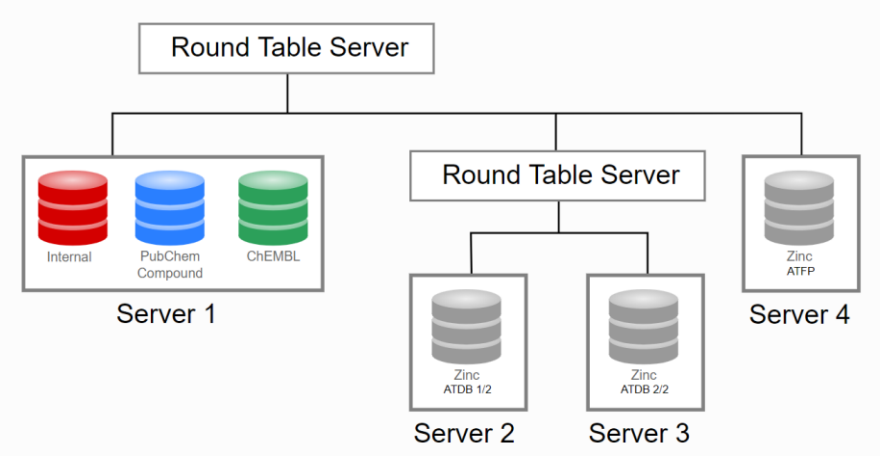
Page Request Time = 61 ms

https://www.youtube.com/watch?v=NmmES_mNF9w

ROUND-TABLE: FEDERATED SEARCH

Movie quote: “First rule in government spending: why build one when you can have two at twice the price?” - S.R. Hadden, Contact (1997).

- Round-table allows Arthor databases to be distributed and split across multiple hosts (on a cloud or local area network) and virtualized as a single DB.



- Round table eliminates Arthor’s 4B row limit.

THE BIGGER PICTURE

- Search is just one aspect of ultra-large chemical database management.
  - Data loading/indexing/fingerprinting is a critical step.
  - Graph canonicalization for deduplication/hashing.
  - molhash for lookup, tautomer, formula, scaffold search.
  - Join performance with external databases rate limiting.
  - Calculation of MW, LogP, Rotb, HBA, HBD, Ro5, PAINS.
  - Pre-sorting SMILES (by HAC/MW) probably bottleneck.
  - Clustering and Diversity selection*

* https://github.com/rdkit/UGM_2019/blob/master/Presentations/Sayle_Clustering.pdf

I FEEL THE NEED, THE NEED FOR SPEED

- An important ingredient is canonical SMILES.
  - RDKit 2019 6.82 Kmol/s
  - InChI (Open Babel) 7.32 Kmol/s
  - Open Babel 10.3 Kmol/s
  - OpenEye OEChem 50 Kmol/s
  - NextMove Software 113 Kmol/s
- PAINS filtering.
  - Arthor takes 32s to check PubChem's 100M compounds.

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.

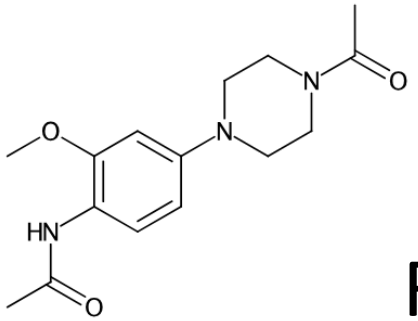
INFINITE PEPTIDE SIMILARITY SEARCH

- Blastp results (BLOSUM62) to 14AA sequence Mastoparan-L

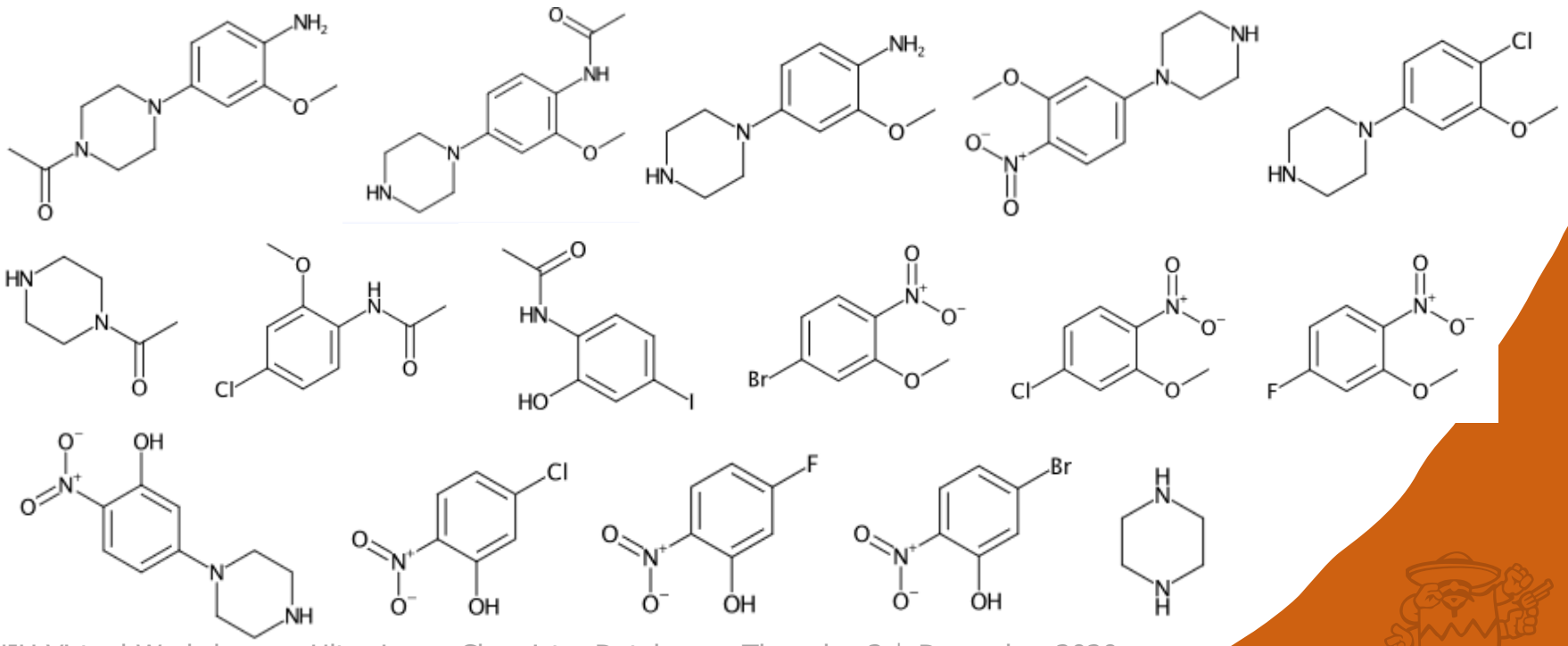
```
Query: INLKALAALAKKIL  Dist=0
Hit 1: VNLKALAALAKKIL  Dist=1
Hit 2: INLKALAALAKKVL  Dist=1
Hit 3: VNLKALAALAKKVL  Dist=2
Hit 4: LNLKALAALAKKIL  Dist=2
Hit 5: INLKALAALAKKLL  Dist=2
Hit 6: INIKALAALAKKIL  Dist=2
Hit 7: INMKALAALAKKIL  Dist=2
Hit 8: INLKAIAALAKKIL  Dist=2
Hit 9: INLKAMAALAKKIL  Dist=2
Hit 10:INLKALAAIAKKIL  Dist=2
Hit 11:INLKALAAMAKKIL  Dist=2
Hit 12:INLKALAALAKKII  Dist=2
Hit 13:INLKALAALAKKIM  Dist=2
Hit 14:MNLKALAALAKKIL  Dist=3
```

RETROSYNTH BUILDING BLOCK SEARCH

- Query:



Results:

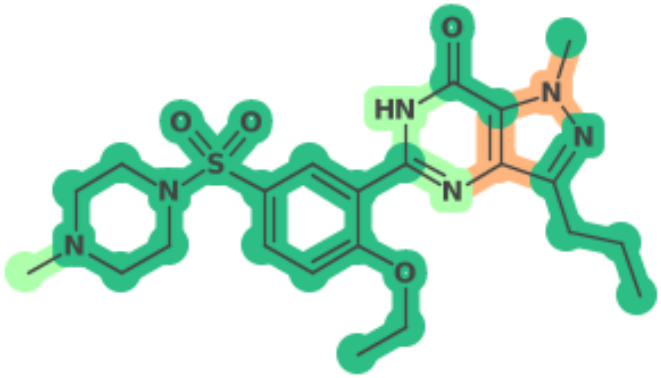


FIGHTING BIG DATA WITH BIGGER DATA

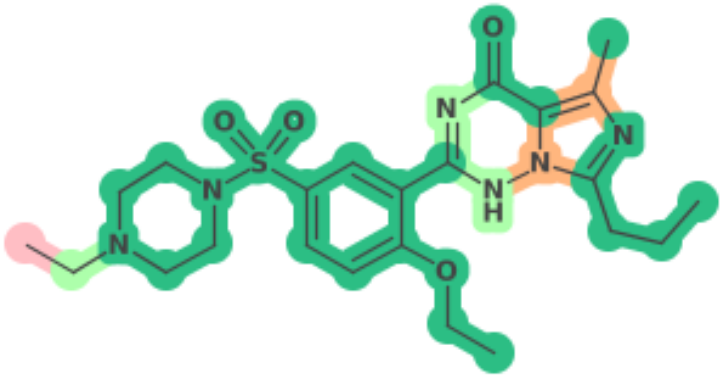
- Observe that with pre-indexing (sorting), lookup & duplicate removal go from O(N) to O(log N).
- Similarly, tautomer, Bemis-Murko scaffold and matched pair search can be sped up by pre-indexing.
- SmallWorld pre-indexes subgraphs to simplify Maximum Common Substructure from NP to ~O(1).
- Graph-edit distance searches (of any size database) now only requires ~10Mbytes of bandwidth.
- The catch: the spatial data structure requires >40TB.
- The good news is this is only computed once.

EXAMPLE EDIT OPERATIONS

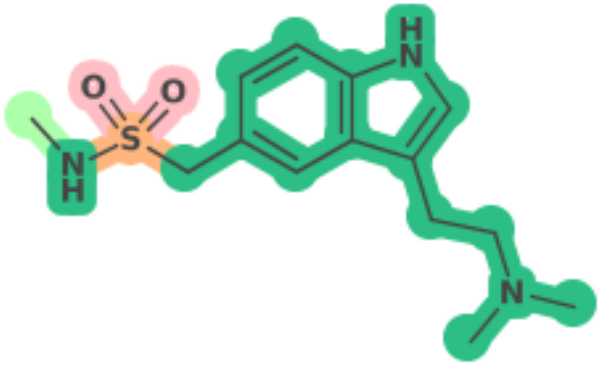
Sildenafil (Viagra)



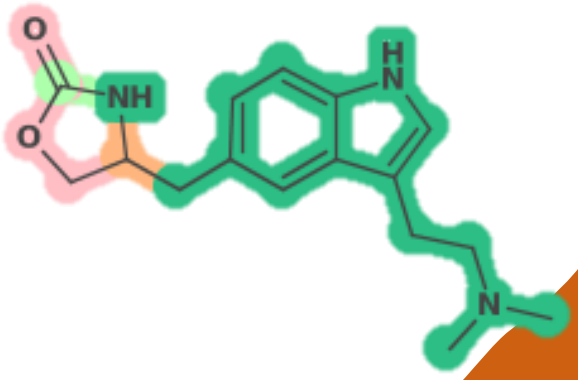
Vardenafil (Levitra)



Sumatriptan (Imitrex)



Zolmitriptan (Zomig)



EDIT DISTANCE

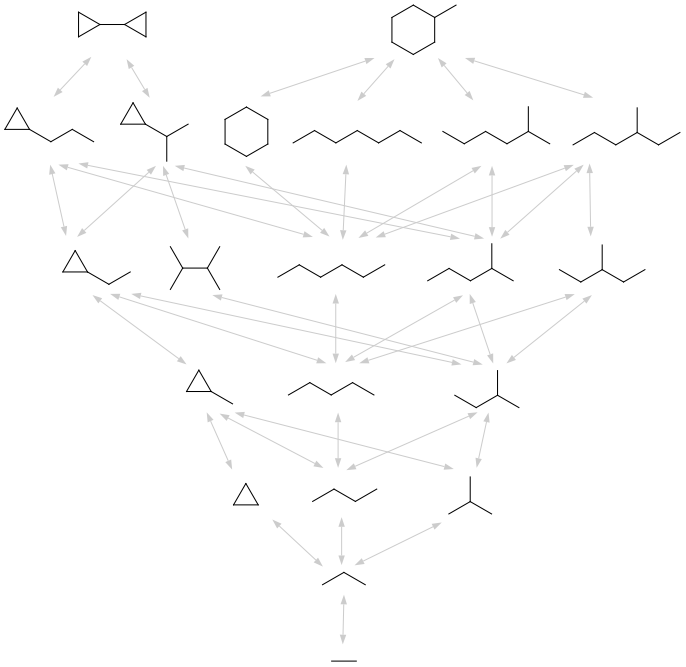
- **Edit Distance** is a measure of similarity (dissimilarity) between two discrete mathematical objects (formally a metric space).
  - String Edit Distance is a similarity metric between strings.
  - Tree Edit Distance is a similarity metric between trees.
  - **Graph Edit Distance** (GED) is a similarity 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).
- Traditionally, 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

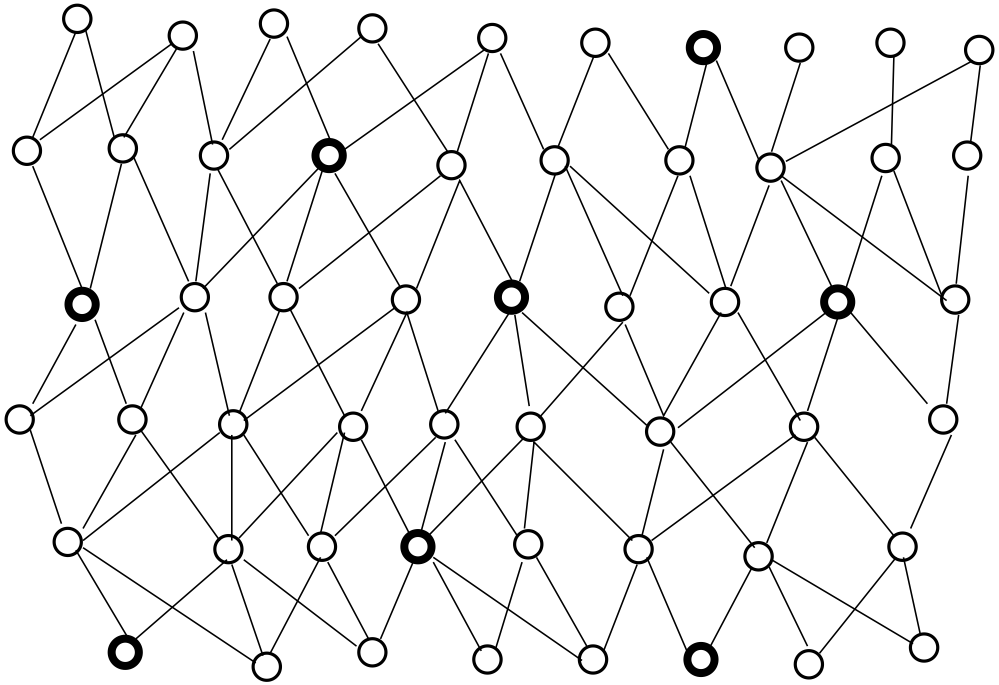
A MAP OF CHEMICAL (GRAPH) SPACE

- 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.



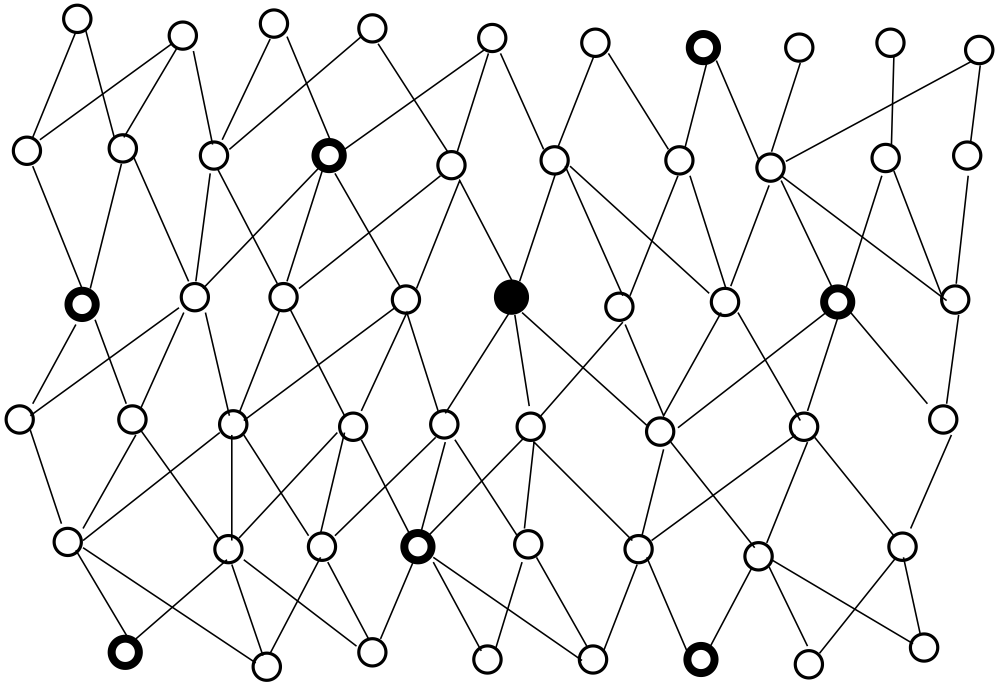
- Currently contains 380 billion vertices and 2.8 trillion edges.

SMALLWORLD SEARCH



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

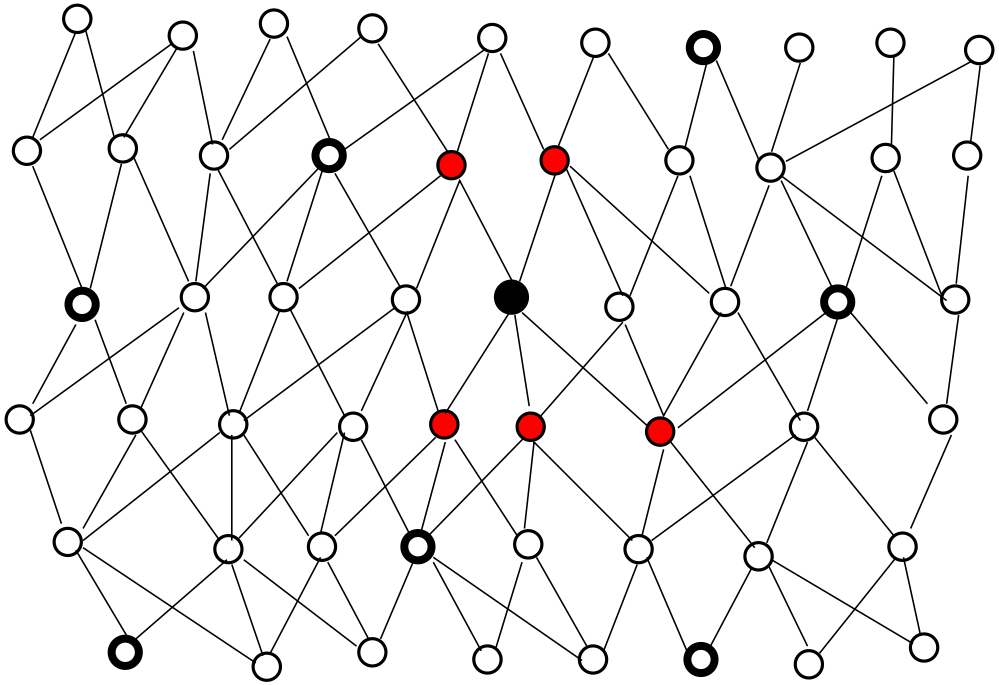
SMALLWORLD SEARCH



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.

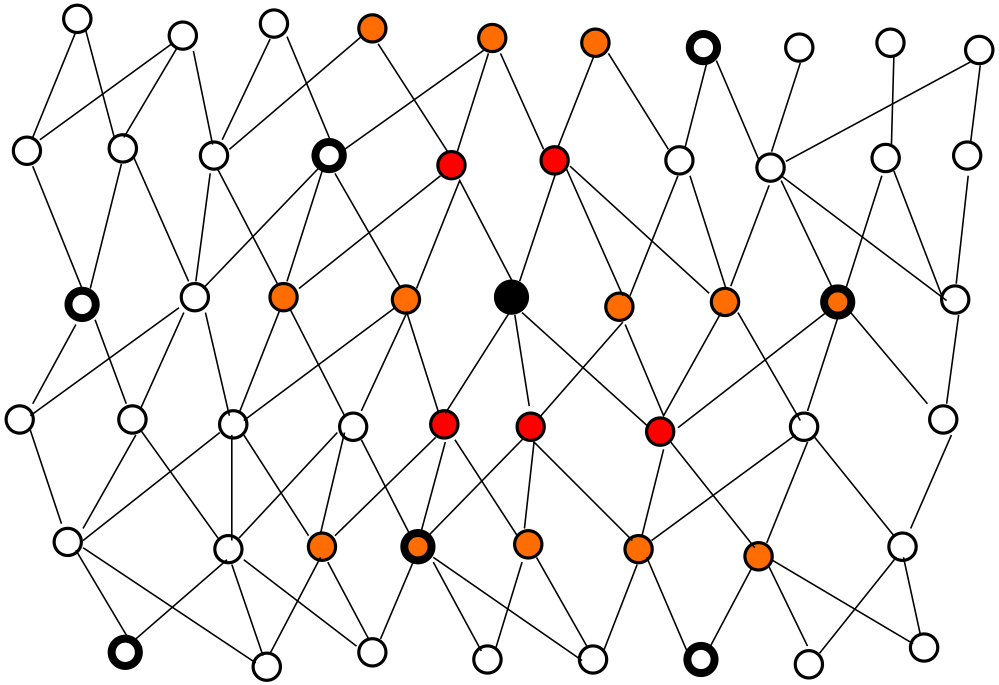
SMALLWORLD SEARCH



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”.

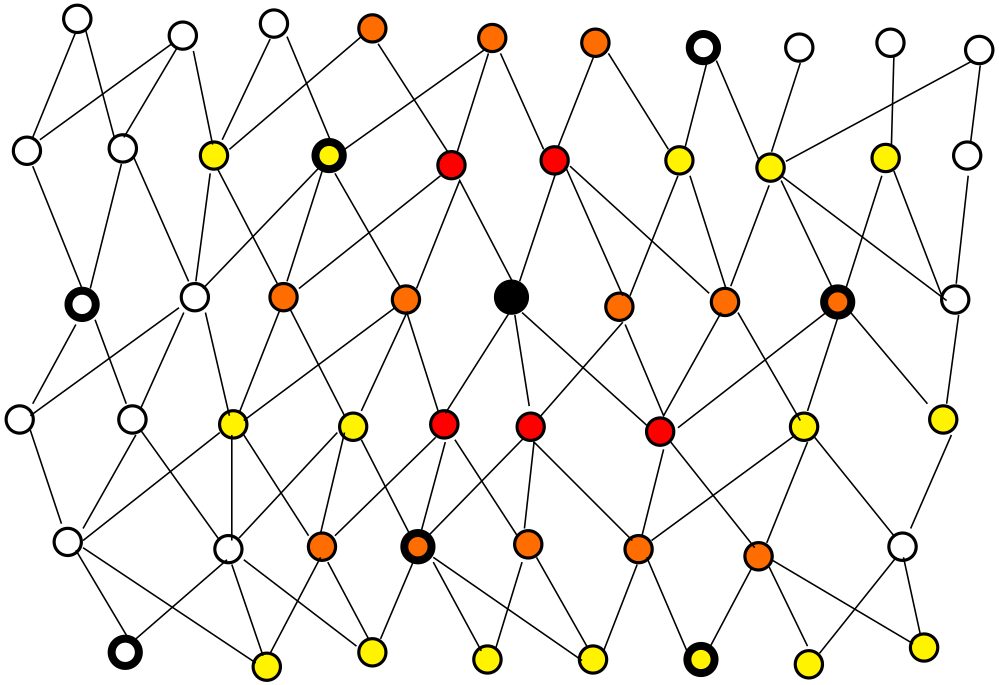
SMALLWORLD SEARCH



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.

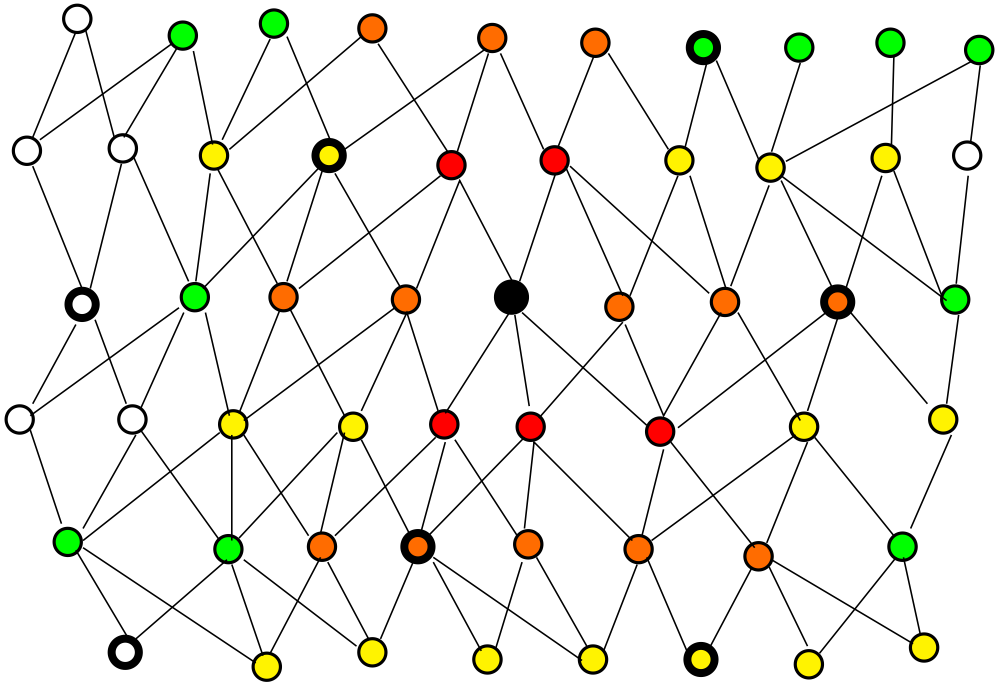
SMALLWORLD SEARCH



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.

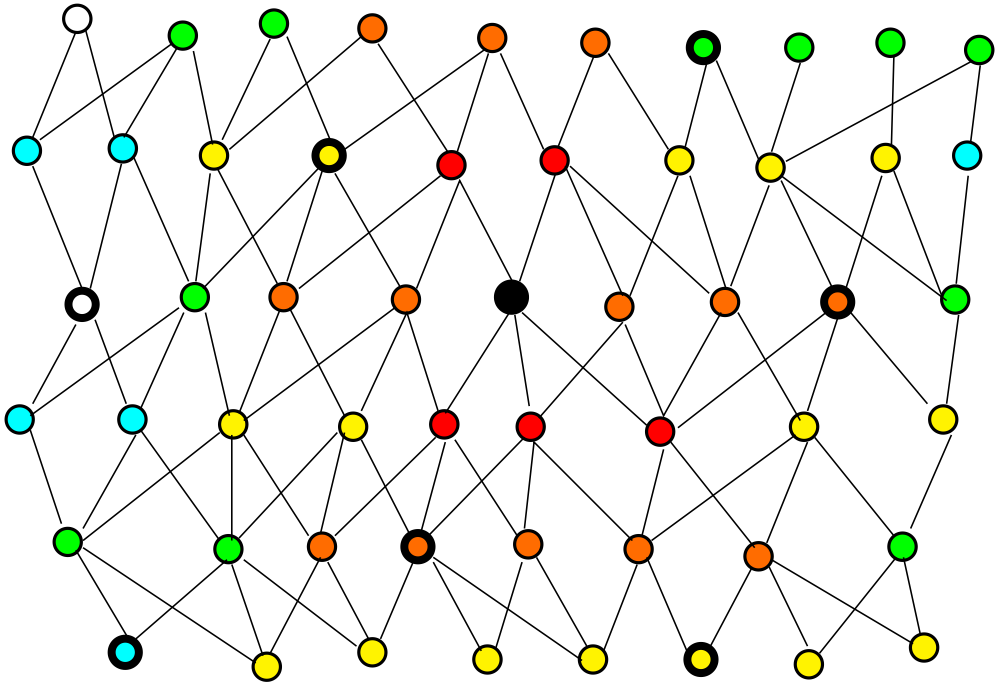
SMALLWORLD SEARCH



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

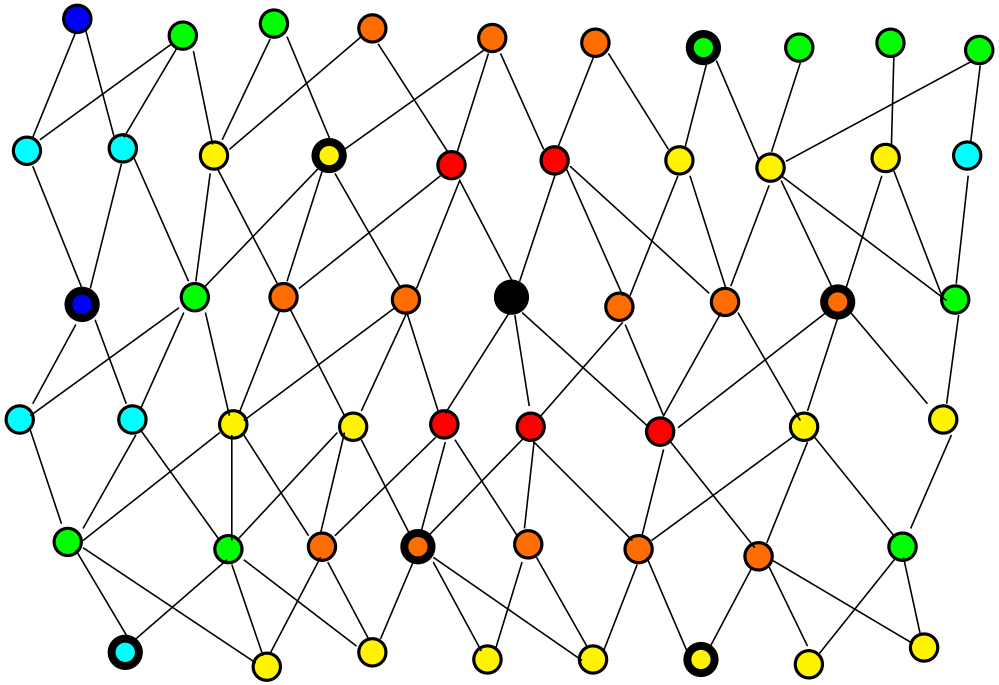
The search terminates once sufficient mapped neighbours have been found (or a suitable iteration limit is reached).

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

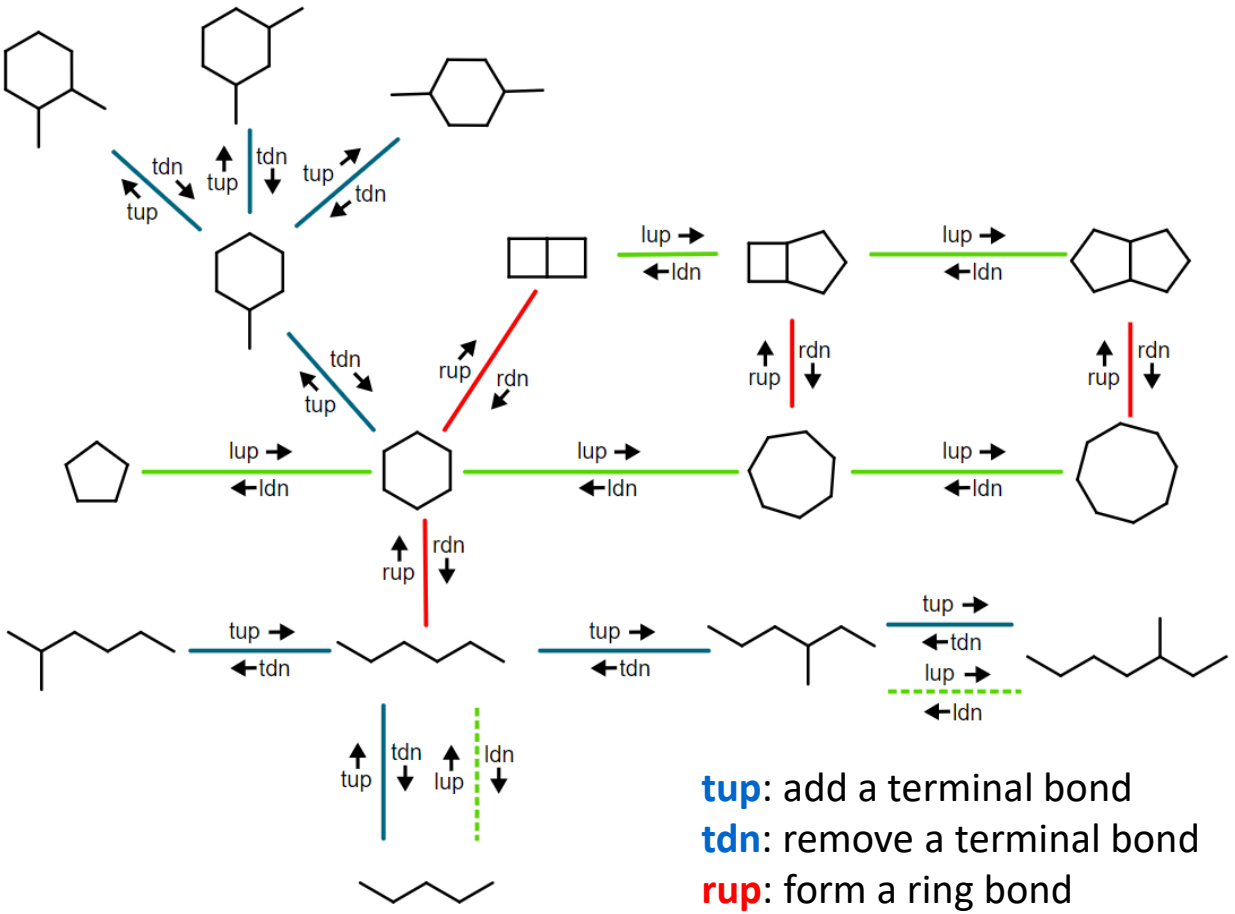


Dist	WF	New	Hits
0	1	1	1
1	5	0	1
2	13	2	3
3	16	2	5
4	11	1	6
5	6	1	7
6	2	1	8

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

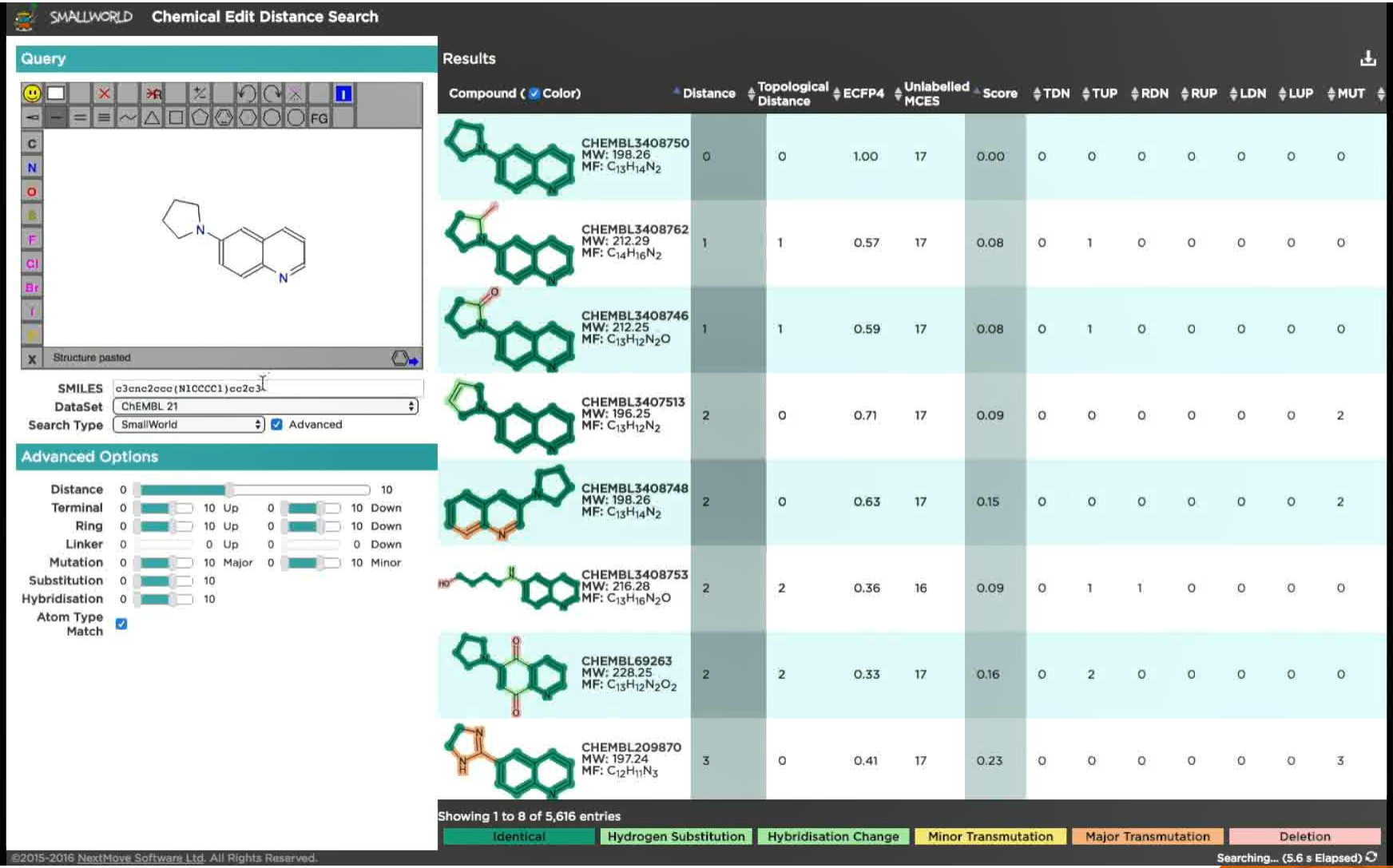
https://graph500.org/

TOPOLOGICAL EDIT/EDGE TYPES

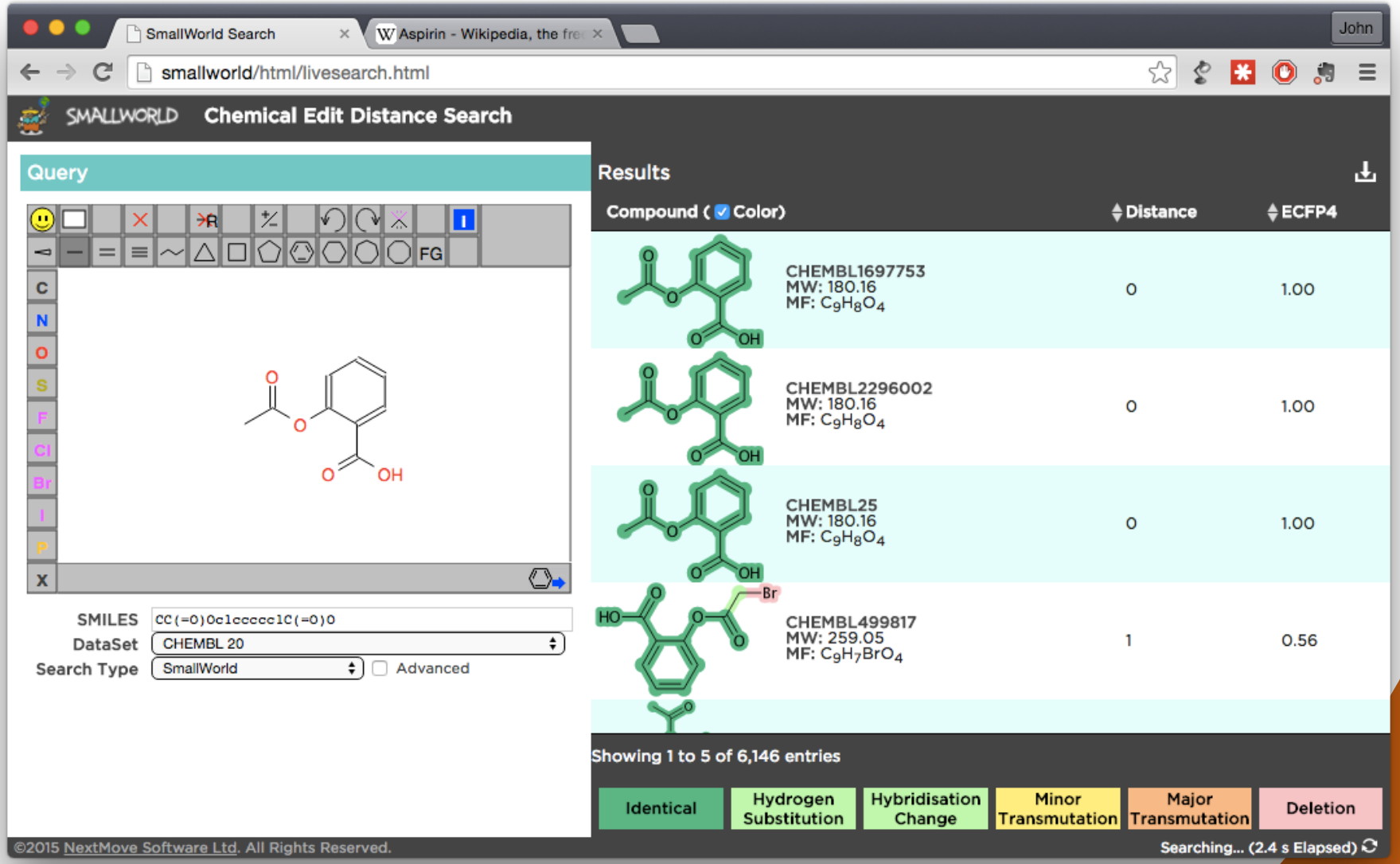


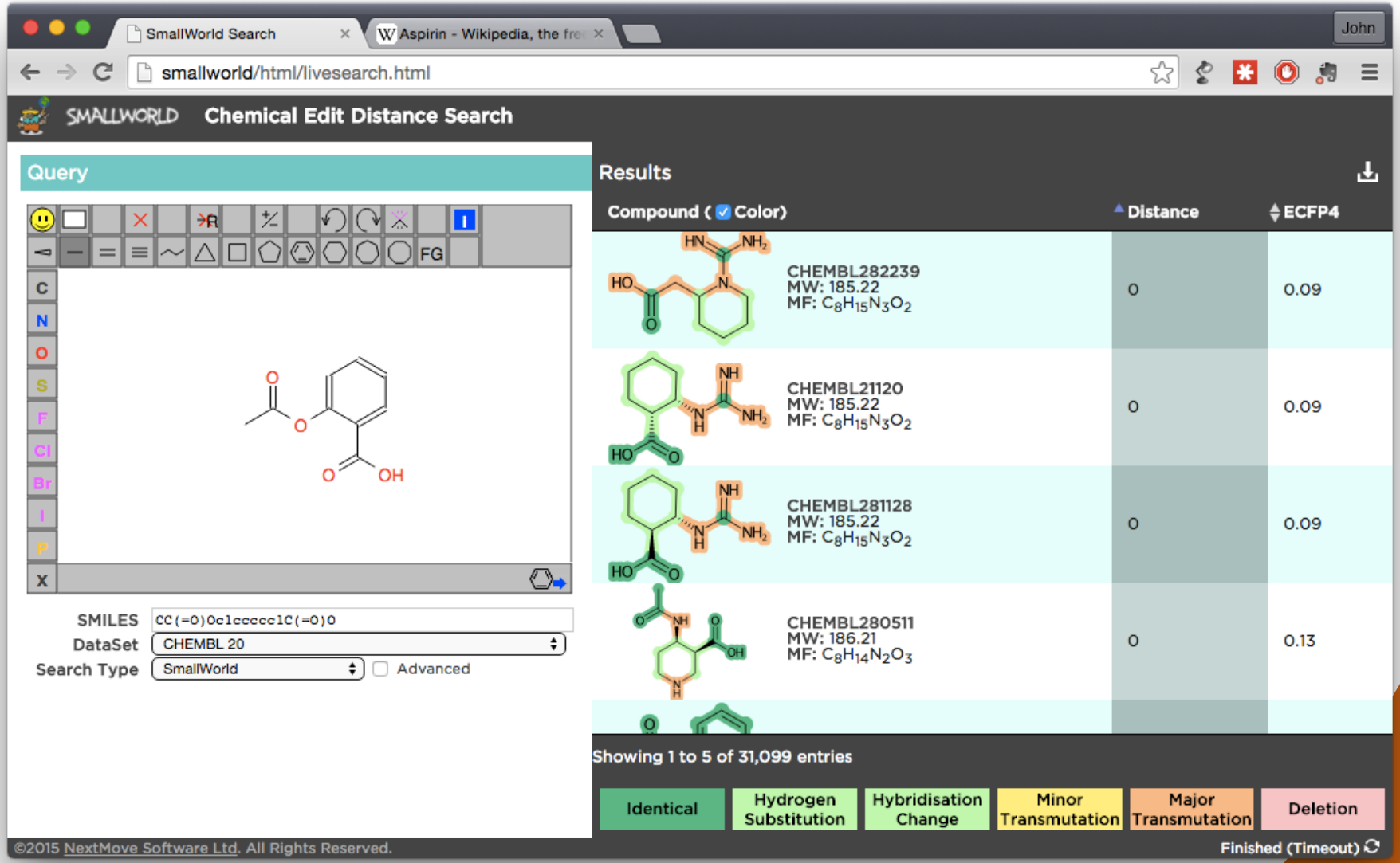
tup: add a terminal bond
tdn: remove a terminal bond
rup: form a ring bond
rdn: break a ring bond
lup: insert a (degree 2) linker node
ldn: remove a (degree 2) linker node

SMALLWORLD



https://www.youtube.com/watch?v=hZ4QyQSeSWg


SmallWorld Search
Aspirin - Wikipedia, the fre
John
smallworld/html/livesearch.html
SMALLWORLD
Chemical Edit Distance Search
Query
SMILES
CC(=O)Oc1ccccc1C(=O)O
DataSet
CHEMBL 20
Search Type
SmallWorld
Advanced
Results
Compound (Color)
Distance
ECFP4
CHEMBL1697753
MW: 180.16
MF: C9H8O4
0
1.00
CHEMBL2296002
MW: 180.16
MF: C9H8O4
0
1.00
CHEMBL25
MW: 180.16
MF: C9H8O4
0
1.00
CHEMBL499817
MW: 259.05
MF: C9H7BrO4
1
0.56
Showing 1 to 5 of 6,146 entries
Identical
Hydrogen Substitution
Hybridisation Change
Minor Transmutation
Major Transmutation
Deletion
©2015 NextMove Software Ltd. All Rights Reserved.
Searching... (2.4 s Elapsed)


SmallWorld Search
Aspirin - Wikipedia, the fre
John
smallworld/html/livesearch.html
SMALLWORLD
Chemical Edit Distance Search
Query
FG
C
N
O
S
F
Cl
Br
I
P
X
SMILES
CC(=O)Oc1ccccc1C(=O)O
DataSet
CHEMBL 20
Search Type
SmallWorld
Advanced
Results
Compound (Color)
Distance
ECFP4
CHEMBL282239
MW: 185.22
MF: C8H15N3O2
0
0.09
CHEMBL21120
MW: 185.22
MF: C8H15N3O2
0
0.09
CHEMBL281128
MW: 185.22
MF: C8H15N3O2
0
0.09
CHEMBL280511
MW: 186.21
MF: C8H14N2O3
0
0.13
Showing 1 to 5 of 31,099 entries
Identical
Hydrogen Substitution
Hybridisation Change
Minor Transmutation
Major Transmutation
Deletion
©2015 NextMove Software Ltd. All Rights Reserved.
Finished (Timeout)

PREVIOUS 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.

GRAPH DATABASE FABRICATION

- The compressed “raw” source representation of SmallWorld is 14TB gzipped; one ASCII line (of two SMILES) for each edge, i.e. 2.8 trillion text lines.
- Hypothetically, these 2.8T triples could be loaded into a database such as Oracle, Virtuoso or Neo4j.
- Instead, we “compile” this graph database down to a 40TB form that is efficiently searched at run-time.
- This 40TB can be gzipped to 20TB for delivery to customers on two £200 external USB disks (like a subscription service).

CHALLENGES FOR 21ST CENTURY

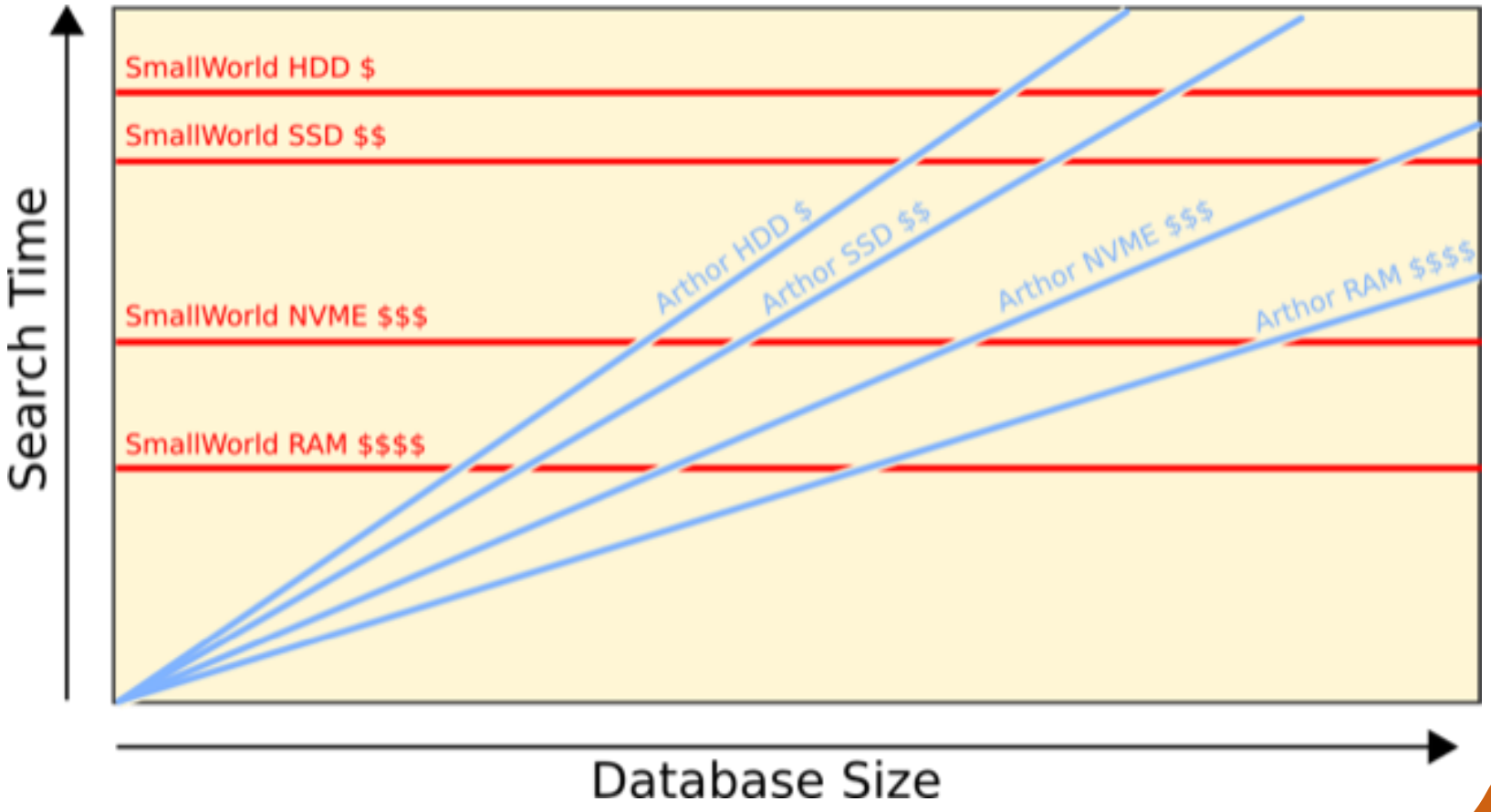
- In SmallWorld, latency not bandwidth (nor compute) becomes the limiting factor.
- Evaluating a proof-of-concept SmallWorld deployment at a pharma company in Boston, on a shared server using standard local network storage, a benchmark searching for the 100 nearest neighbors of 18 drug like molecules against Enamine REAL 2019 (1.3B) took 9 minutes from a cold start. They were impressed with ~30s per search. Re-running the benchmark immediately (a warm start executing exactly the same number of instructions) completed in 11s, around a half second per mol.
- The difference in the times is due to storage latency and OS file caching.
- SmallWorld only requires 10MB of bandwidth, but randomly within ~41TB.
- The company's full deployment uses a 2TB server with attached SSDs.

SUMMARY

- Algorithmic improvements and Moore's law advances in hardware should allow traditional cheminformatics and bioinformatics search techniques to be applied for the time being, but ultimately next generation approaches will be required to handle multi-billion compound databases.

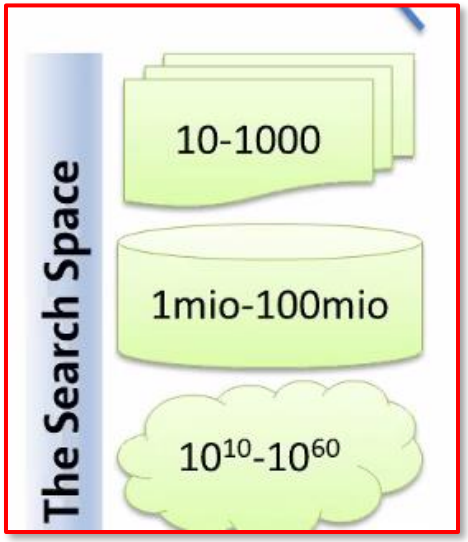
- First master computation, then bandwidth, then latency.

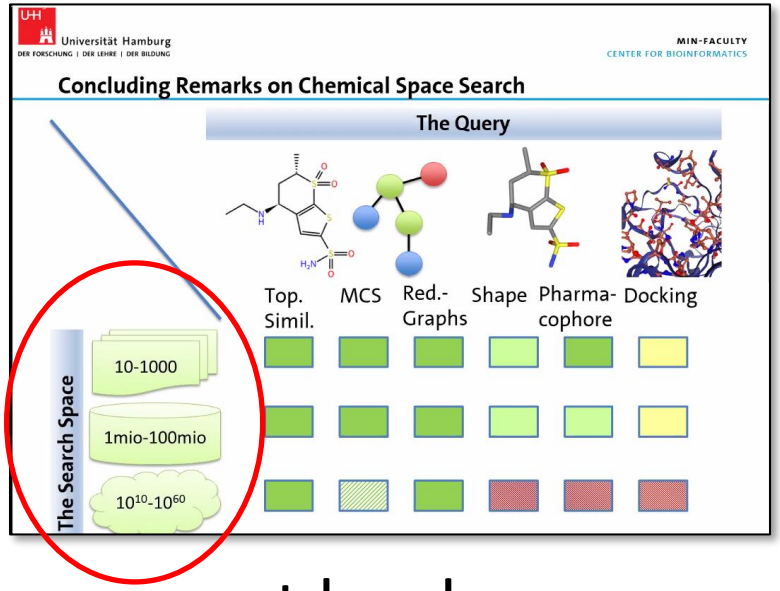
EXECUTIVE SUMMARY



THE END OF THE WORLD IS NIGH?

- From Matthias Rarey's talk yesterday:





- 10^{10}-10^{12} is manageable on current hardware.
- It needs to be, 10^{60} uses 10^{30} building blocks in a two component (proximal) library, or $6x10^{10}$ in 5 steps.

ACKNOWLEDGEMENTS

- Andrew Grant, AstraZeneca.
- John Irwin, ZINC group, UCSF.
- Yurii Moroz, Enamine and ChemSpace.
- Darren Green, GSK.
- Pat Walters, Relay Therapeutics.
- Evan Bolton, PubChem Group, NCBI.
- Marc Nicklaus, NCI.
- Gergely Zahoranszky-Kohalmi, NCATS.
- Noel O'Boyle, Heptares.
- Andrew Dalke, Dalke Scientific Software.
- Christos Nicolaou, Eli Lilly.
- Oliver Korb, Roche.
- Jose Batista, OpenEye Scientific Software.
- Jameed Hussain, Dotmatics.

J. ANDREW GRANT (1963-2012)

Me and Andy at OpenEye EuroCUP 2008