

PubChem3D: Relative Diversity of Shape

Evan Bolton, Ph.D.

NCBI/NLM/NIH

CUP IX

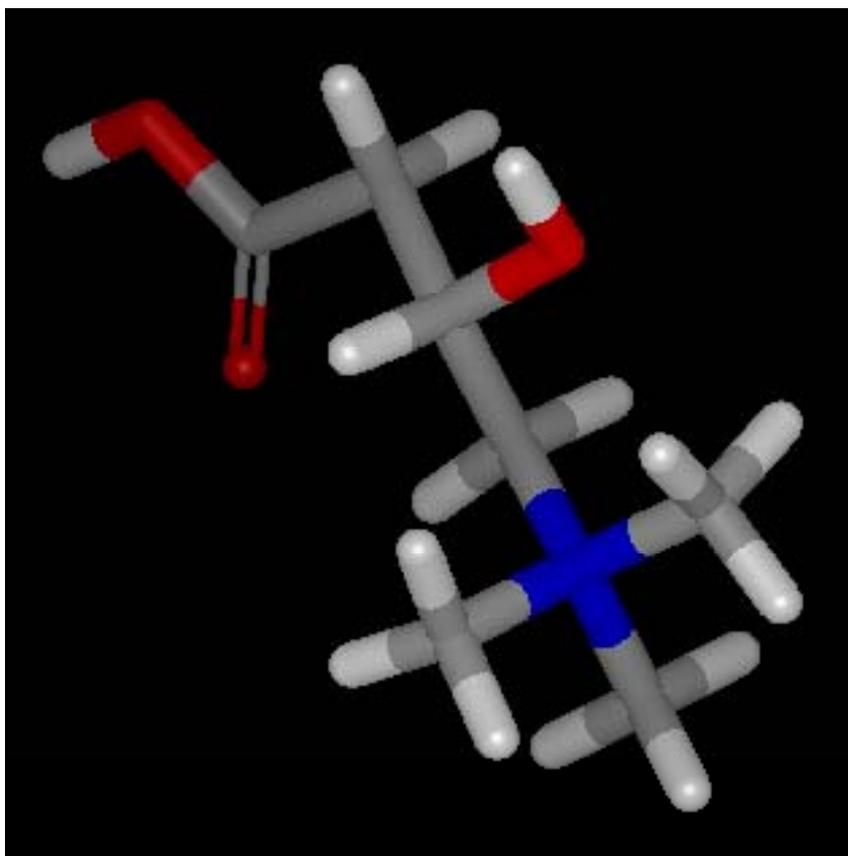
bolton@ncbi.nlm.nih.gov



Why 3D?

- **2D**, while useful, isn't enough
- Adds a **new** dimension to data mining
- Ability to “scaffold hop”
- Enable ability to cluster/neighbor by...
 - Shape ... Pharmacophore ... Electrostatics
- Different insights from same (biological) data
- Possible to ...
 - ... better integrate with other NCBI projects/data
 - ... consider advanced Virtual Screening methods
 - ... etc.

3D Description of



OMEGA2 Validation Study

- Most default values are great for general use
 - ... Sufficient default fragment sampling
 - ... Fragment force field variation provides little benefit
 - ... Conformer force field default more accurate than full MMFF94
 - ... Reasonable default energy windows for Fragment/Conformer

However ...

- ... Maximum total Conformer Count (100k) may limit overall accuracy
- Use of full MMFF94 force field produces fewer conformers but substantially reduces accuracy (but less so as energy increases...)
- Average RMSD and Shape accuracy linearly “degrades” with size and flexibility
- Shape accuracy is less demanding on energy window than RMSD

How do we do 3D?

- OMEGA2 (v2.1 and v2.2.1 C++ API)
- MMFF94s minus coulombic terms (NoEstat)
- 25 kcal/mol Energy Window (frag and gen)
- Maximum of 100,000 conformers generated
- RMSD Sampling ... but at what RMSD?
- Post Process Cleanup
 - Maximum 500 conformers (recluster at higher RMSD)
 - Energy minimization of hydrogen locations
 - Prune “bumps” (25 kcal/mol threshold)

“Molecular Flexibility” Measure

“Effective Rotors” ... takes into account ring flexibility

$$nr_{\text{effective}} = nr + \frac{nnara}{5}$$

- *nr* is the number of rotatable bonds
- *nnara* is the number of “non-aromatic” SP₃-hybridized ring atoms

“Effective Rotors” Concept

$$\ln(NC) = 0.02 + 1.007 \cdot \ln(NC_0) - \ln\left(\frac{RMSD}{RMSD_0}\right) (0.513 \cdot nr + 0.108 \cdot nnara)$$

$$R^2 = 0.92 \text{ and } RSE = 0.57, \text{ with } N = 6,766$$

where:

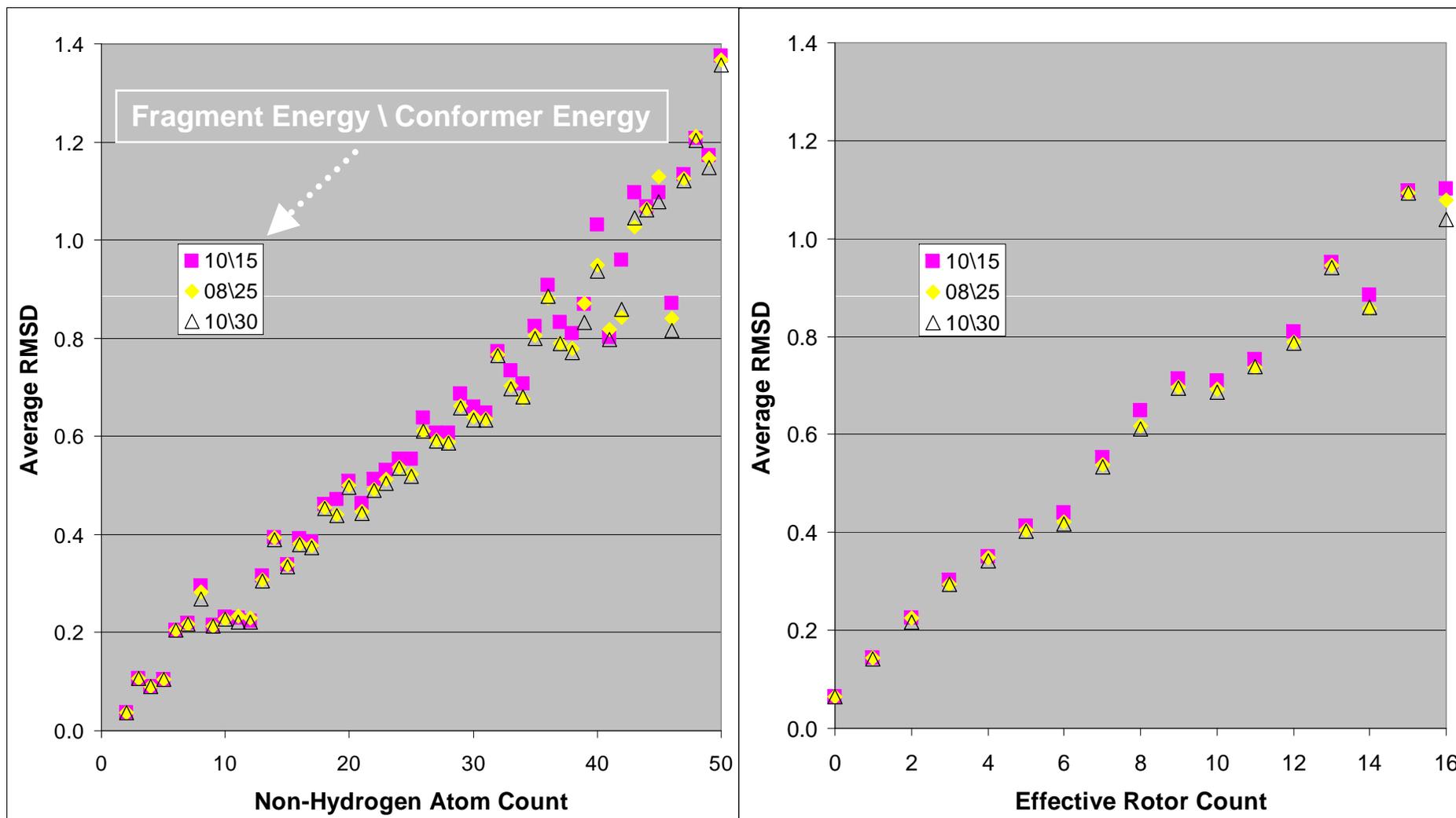
- **RMSD** is the desired root-mean-square-deviation
- **NC** is the number of conformers estimated at **RMSD**
- **RMSD₀** is 2.0 Å
- **NC₀** is the actual number of conformers produced at RMSD 2.0 Å
- **nr** is the number of rotatable bonds
- **nnara** is the number of “non-aromatic” SP₃-hybridized ring atoms

Borodina Y, Bolton E, Fontaine F, Bryant S

Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space.

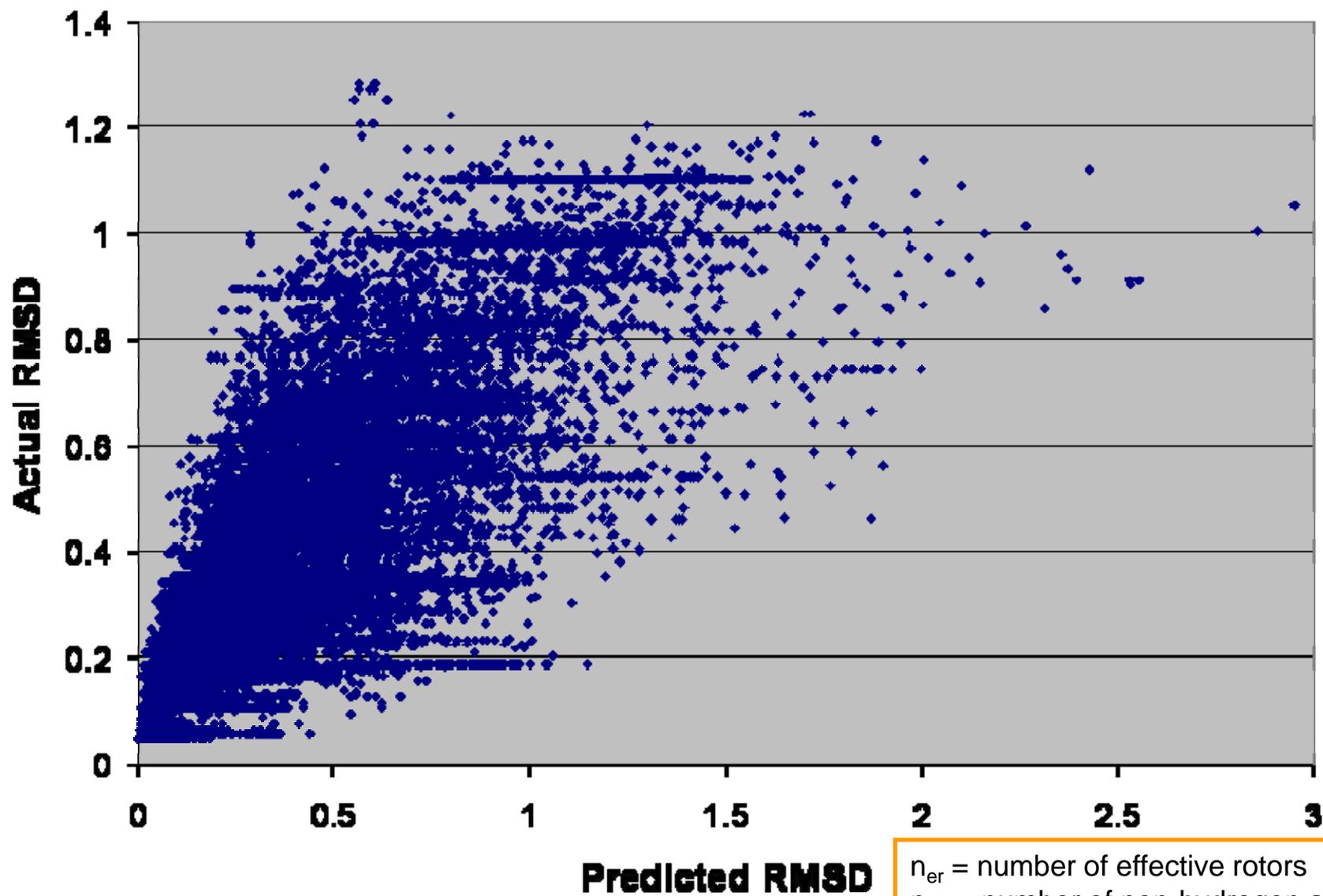
J Chem Inf Model. 2007 Jul-Aug;47(4):1428-37. PMID: 17569521

OMEGA2 RMSD Accuracy



Average RMSD values for all 25,972 3-D reference structures as a function of Molecular Size and Flexibility

Predicting RMSD Accuracy

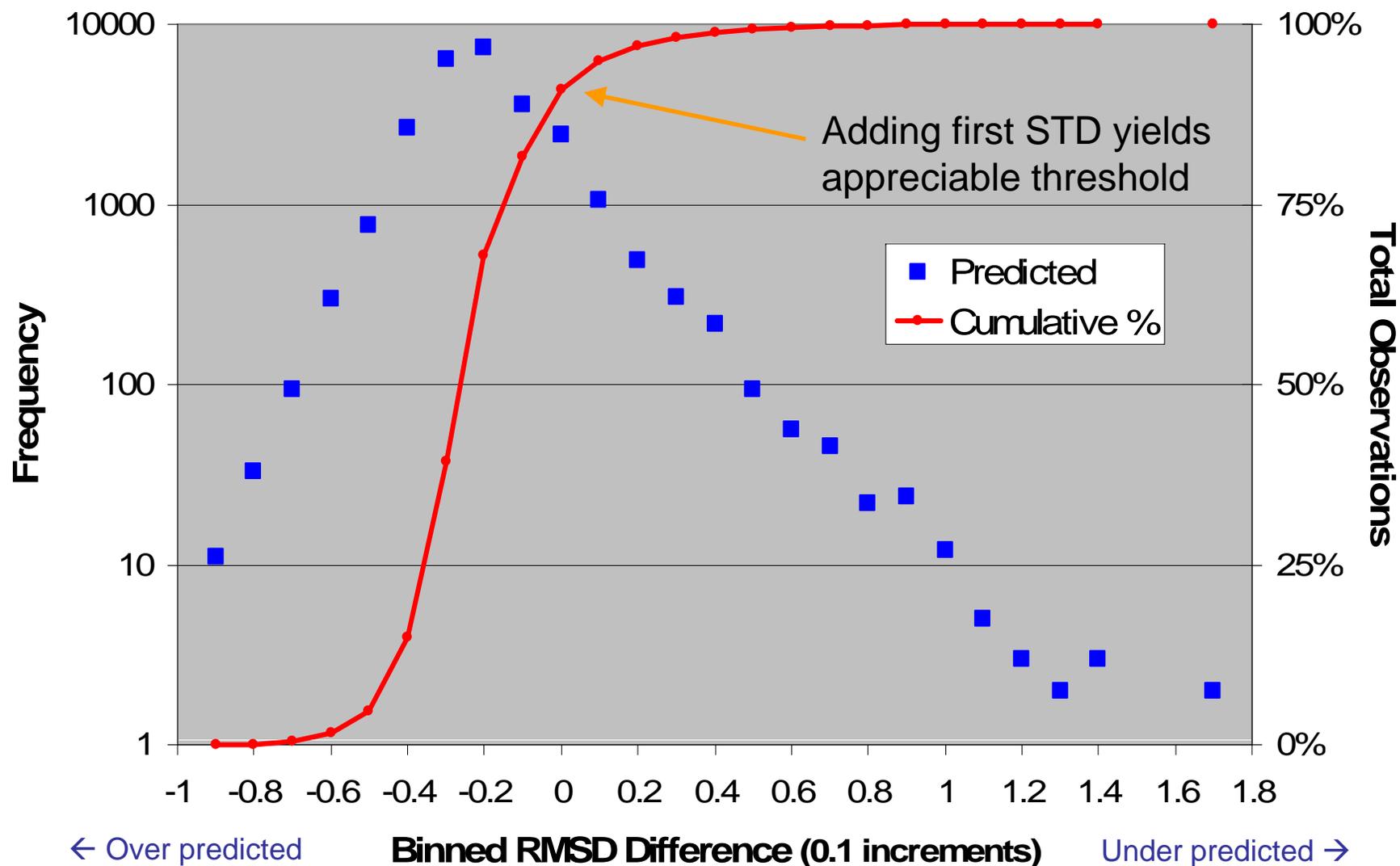


n_{er} = number of effective rotors
 n_{ha} = number of non-hydrogen atoms

$$\text{RMSD} = 0.029 + 0.040 * n_{er} + 0.0099 * n_{ha} \quad R^{**2} = 0.65 \text{ +/- } 0.19 \quad N=25,972$$

$$\text{RMSD} = 0.19 + 0.029 + 0.040 * n_{er} + 0.0099 * n_{ha}$$

Creating a PubChem3D RMSD Threshold



RMSD sampling threshold?

$$\text{RMSD} = 0.029 + 0.040 * n_{er} + 0.0099 * n_{ha}$$

$R^{*2} = 0.65 \pm 0.19 \quad N=25,972$

RMSD = 0.19 + RMSD (and round up to nearest 0.2 incr)

RMSD_{pred} > RMSD_{experiment} 91% of the time

0.1 + RMSD_{pred} > RMSD_{experiment} 95% of the time

n_{er} = number of effective rotors
 n_{ha} = number of non-hydrogen atoms

How do we do 3D?

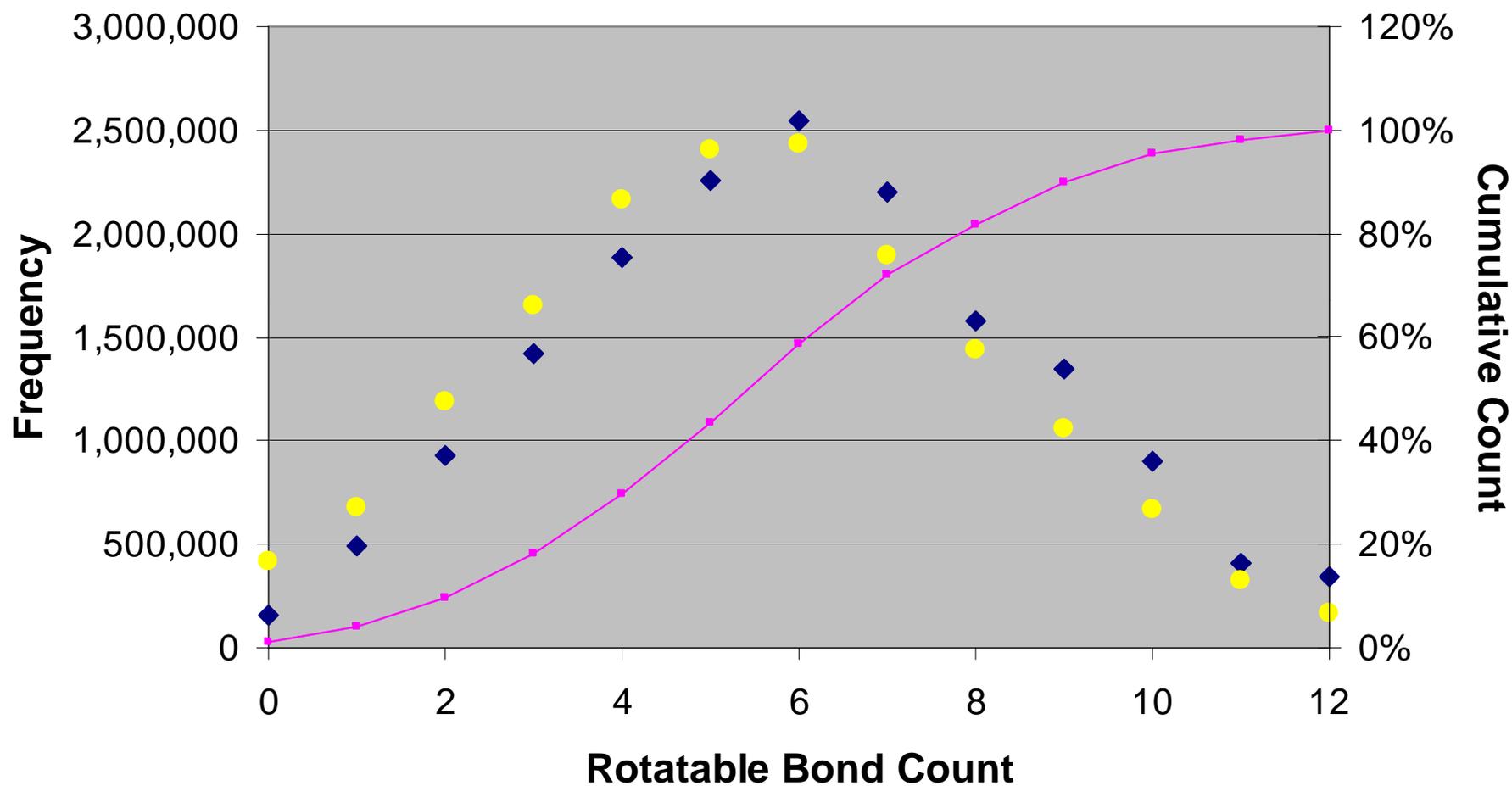
- OMEGA2 (v2.1 and v2.2.1 C++ API)
- MMFF94s minus coulombic terms (NoEstat)
- 25 kcal/mol Energy Window
- Maximum of 100k conformers generated
- **RMSD Sampling** (in 0.2 incr. based on structure)
- Post Process Cleanup
 - Maximum 500 conformers (recluster at higher RMSD)
 - Energy minimization of hydrogen locations
 - Prune “bumps” (25 kcal/mol threshold)

PubChem3D “Scope”

- Non-hydrogen Atoms ≤ 50
 - 18.6 million of 19.1 million “*public*” CIDs (97.3%) (24.7m total CIDs)
- Rotatable Bonds ≤ 15
 - 18.5 million (96.8%)
- Undefined Stereo Centers ≤ 5
 - Atom 18.9 million (99.0%)
 - Bond 19.1 million (99.9%)
- Covalent Unit Count $= 1$
 - 18.2 million (95.3%)
- Overall Potential Coverage of “Live” portion of PubChem
 - 17.3 million (90.6%)

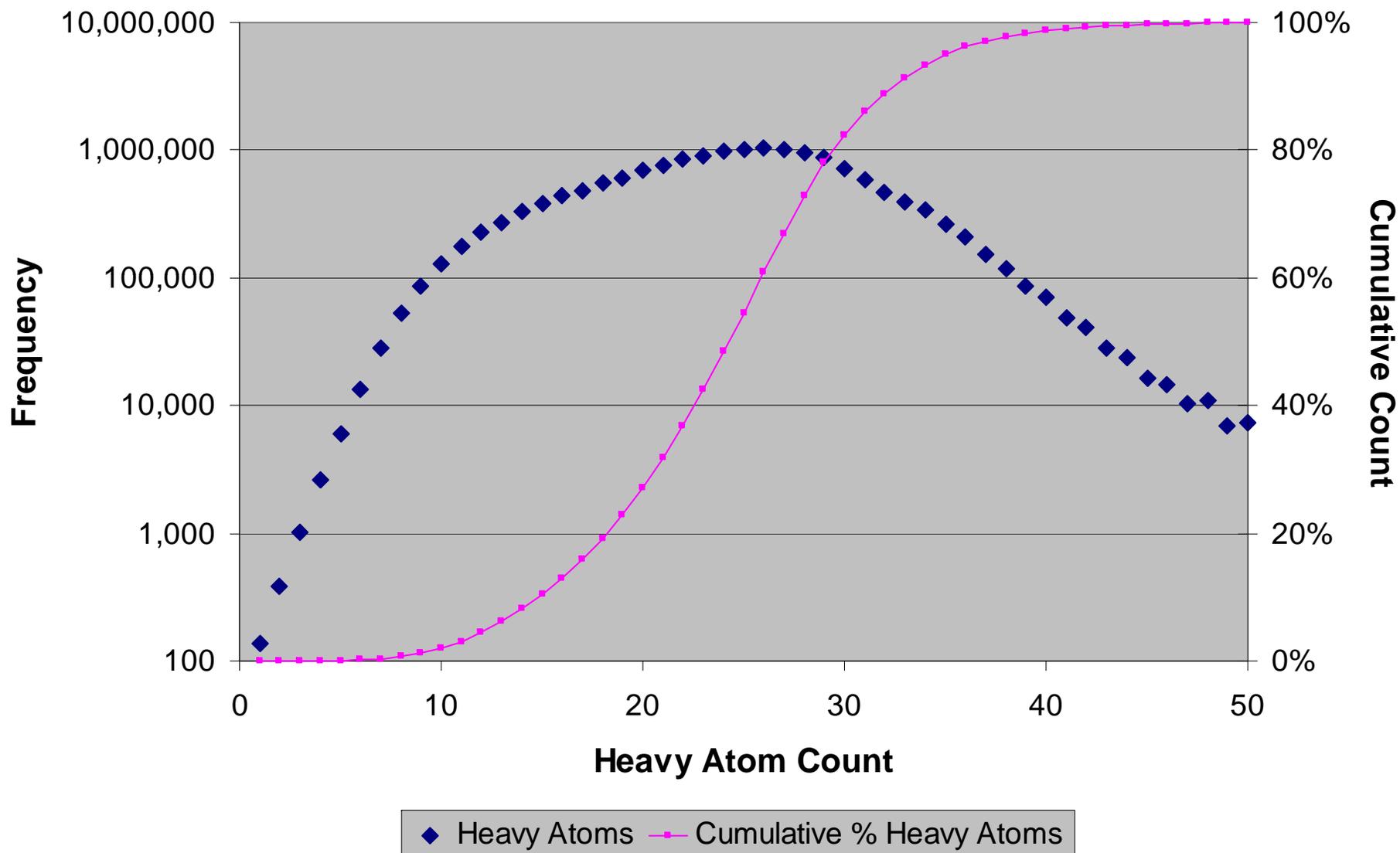


PubChem3D Compound Rotatable Bonds

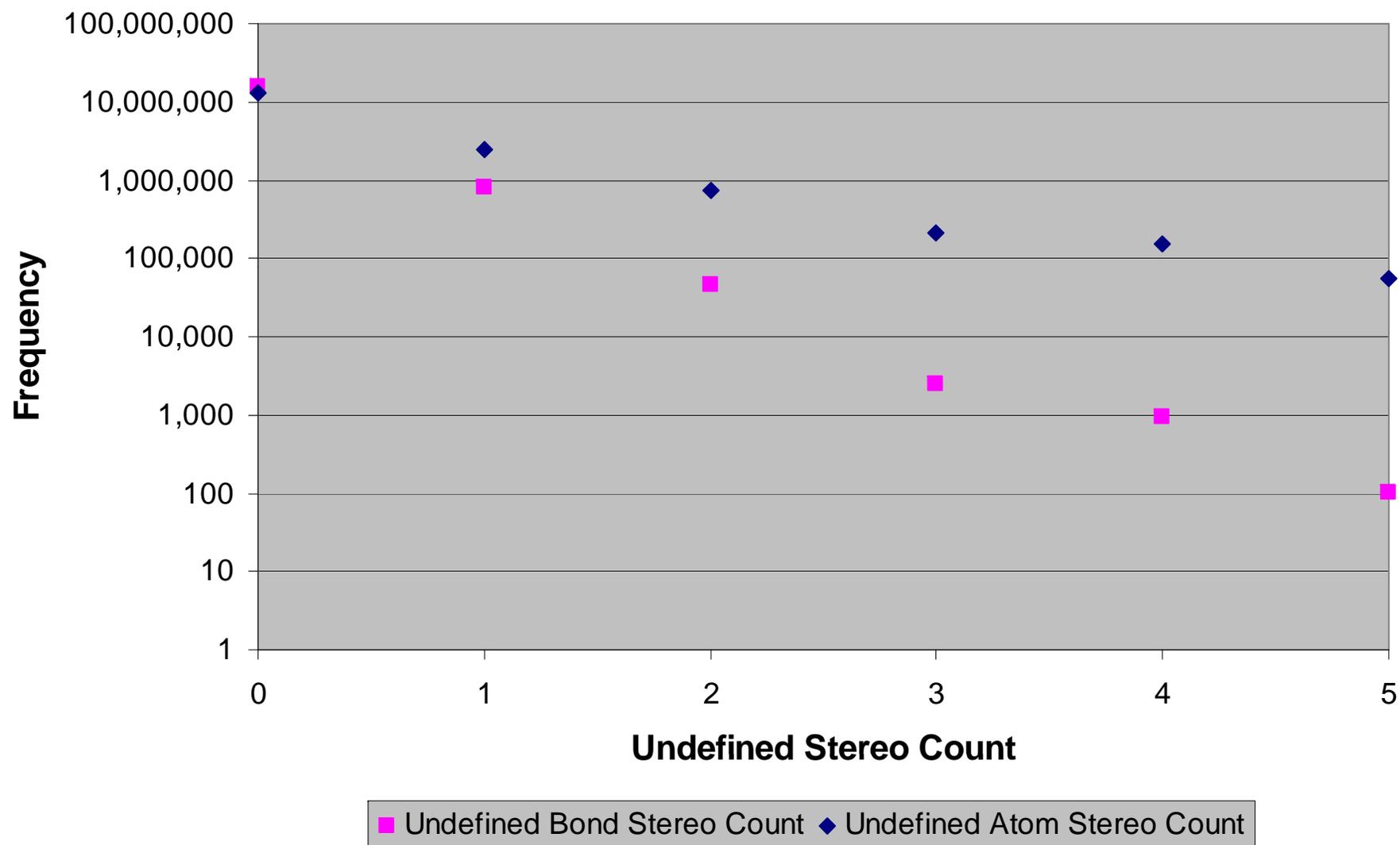


◆ Effective Rotors ● Rotatable Bonds — Cumulative % Effective Rotors

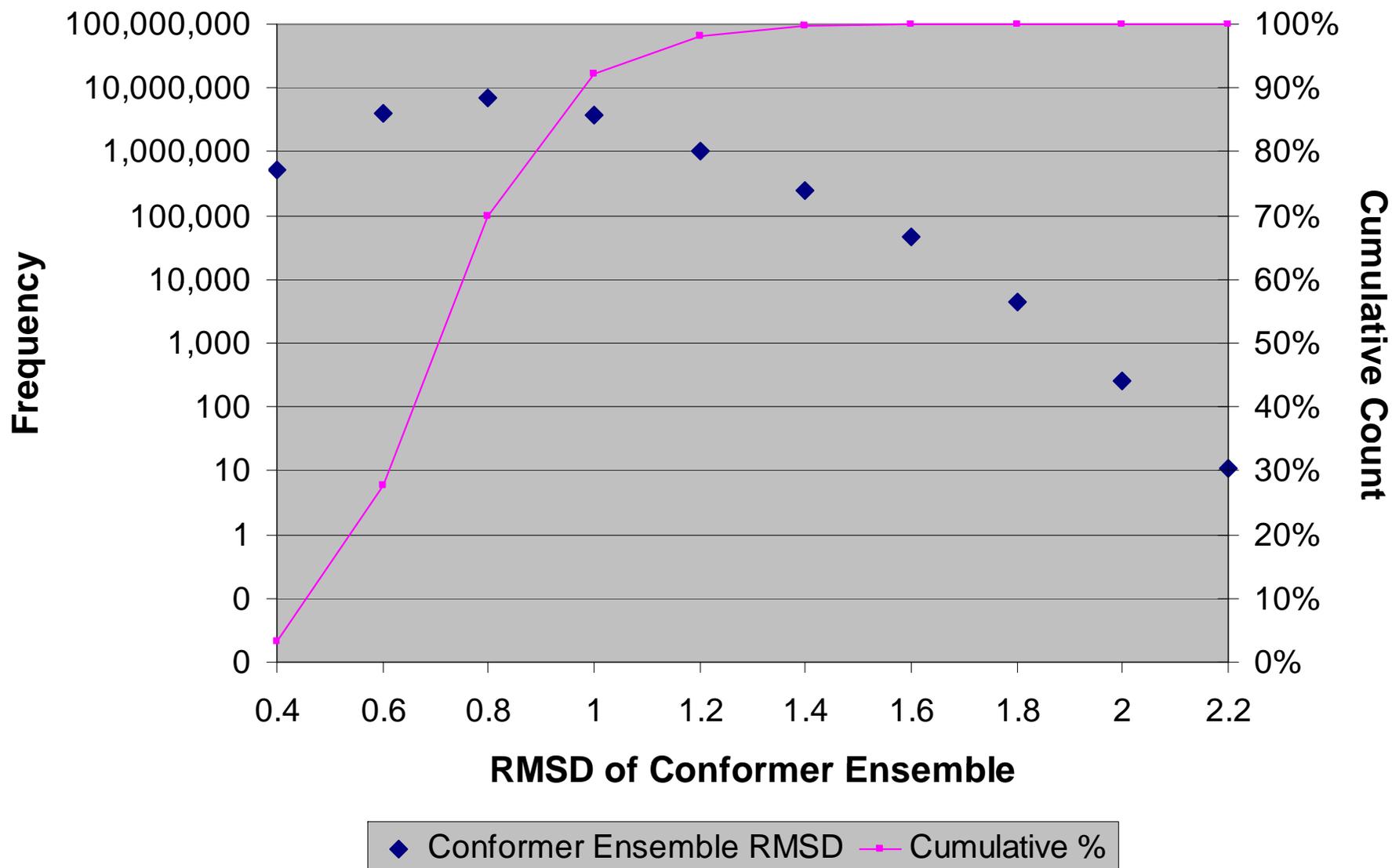
PubChem3D Compound Heavy Atom Count



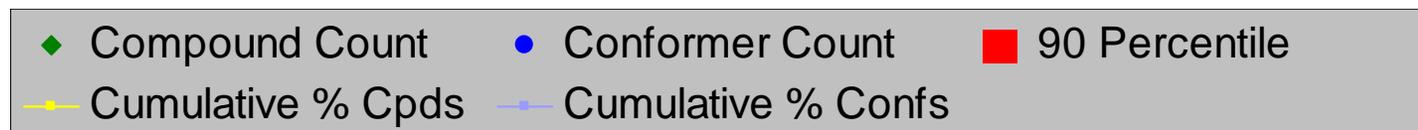
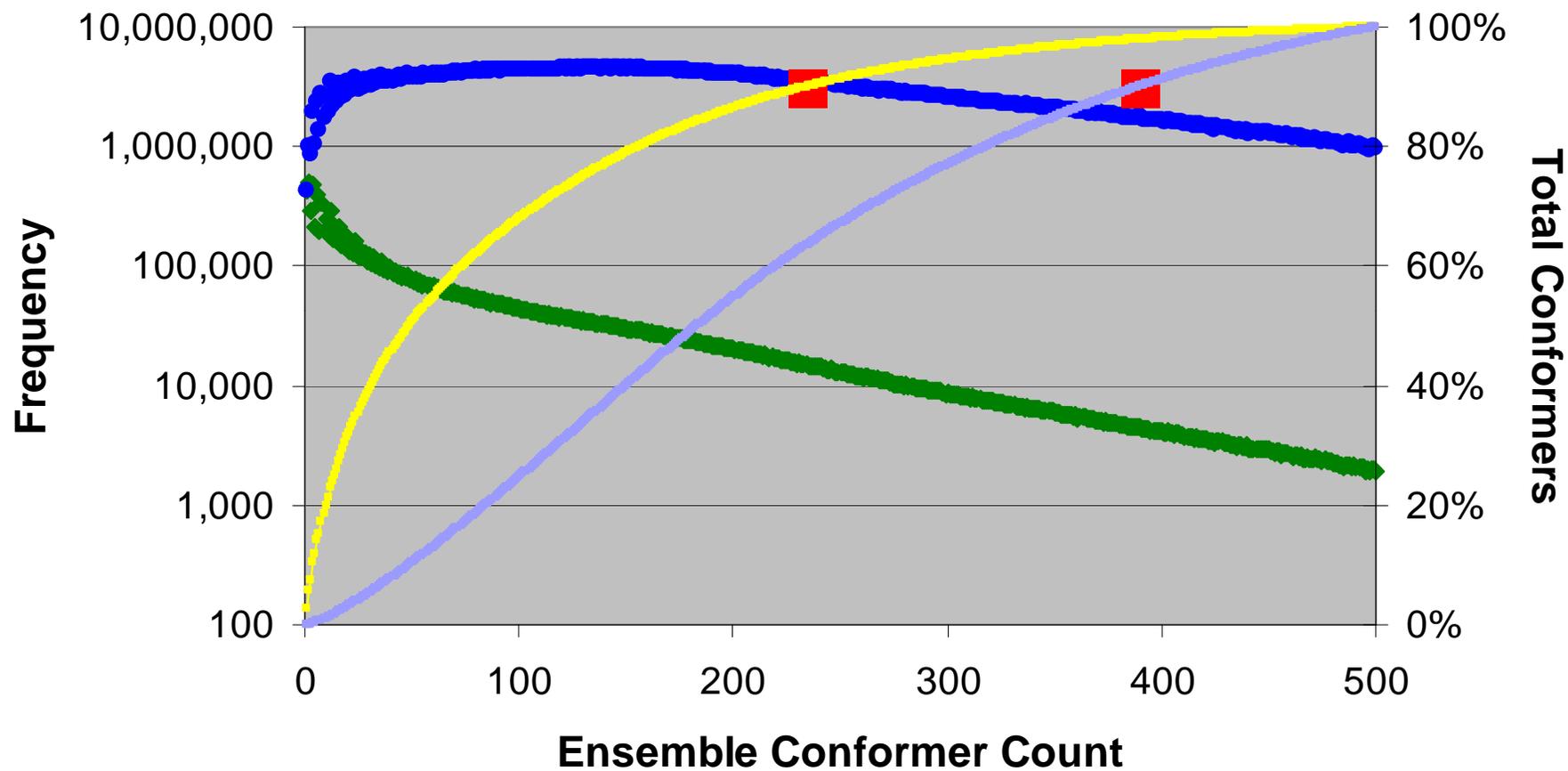
PubChem3D Compound Undefined Stereo



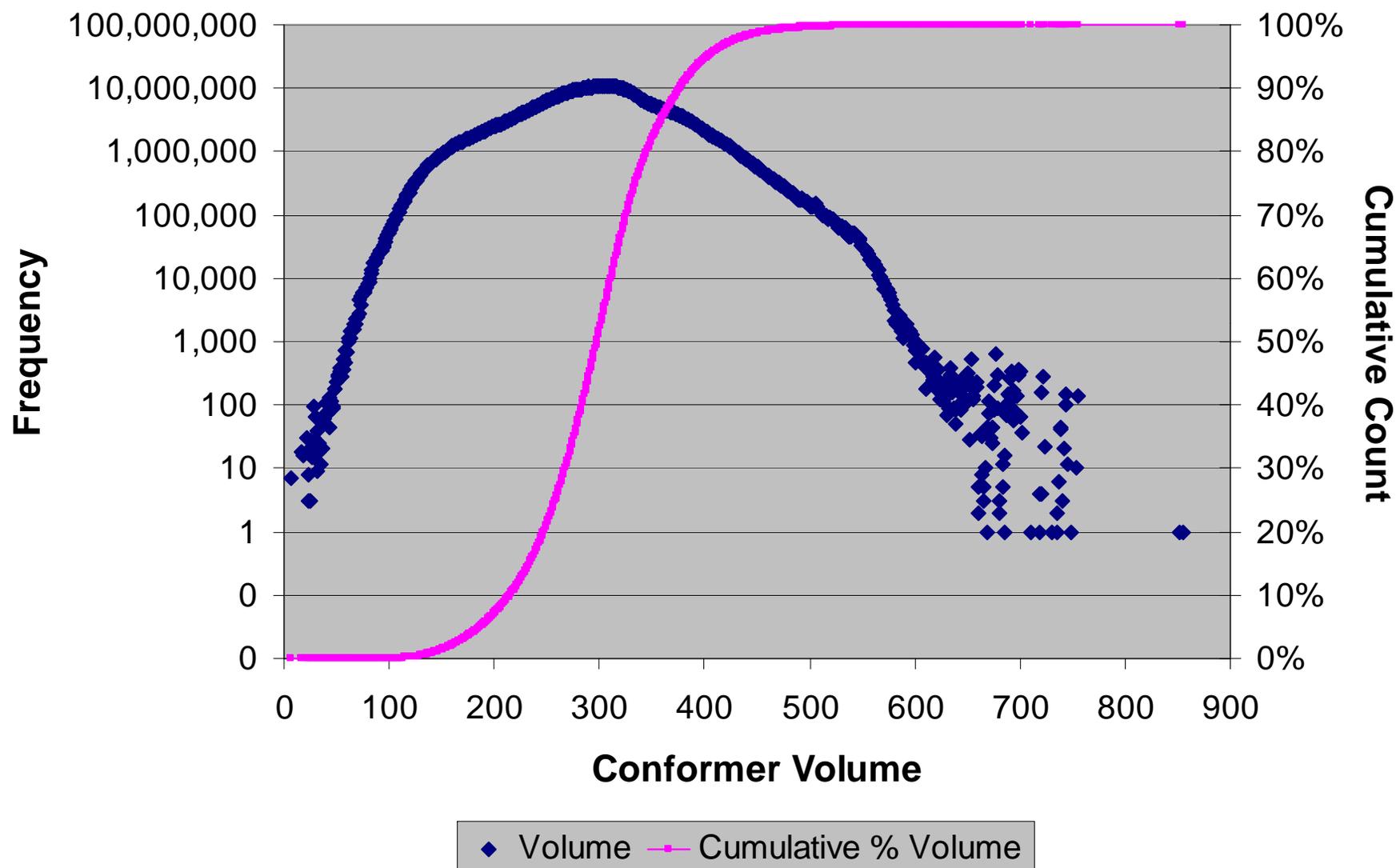
PubChem3D Conformer Ensemble RMSD



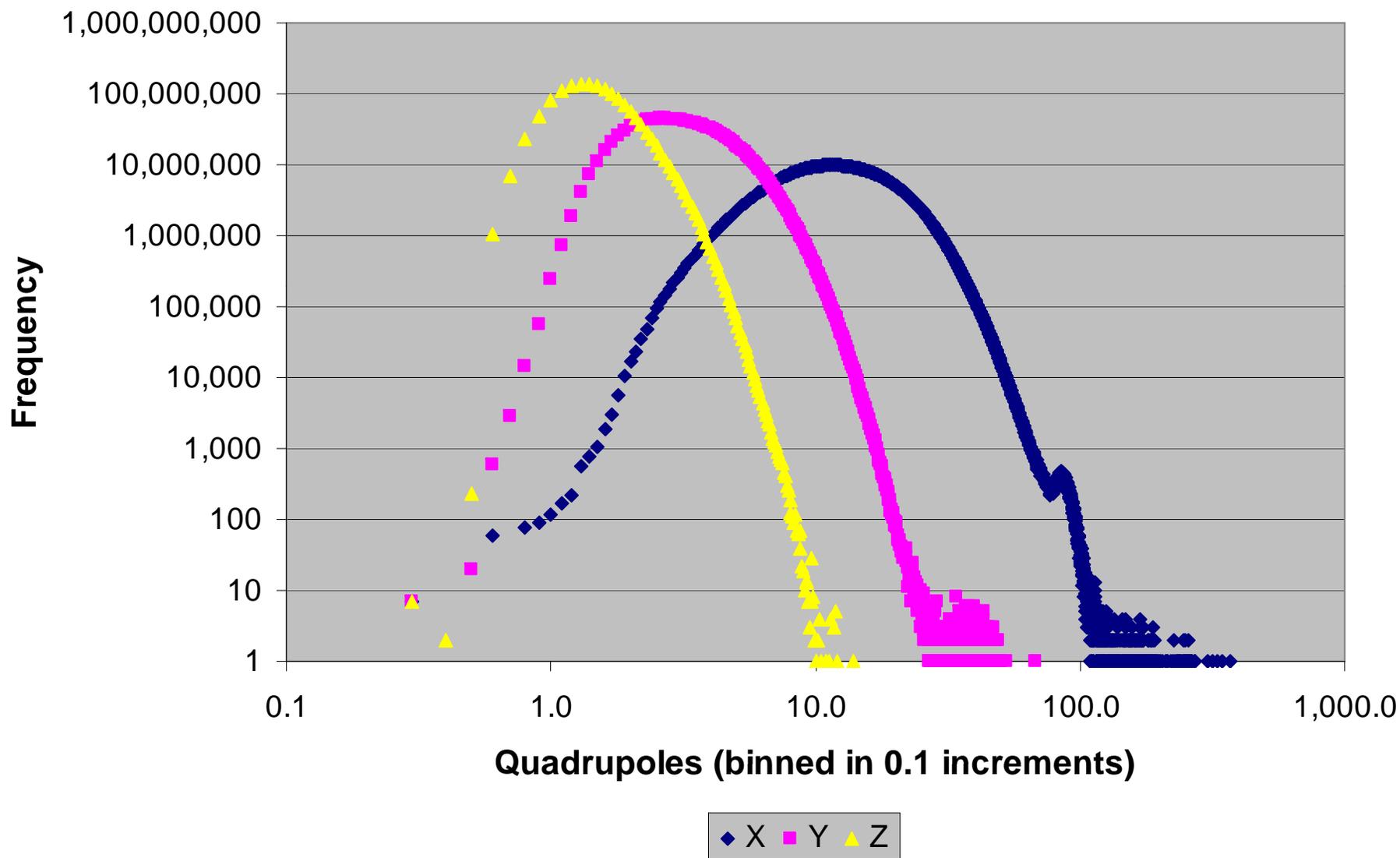
PubChem3D Conformer Ensemble Size



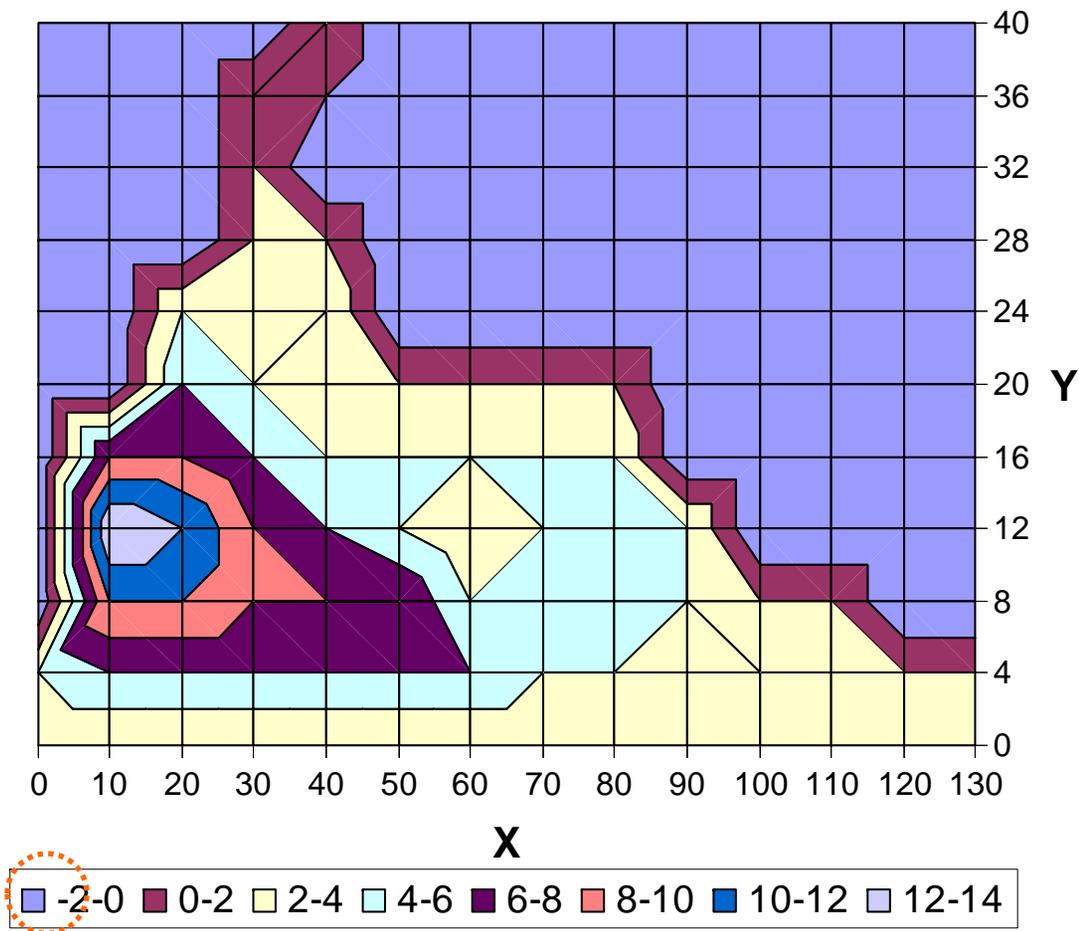
PubChem3D Conformer Volume



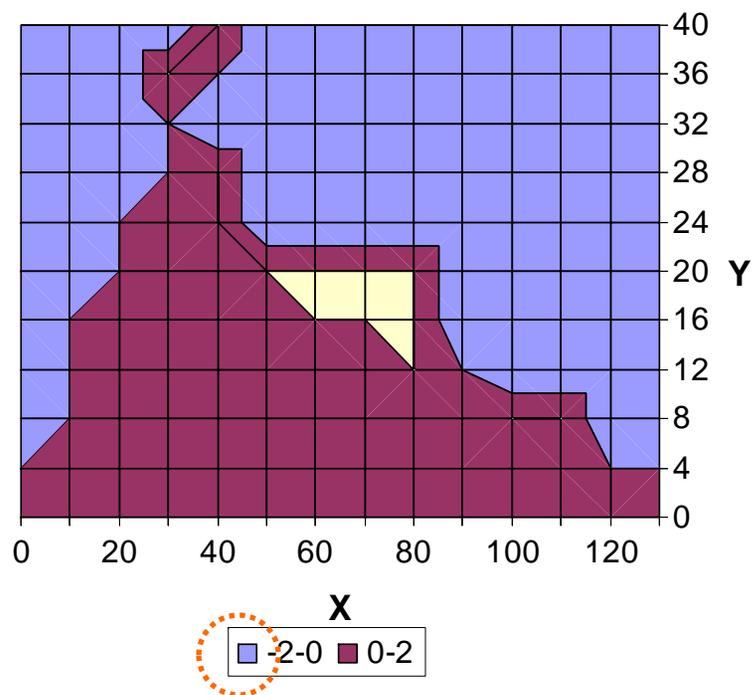
PubChem3D Conformer Quadrupoles



PubChem3D Quadrupole Maximums



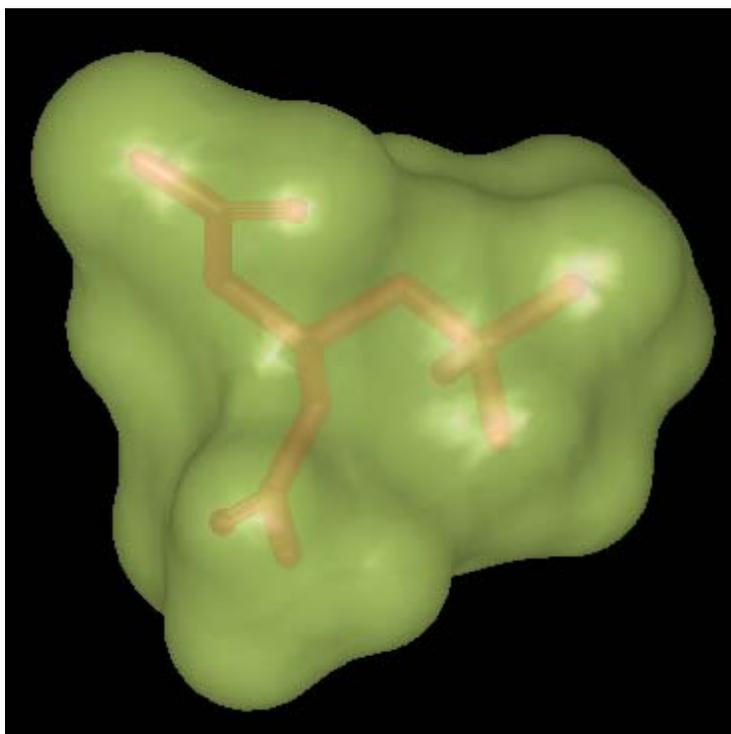
PubChem3D Quadrupole Minimums



PubChem3D Data “Dissemination”

- FTP download
 - ftp.ncbi.nlm.nih.gov/pubchem/Compound_3D
 - Single conformer per compound (only)
 - MMFF94 charges
 - MMFF94s Energy (No_Estat)
 - Steric: Volume, Quadrupoles, Octopoles
 - Effective Rotor Count
 - RMSD of conformer model

Shape Description of



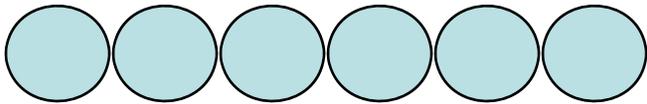
How do we “do” shape?

- `OEChem::OESuppressHydrogens`
- `OEChem::OEAssignBondiVdWRadii`

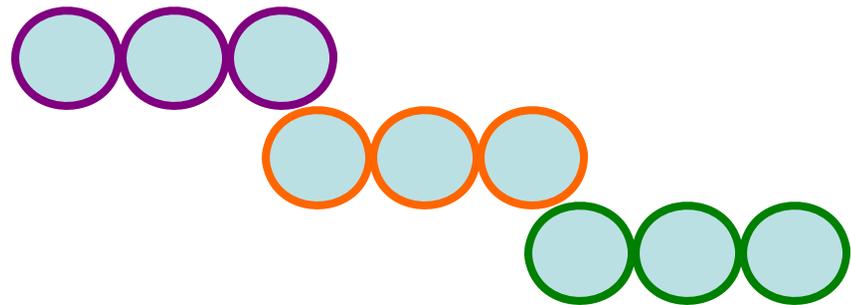
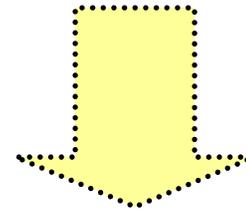
- `OEShape::OEOverlapMethod::Analytic`
- `OEShape::OEOverlapRadii::All`

Cluster vs. PreCluster

“Clustered”



“PreClustered”



How to cluster billions of conformers?

- Cluster by unique volume
 - Precluster conformers to achieve a set of <6,000 conformers (**Basis shapes**)
 - Cluster basis shapes (**Reference shapes**)
 - While count of ref shapes >128, recluster
 - Recluster (ref + basis shapes) to fill holes
- For all 706 unique volumes and ~1.5b confs:
 - 77,409 reference shapes
 - 1,376,218 basis shapes

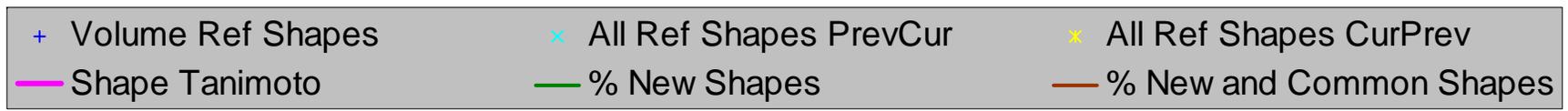
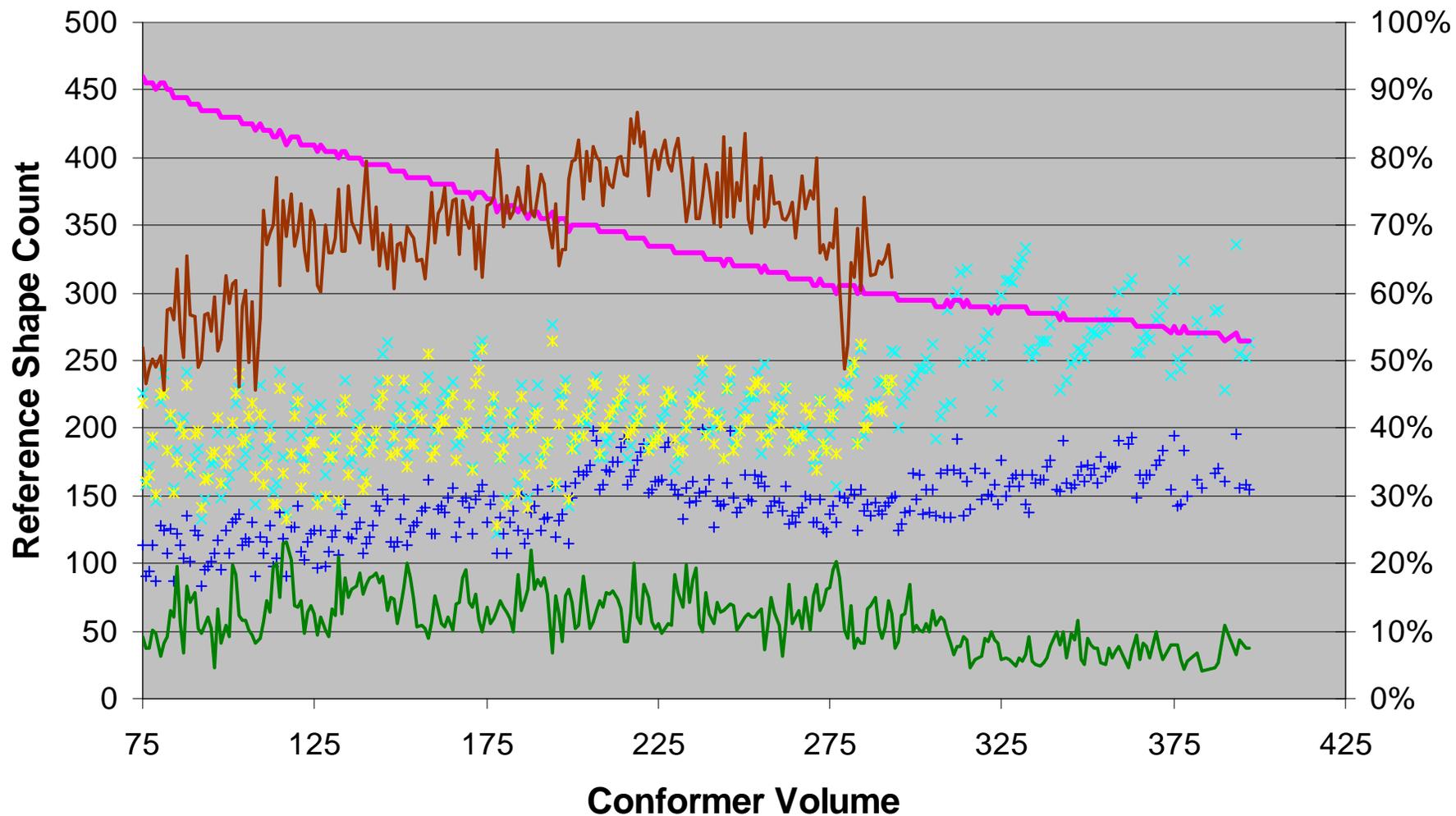
How unique is shape space?

- For volume < current volume
 1. Pool all reference shapes
 2. Precluster at current volume ST
 3. Cluster unique at current volume ST
 4. Pool all basis shapes
 5. Precluster basis shapes at current volume ST
 6. Cluster unique and preclustered basis shapes
 7. Recluster with current volume ref shapes
 8. Fill cluster holes with current vol basis shapes
 9. Count # refs from current vol ref/basis shapes
- Repeat... but use current volume shapes first

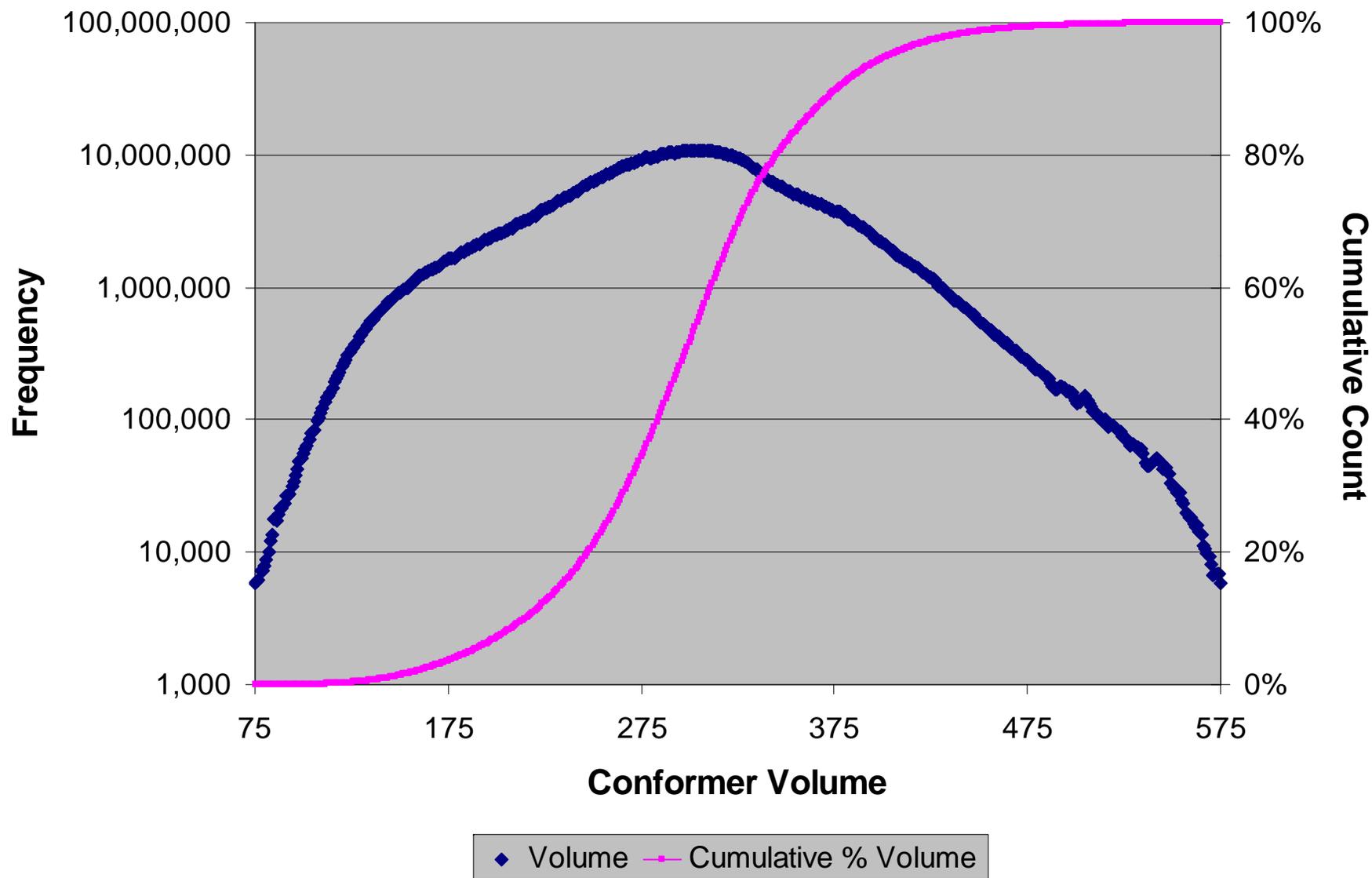
Example: Volume=200

- For volume < current volume (1-199 < 200)
 - Pool all reference shapes (16,739 shapes)
 - PreCluster at current volume ST (735 refs at 70%)
 - Cluster at current volume ST (131 refs at 70%)
 - Pool all basis shapes (251,809 shapes)
 - Precluster basis shapes at current vol ST (1,725 shapes)
 - Cluster unique and preclustered basis shapes (167 refs at 70%)
 - Recluster with current volume ref shapes (170 refs at 70%)
 - Fill cluster holes with current vol basis shapes (191 refs at 70%)
 - Count # refs from current vol ref/basis shapes (27 unique refs)

Unique Shape Expansion by Volume and ST



PubChem3D Conformer Volume Distribution



Conclusions

- Shape space of small molecules can be reduced to a small representative set
- Each increase in conformer volume contributes ~12% more shape diversity
- Each unique volume contains ~70% of total shape diversity within all lesser volumes

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

