Precomputed 3D Similarity of Small Molecules: PubChem3D!

Evan Bolton, Ph.D.
NCBI/NLM/NIH
CUP X
bolton@ncbi.nlm.nih.gov

PubChem3D Release

http://pubchem.ncbi.nlm.nih.gov/release3d.html

NCBI PubChem

PubChem Compound GO

PubMed | Entrez | Structure | PubChem | Help

PubChem » Help » 3D Release Note

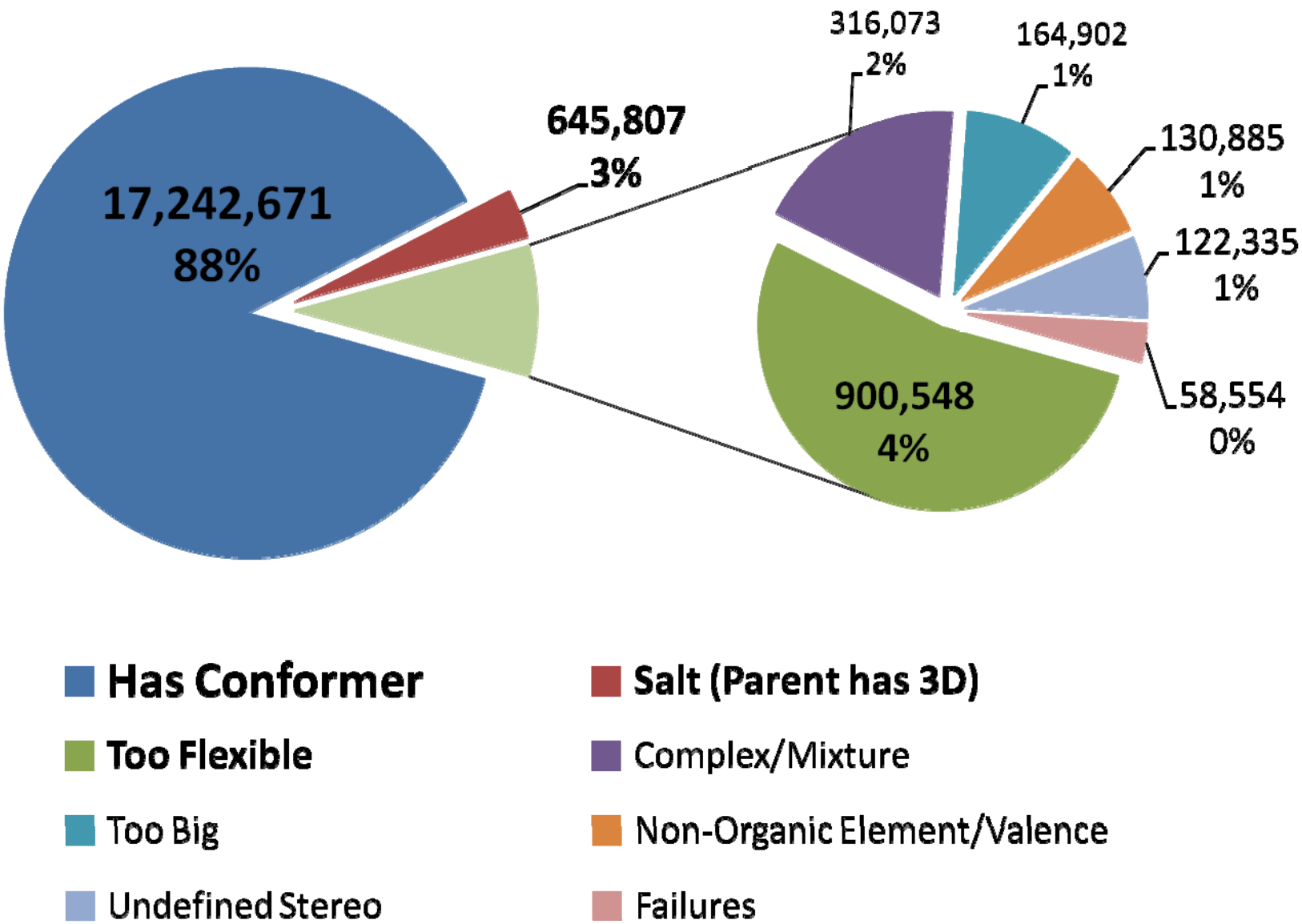
PubChem3D Release Notes

February 2009

PubChem generates [1-3] a theoretical 3D description of each compound in the PubChem Compound database that is not too large (<= 50 non-hydrogen atoms), is not too flexible (<= 15 rotatable bonds), consists of only organic elements (H, C, N, O, F, P, S, Cl, Br, and I), has only a single covalent unit (i.e., not a salt or a mixture), and contains only atom types recognized by the MMFF94s force field [4-5]. At the time of launch, this includes more than 17 million of the 19.5 million records (+87%) in the PubChem Compound database. Considering only the parent forms of salts are considered for 3D (e.g., acetic acid not sodium acetate) and that nearly 0.5 million compounds have a parent with 3D information, ~90% of PubChem Compound has 3D information.

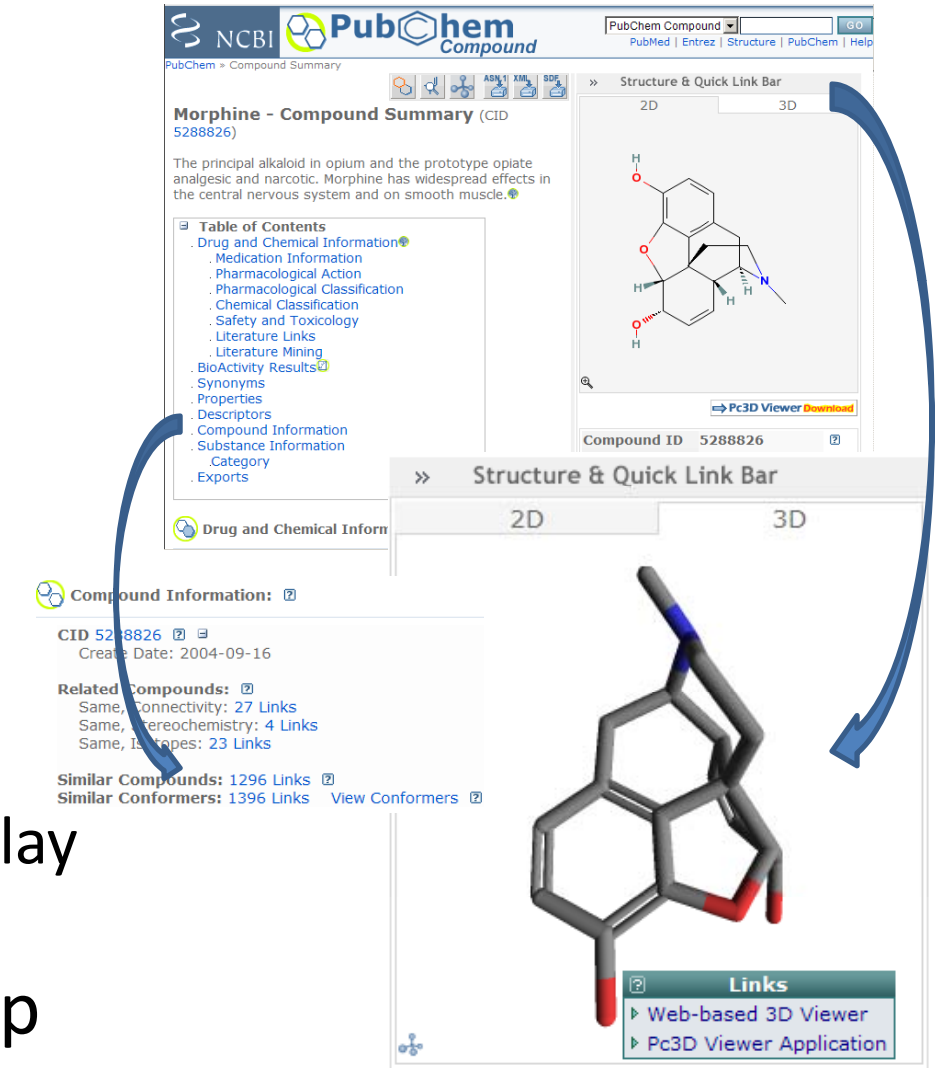
Each provided theoretical 3D conformer is not at an energy minimum and may not represent the lowest energetic form in vacuum, solvent, or a binding pocket. Rather, the theoretical 3D description is a low energy conformer selected from a conformer

Compound 3D Coverage

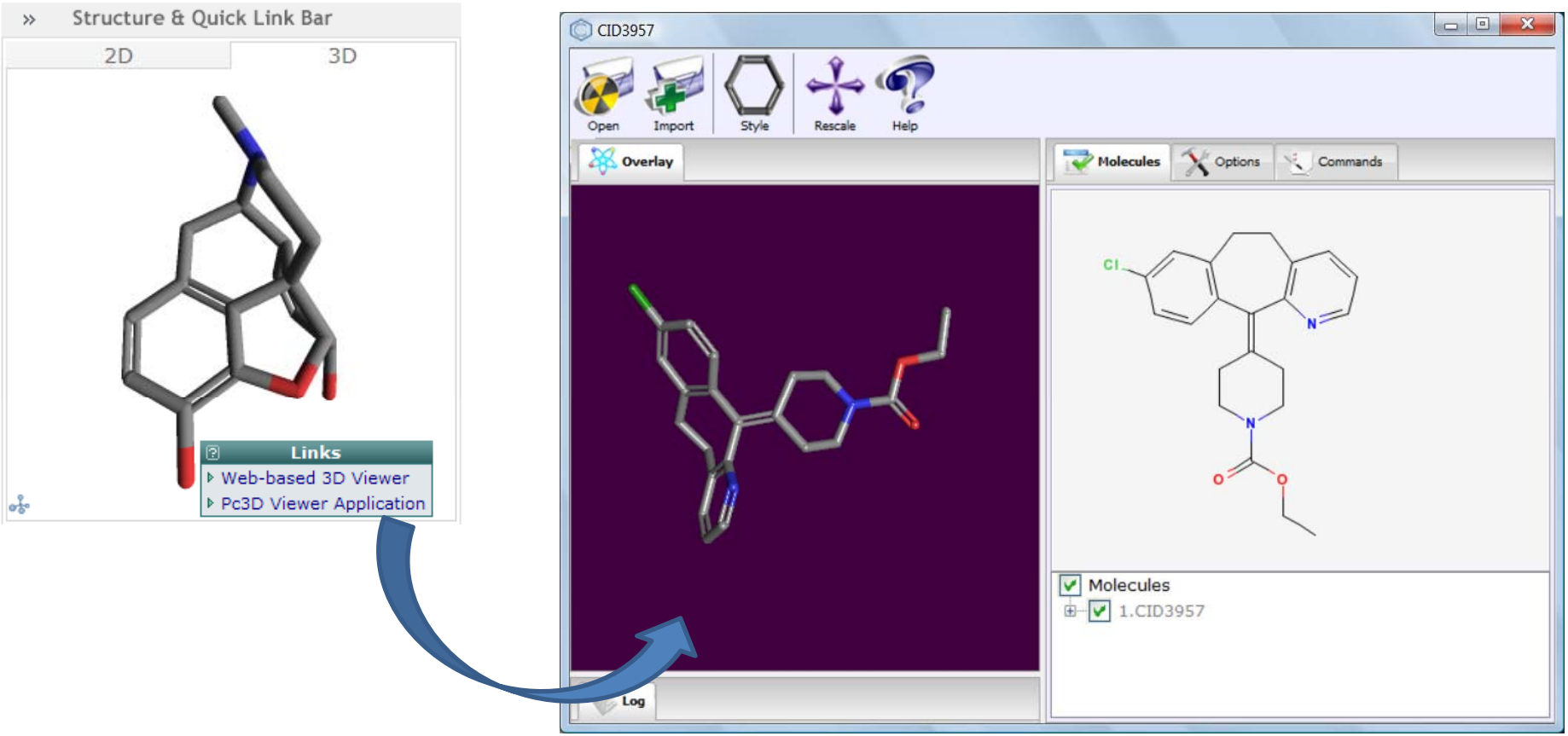


PubChem3D: What does it include?

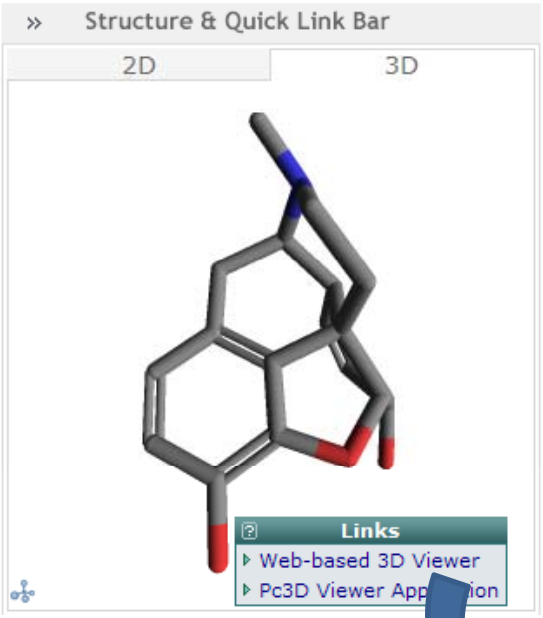
- Single conformer per compound (+17M)
- Download capabilities
  - FTP site or on-demand
- New indexes
  - Has 3D Conformer
  - Vol3D, X3D, Y3D, Z3D
- Helper application
- Web-based viewer
  - Per conformer, per overlay
- Similar conformers neighboring relationship

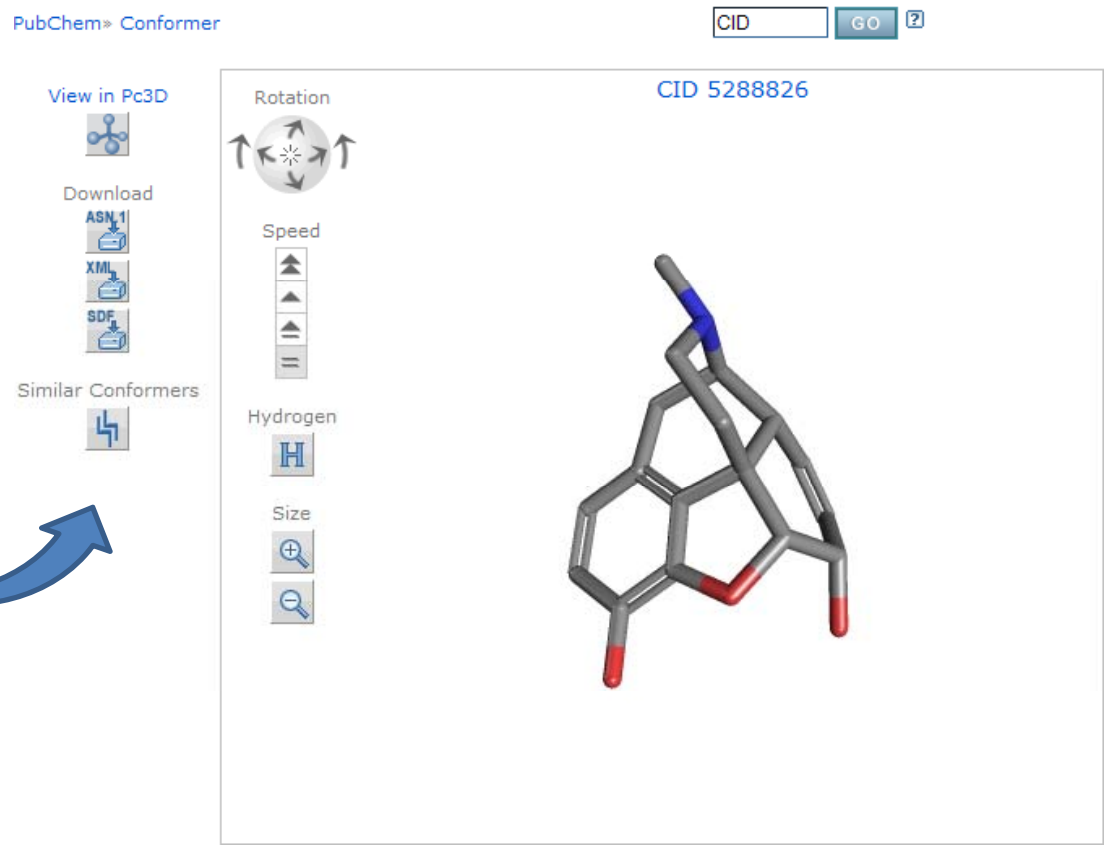


Pc3D Viewer Application

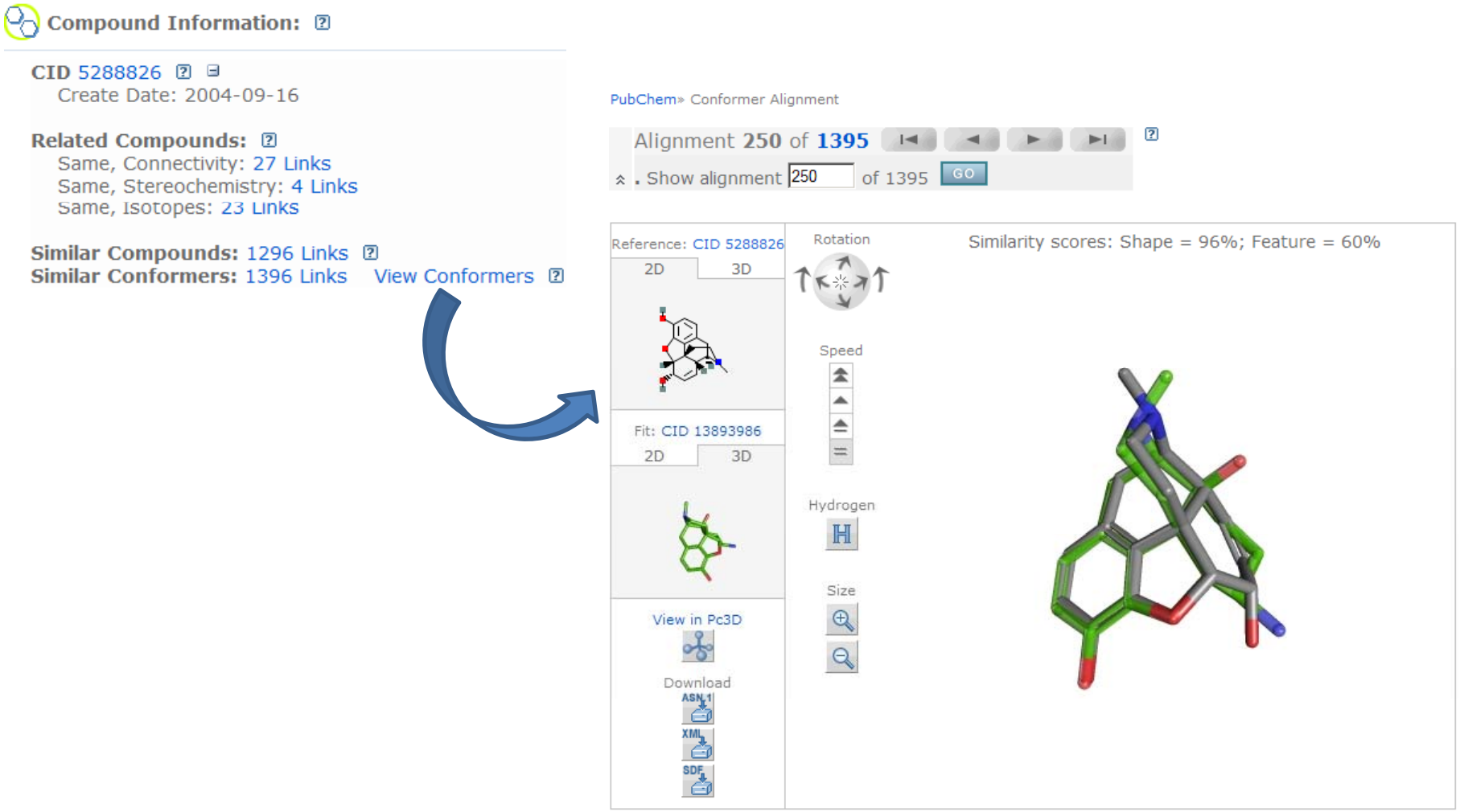


Web-based 3D Viewer





3D Overlay Display



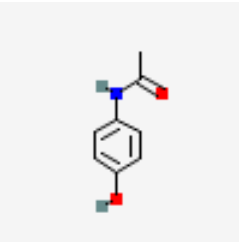
What is a “Similar Conformer”?

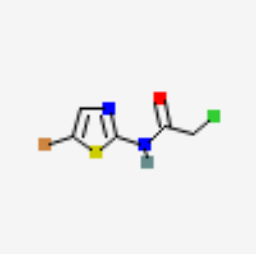
- When two conformers have …

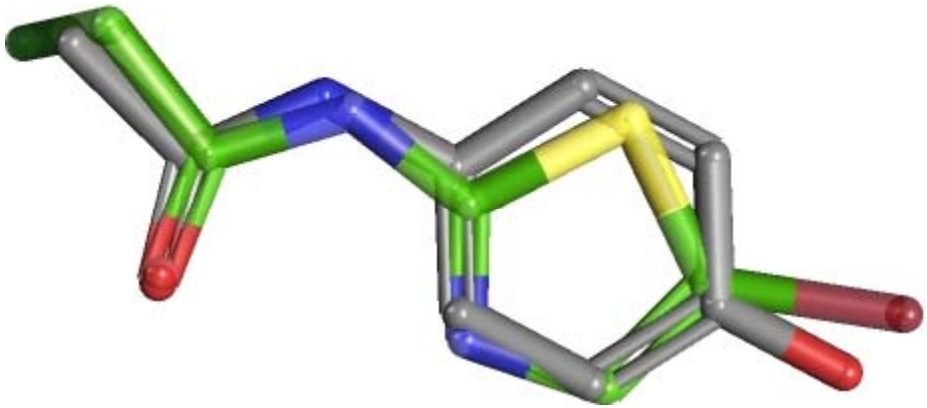
  … similar shapes (ST >= 80)

  … similar features (CT >= 50)

  … BUT only shape optimized

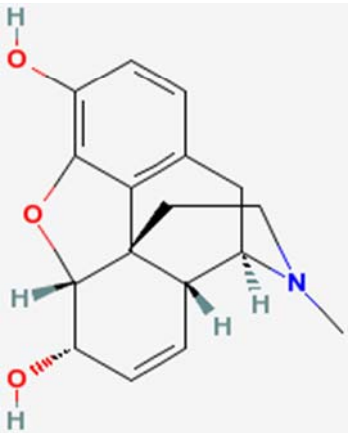


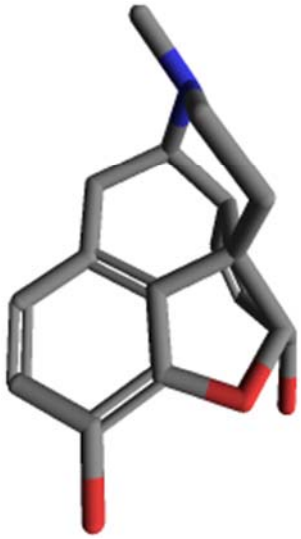




Similarity scores: Shape = 92%; Feature = 54%

2D vs. 3D Similarity Neighbors

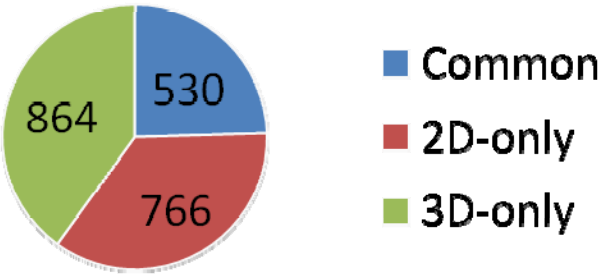




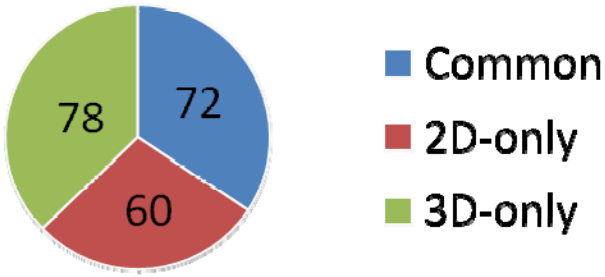
Morphine - Compound Summary (CID 5288826)

The principal alkaloid in opium and the prototype opiate analgesic and narcotic. Morphine has widespread effects in the central nervous system and on smooth muscle.

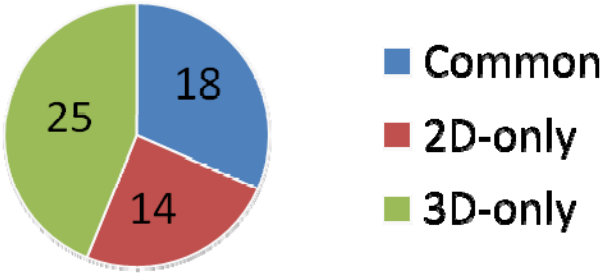
Similarity Links



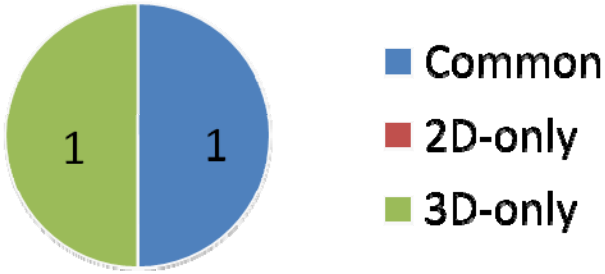
MeSH Links



BioAssay Tested



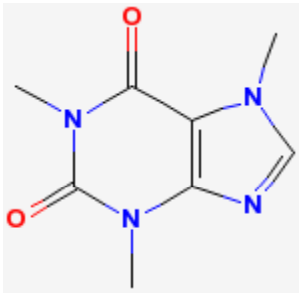
Protein3D Links

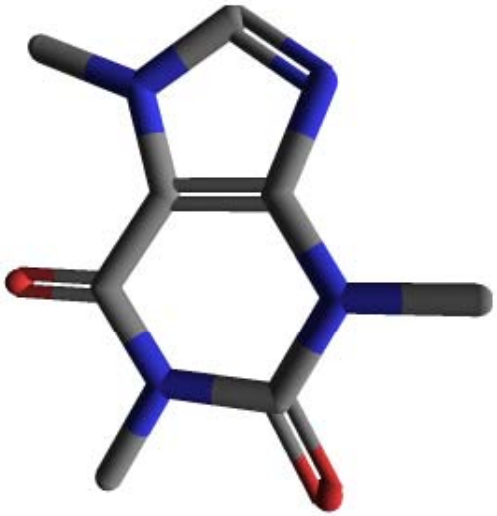


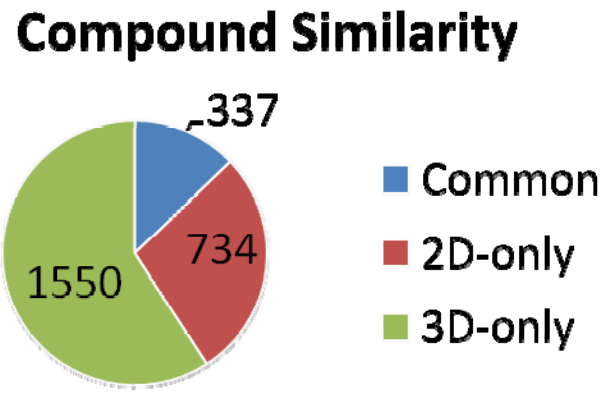
2D vs. 3D Similarity Neighbors

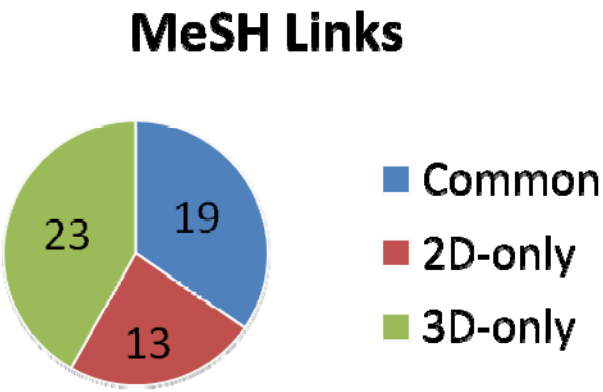
Caffeine - Compound Summary (CID 2519)

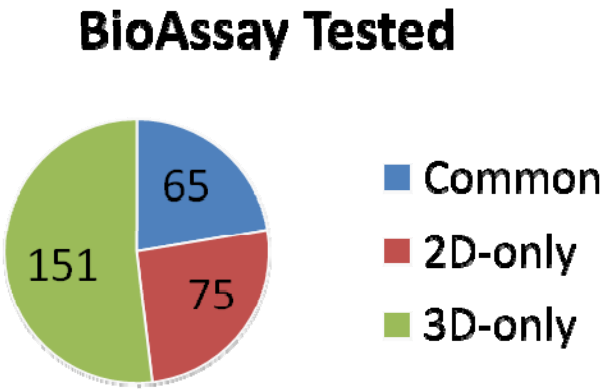
A methylxanthine naturally occurring in some beverages and also used as a pharmacological agent. Caffeine's most notable pharmacological effect is as a central nervous system stimulant, increasing alertness and producing agitation. It also relaxes SMOOTH MUSCLE, stimulates CARDIAC MUSCLE, stimulates DIURESIS, and appears to be useful in the treatment of some types of

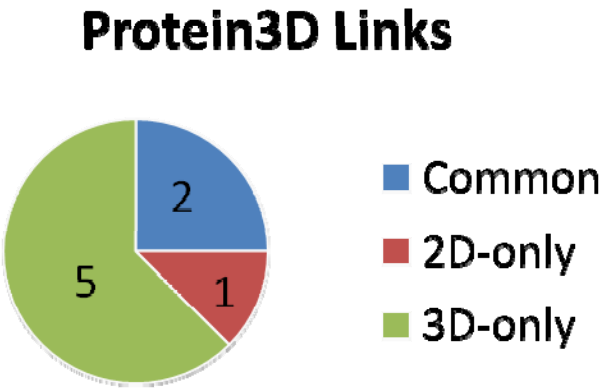








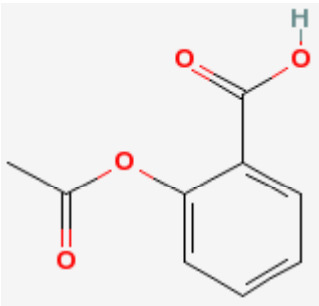


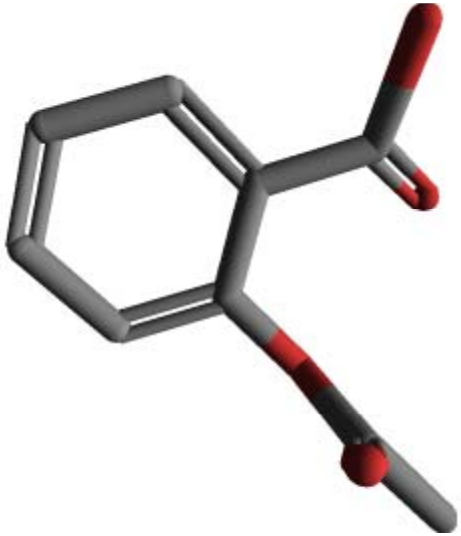


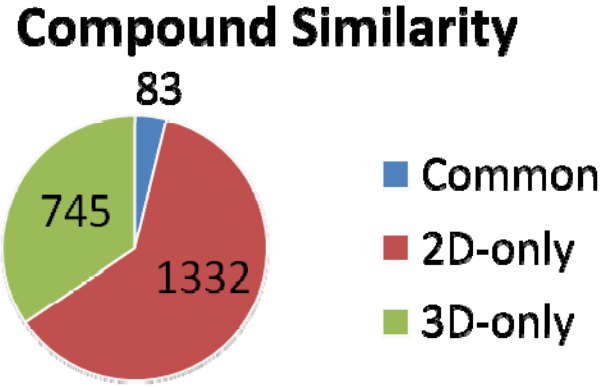
2D vs. 3D Similarity Neighbors

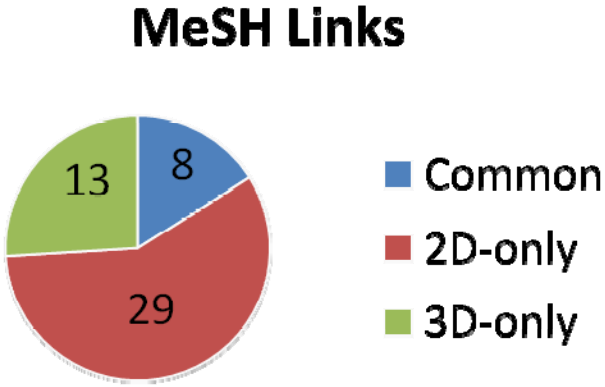
Aspirin - Compound Summary (CID 2244)

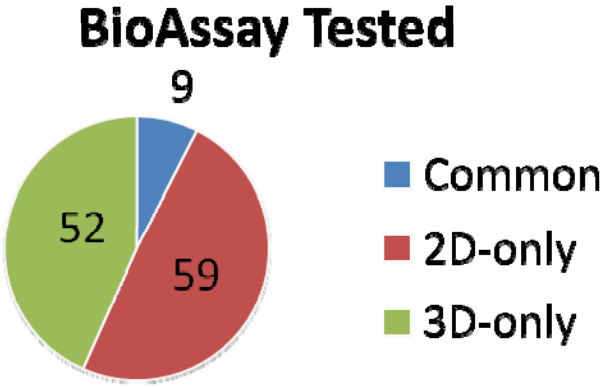
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)

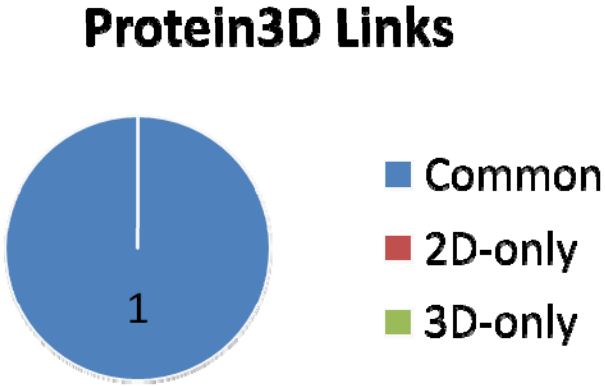












History and Filters… powerful together

NCBI PubChem Compound

My NCBI
Welcome eebolton. [Sign Out]

Search PubChem Compound for #6 AND "pccompound structure"[filter] Preview Go Clear Save Search

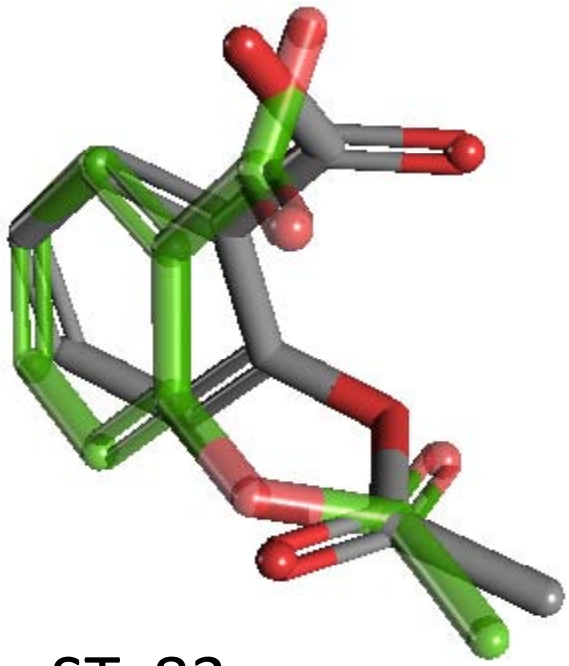
Limits | Preview/Index | History | Clipboard | Details

See Details. No items found.

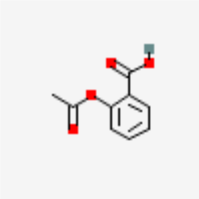
- Search History will be lost after eight hours of inactivity.
- Search numbers may not be continuous; all searches are represented.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use #search, e.g., #2 AND #3 or click query # for more options.

Search	Most Recent Queries	Time	Result
#15	Search **#6 AND "pccompound structure"[filter]**	10:35:41	0
#14	Search **#5 AND "pccompound structure"[filter]**	10:35:38	0
#13	Search **#4 AND "pccompound structure"[filter]**	10:35:33	1
#12	Search **#6 AND "pccompound pcassay"[filter]**	10:35:16	52
#11	Search **#5 AND "pccompound pcassay"[filter]**	10:35:13	59
#10	Search **#4 AND "pccompound pcassay"[filter]**	10:35:07	9
#9	Search **#6 AND "pccompound mesh"[filter]**	10:34:06	13
#8	Search **#5 AND "pccompound mesh"[filter]**	10:34:02	29
#7	Search **#4 AND "pccompound mesh"[filter]**	10:33:57	8
#6	Search **#3 NOT #4**	10:33:44	745
#5	Search **#2 NOT #4**	10:33:39	1332
#4	Search **#2 AND #3**	10:33:28	83
#3	**Similar Conformers** for **PubChem Compound** (Select **2244**)	10:33:07	828
#2	**Similar Compounds** for **PubChem Compound** (Select **2244**)	10:33:01	1415
#1	Search **aspirin**	10:32:54	71

Single Conformer vs. Multi Conformers

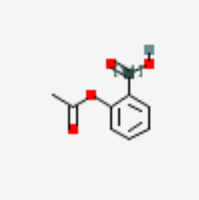


CID: 2244 Related Structures, BioAssays, Literature, Other Links



aspirin; Acylpyrin; Colfarit ...
IUPAC: 2-acetyloxybenzoic acid
MW: 180.157420 g/mol | **MF**: $C_9H_8O_4$
Tested in BioAssays: All: 241, Active: 2; BioActivity Analysis
Cyclooxygenase Inhibitors... more

CID: 450661 Related Structures, Literature

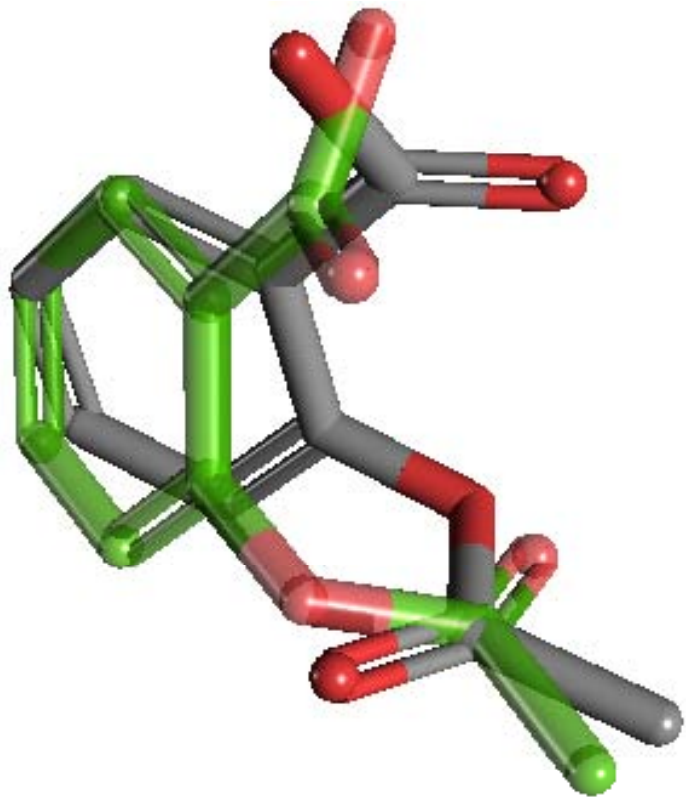


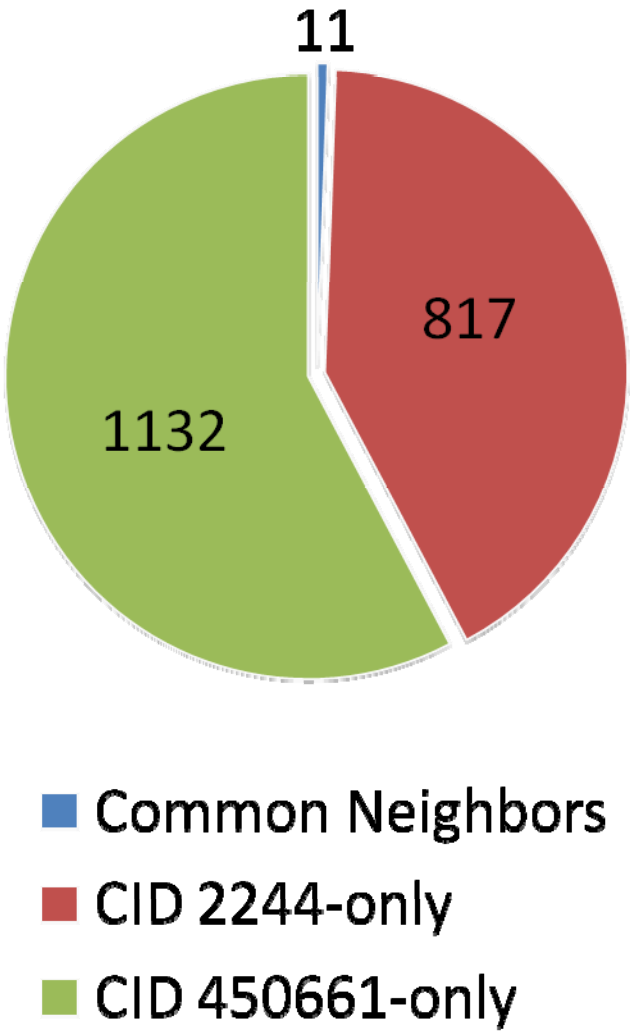
aspirin; MOLI001189; CID450661
IUPAC: 2-acetyloxybenzoic acid
MW: 179.158154 g/mol | **MF**: $C_9H_8O_4$
Cyclooxygenase Inhibitors... more

ST: 83
CT: 27

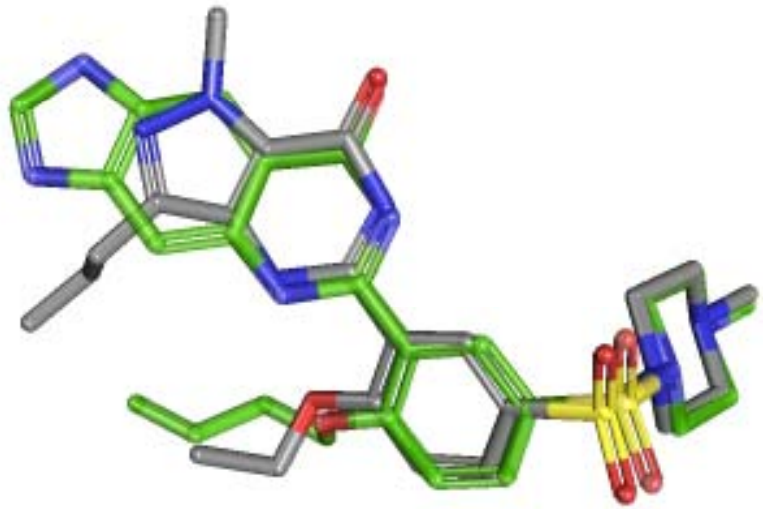
Energy separation <0.5 kcal/mol

Different conformers of structure can give orthogonal neighbor sets





3D Neighbors



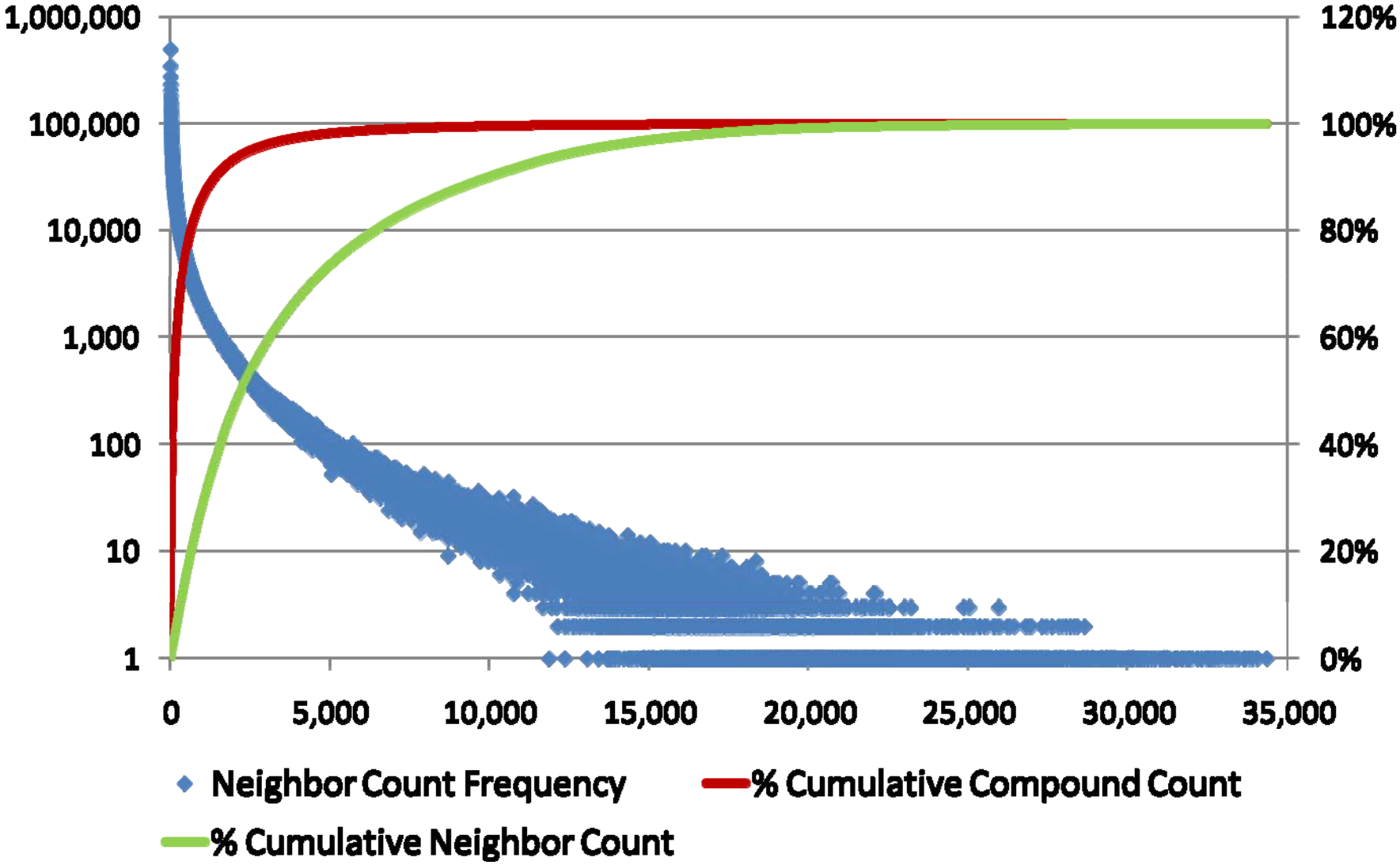
- Lots of them…

  +8B using ST/CT constraints 80/50

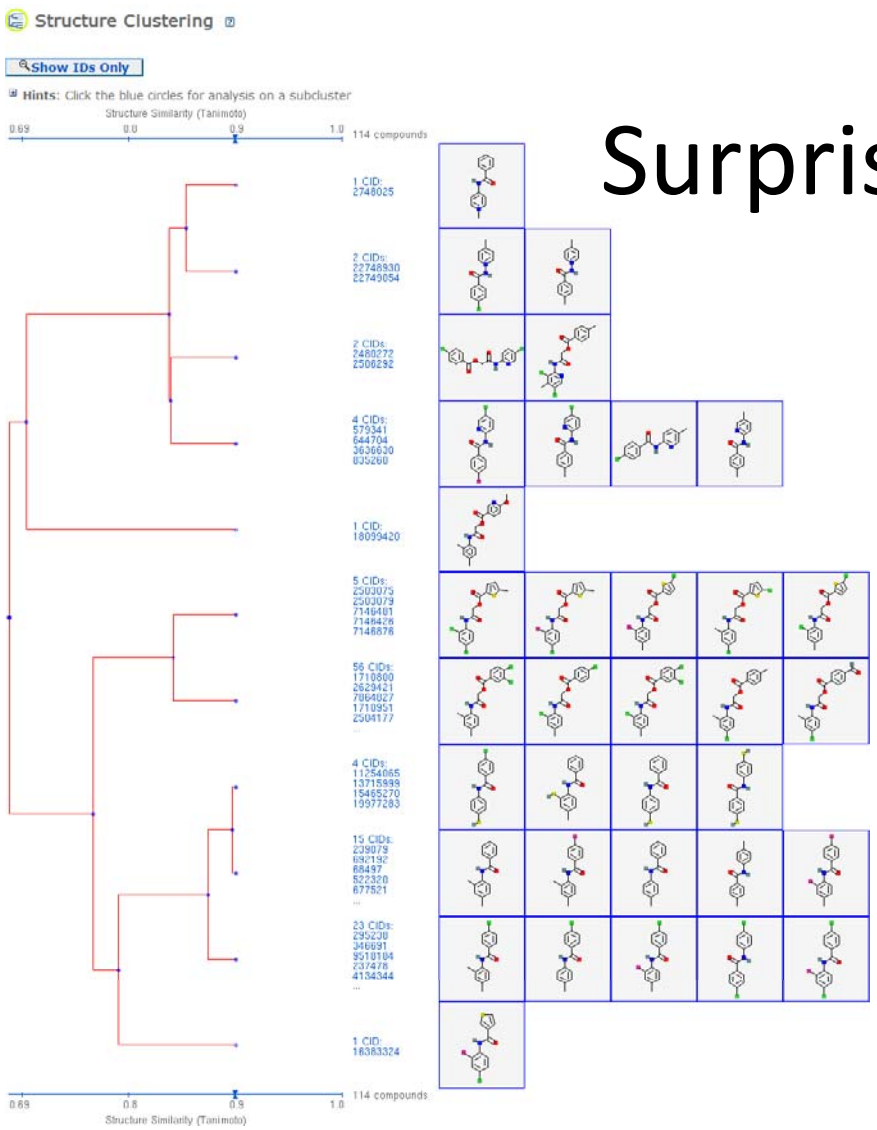
  ~500 per Compound

- Each change in constraints (by +/- 5) changes neighbor count by about a factor of 2

3D Neighbor Distribution

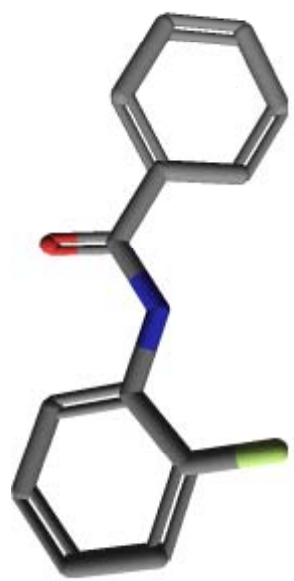


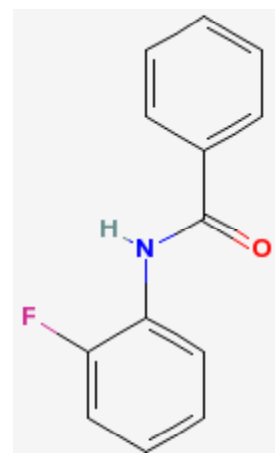
Cases of +30K neighbors



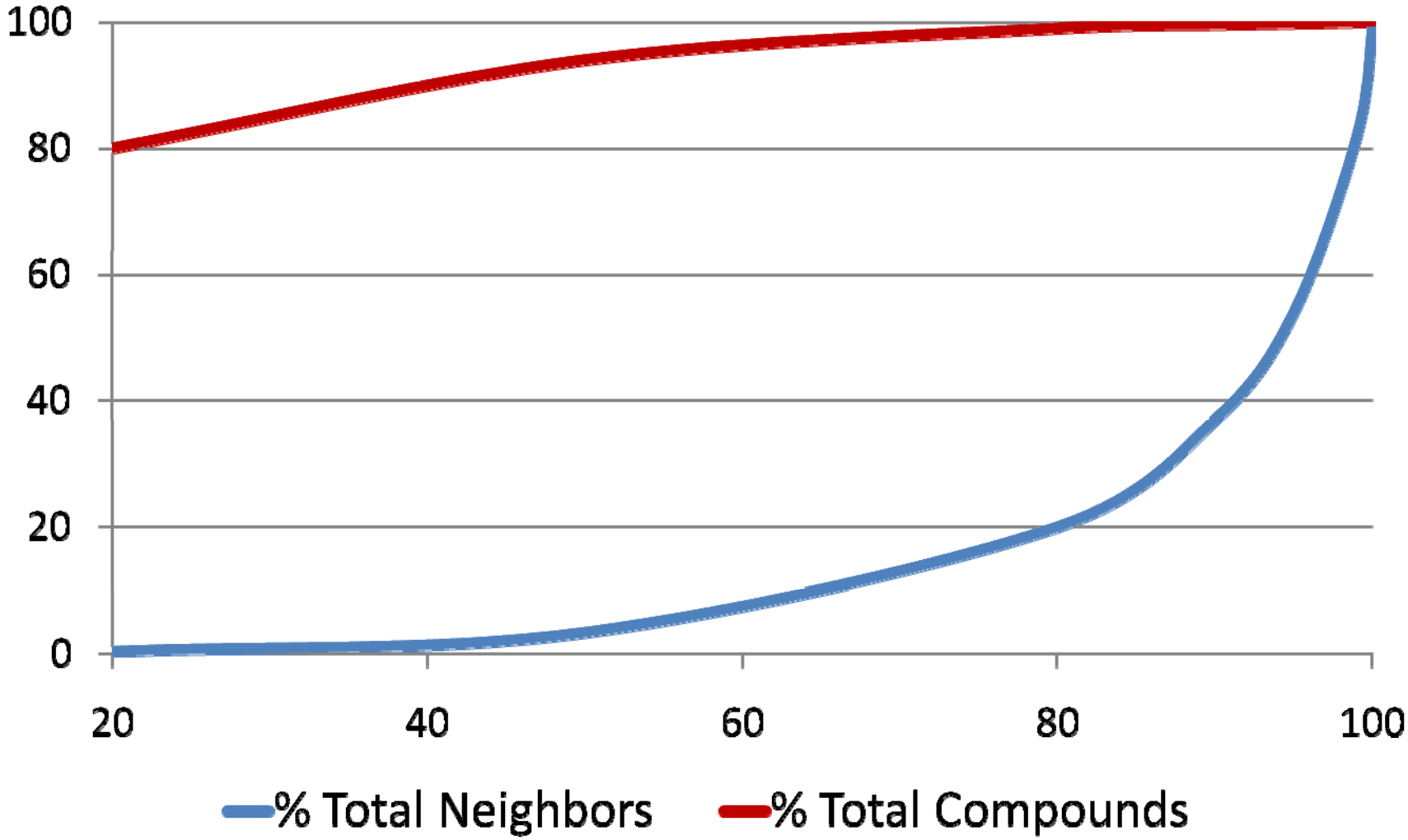
Surprise! ☺ “linked phenyls”

Similar Compounds: 2395 Links
Similar Conformers: 30155 Links





Compound Count vs. Neighbor Count



How did we 3D neighbor PubChem?

- Neighboring a symmetric matrix of size 17M means 146**Q** comparisons

  Oh right… a **Q** is 1,000,000,000,000… e+12

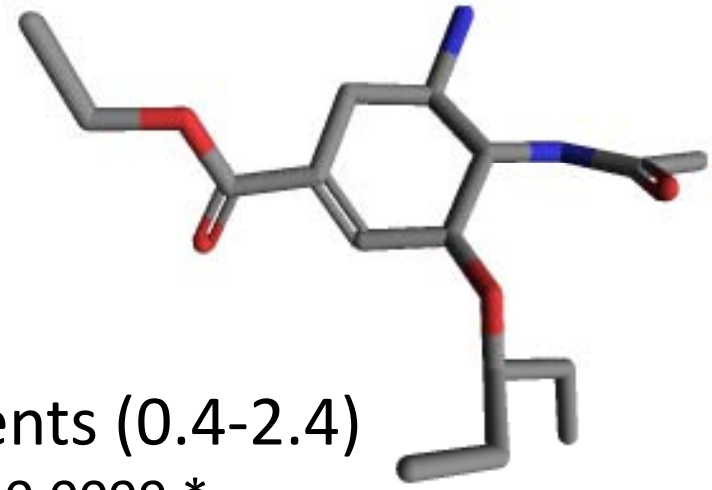
- Brute force @ 1,000 comparisons / sec / core

  = 450K CPU years

This took eight months…

How did we do it?

3D Conformers

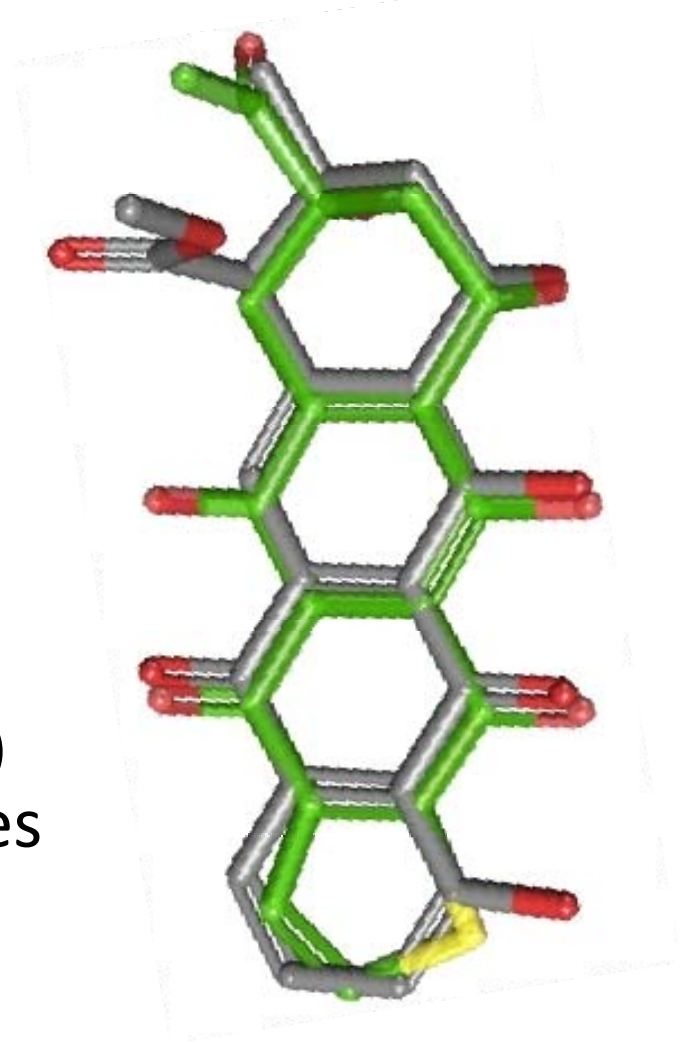


- OMEGA2 C++ API
  - MMFF94s minus coulombic terms
  - 25 kcal/mol energy filtering
  - Max. 100k conformers
- RMSD clustering in 0.2 increments (0.4-2.4)
  - RMSD = 0.219 + 0.040 * n_{eff_rotors} + 0.0099 * n_{heavy_atoms}
  - Max. 500 conformers per model
- Post process
  - Minimize energy of hydrogen locations
  - Prune bumps (+25 kcal/mol)
- Pick single conformer per Compound
  - Lowest energy conformer in entire ensemble

Borodina Y, Bolton E, Fontaine F, Bryant S
Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space.
J Chem Inf Model. 2007 Jul-Aug;47(4):1428-37. PMID: 17569521

3D Neighboring

- OEShape C++ API
  - Bondi radii
  - Heavy atoms only, All radii
  - Analytic methodology
  - All color = true
  - Thresholds: ST >= 80 CT >= 50
- Pre-compute
  - Shape fingerprint
  - Self overlaps (Shape and Feature)
- Three distinct neighboring stages
  - Filter
  - Align
  - Optimize



Stage 1: Filter

- Possible for two conformers to be a neighbor?

  Use pre-computed **A** and **B** self overlap

  Tanimoto Equation: $T = AB / (A + B – AB)$

  $AB = (A + B) * T / (1 + T)$

  if AB > than A or B, cannot be a neighbor

Stage 1: Filter

- Possible for two conformers to be a neighbor?

  Use pre-computed **A** and **B** self overlap

  Tanimoto Equation: $T = AB / (A + B - AB)$

  $AB = (A + B) * T / (1 + T)$

  if AB > than A or B, cannot be a neighbor

  But shape overlap AB can be larger than A and B!

  Actual filter relationship at ST=0.8:

  $0.75 <= 2.25 * A / (A + B) <= 1.50$

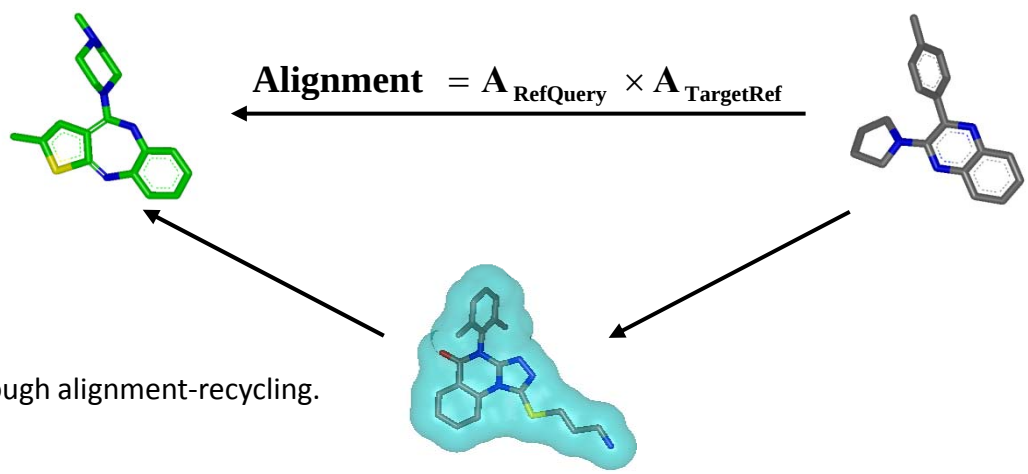
  $0.75 <= 2.25 * B / (A + B) <= 1.50$

  where: $2.25 = (1 + T) / T = 1.8 / 0.8 = 2.25$

Stage 2: Align

- Use pre-computed shape fingerprints
- Find common reference shapes
  - If none in common, then not a neighbor
- Loop over and "replay" alignment to reference shape
- Compute (grid) shape overlap (AB)
- Keep alignment giving best shape overlap
- Compute analytic ST

  ST<75 not a neighbor

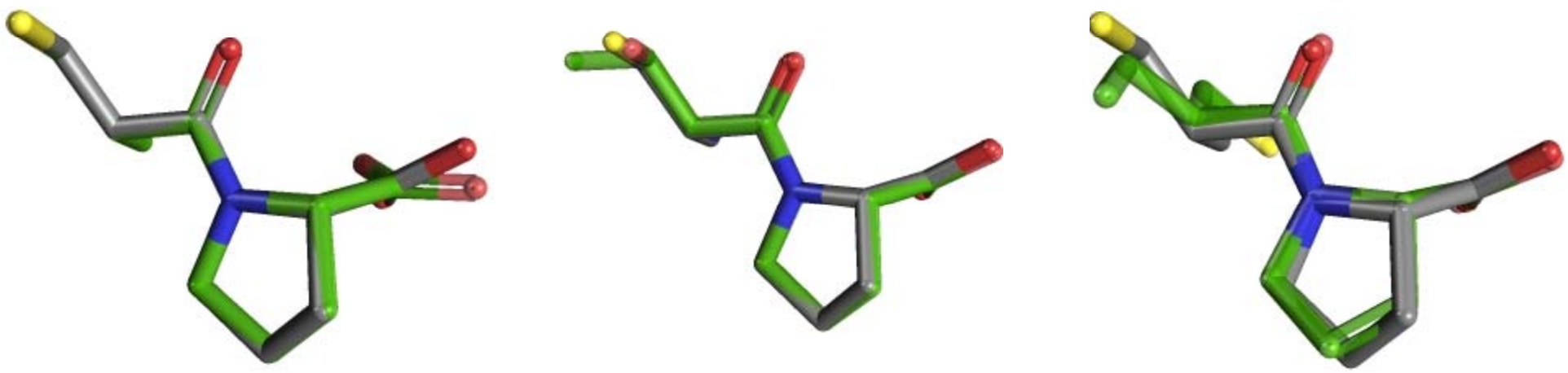


Fontaine F, Bolton E, Borodina Y, Bryant S
Fast 3D shape screening of large chemical databases through alignment-recycling.
Chem Cent J. 2007 Jun 6; 1-12. PMID: 17880744

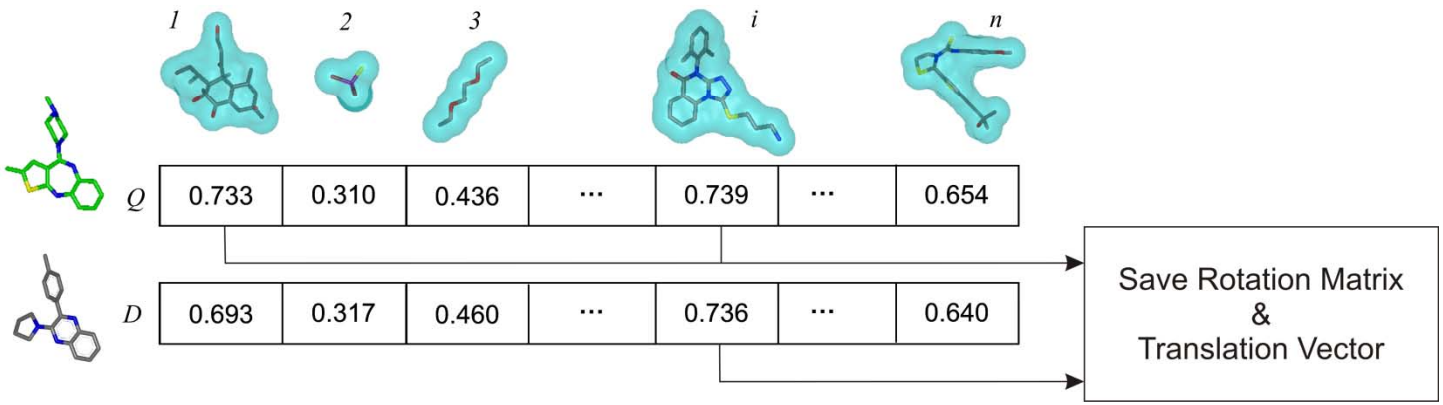
Stage 3: Optimize

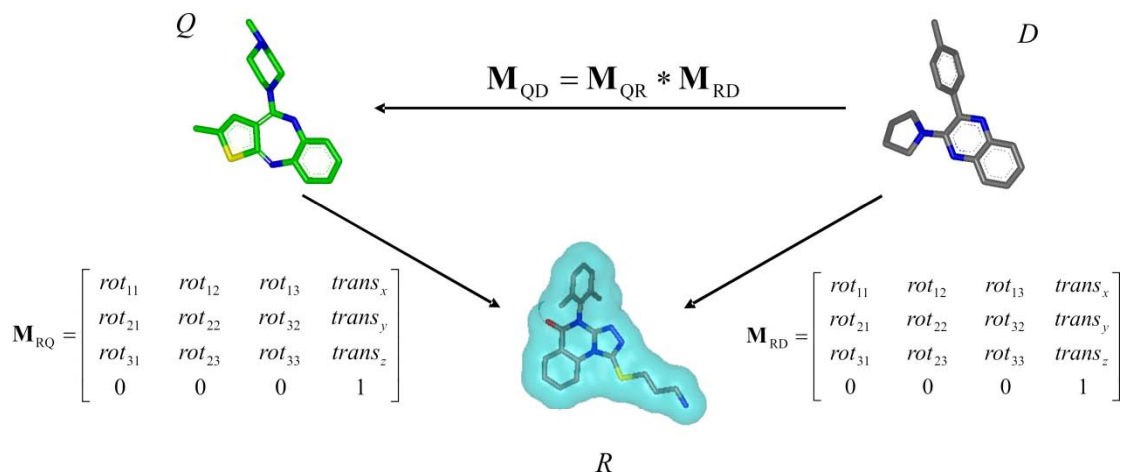
- Use Stage 2 alignment as starting point
- Optimize analytic shape overlay “AsIs”
- Perform single point CT calculation

One optimization not many…



Alignment Recycling





Fontaine F, Bolton E, Borodina Y, Bryant S
Fast 3D shape screening of large chemical databases through alignment-recycling.
Chem Cent J. 2007 Jun 6; 1-12. PMID: 17880744

Shape Fingerprint Definition

Volume Range	ST	Reference Count	Conformer Count
1-165	75	251	40M
166-199	70	274	72M
200-238	65	271	167M
239-285	60	349	448M
286-344	55	361	768M
345-432	50	473	519M
433-999	45	633	116M
		2,612	**Total Refs**

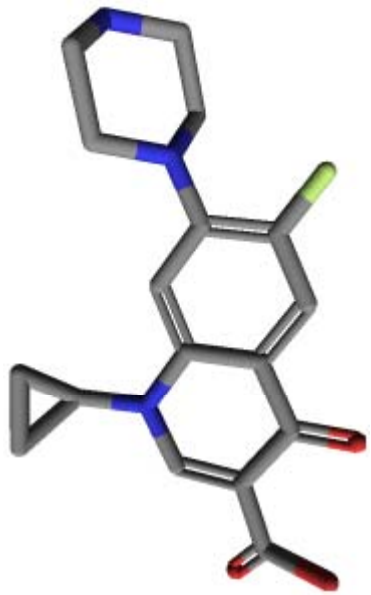
- Represents all unique shapes for 17M compounds and their +2B conformers
- No (useful) correlation between count of common reference shapes and shape similarity
- Can have only one common shape yet be highly similar

Method Efficiency

- Case study… neighbor 5,000 compounds
  - 854,531,025,000 comparisons (5K by 17M)
  - Filtering performance (to remove from consideration)
    - ST Filter: 7% 61B CPU Cost 0%
    - CT Filter: 17% 145B CPU Cost 0%
    - FP Align: 76% 649B CPU Cost 39%

    - ST Optimize: 295M times (0.034% … 1 in 2,941)
    - CT Compute: 56M times (0.0064% … 1 in 15,625)
  - Total time: 30 hours on 1500 cores

  - All this for 954K unique neighbors (382/Compound)



How to neighbor faster?

- 40% (12H -> 6H?)… identify similar and align
  - Avoid consideration of dissimilar shapes
  - Use Quadrupoles?
- 30% (9H -> 4.5H?)… align similar shapes
  - “Analytic” optimization is expensive
  - Optimize on grid then compute analytic overlap?
- 30% (9H -> 3H?)… compute feature similarity
  - Re-computation of CT self overlap is wasteful
  - Compute overlap (AB) only?

Possible Future Directions

- On-demand structure clustering
  - Using shape/feature similarity
- Bioassay data analysis
  - (help) separate actives from inactives
- Search system
- Expose more than one conformer

PubChem Crew ...

Steve Bryant

Yulia Borodina

Jie Chen

Svetlana Dracheva

Lewis Geer

Lianyi Han

Jane He

Siqian He

Karen Karapetian

Sunghwan Kim

Wenyao Shi

Ben Shoemaker

Vahan Simonyan

Tugba Suzek

Paul Thiessen

Valery Tkachenko

Jiyao Wang

Yanli Wang

Jewen Xiao

Jian Zhang

NCBI