PubChem, 3-D?, Omega, “RMSD and Shape”

Evan Bolton, Ph.D.

National Center for Biotechnology Information

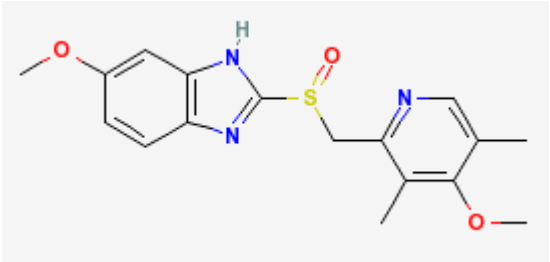
CUP VIII 2007

bolton@ncbi.nlm.nih.gov

Why **3-D**?

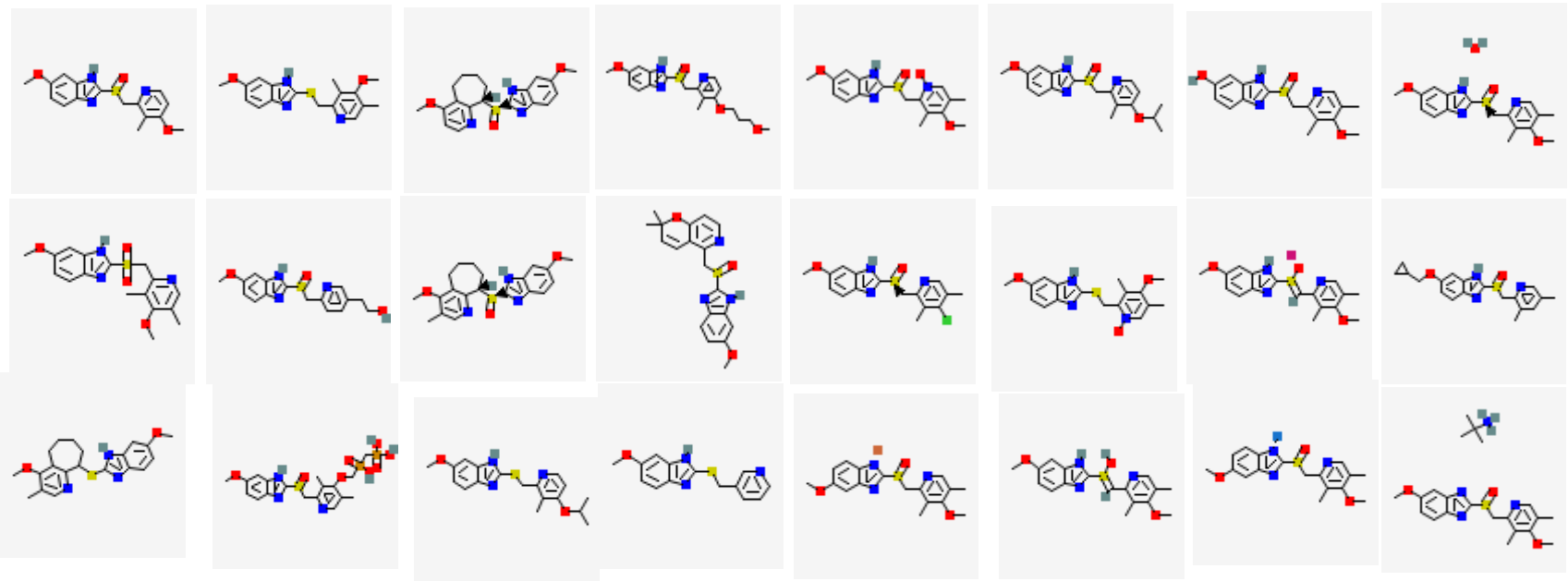
Isn’t **2-D** enough?

Problem with **2-D** Similarity…

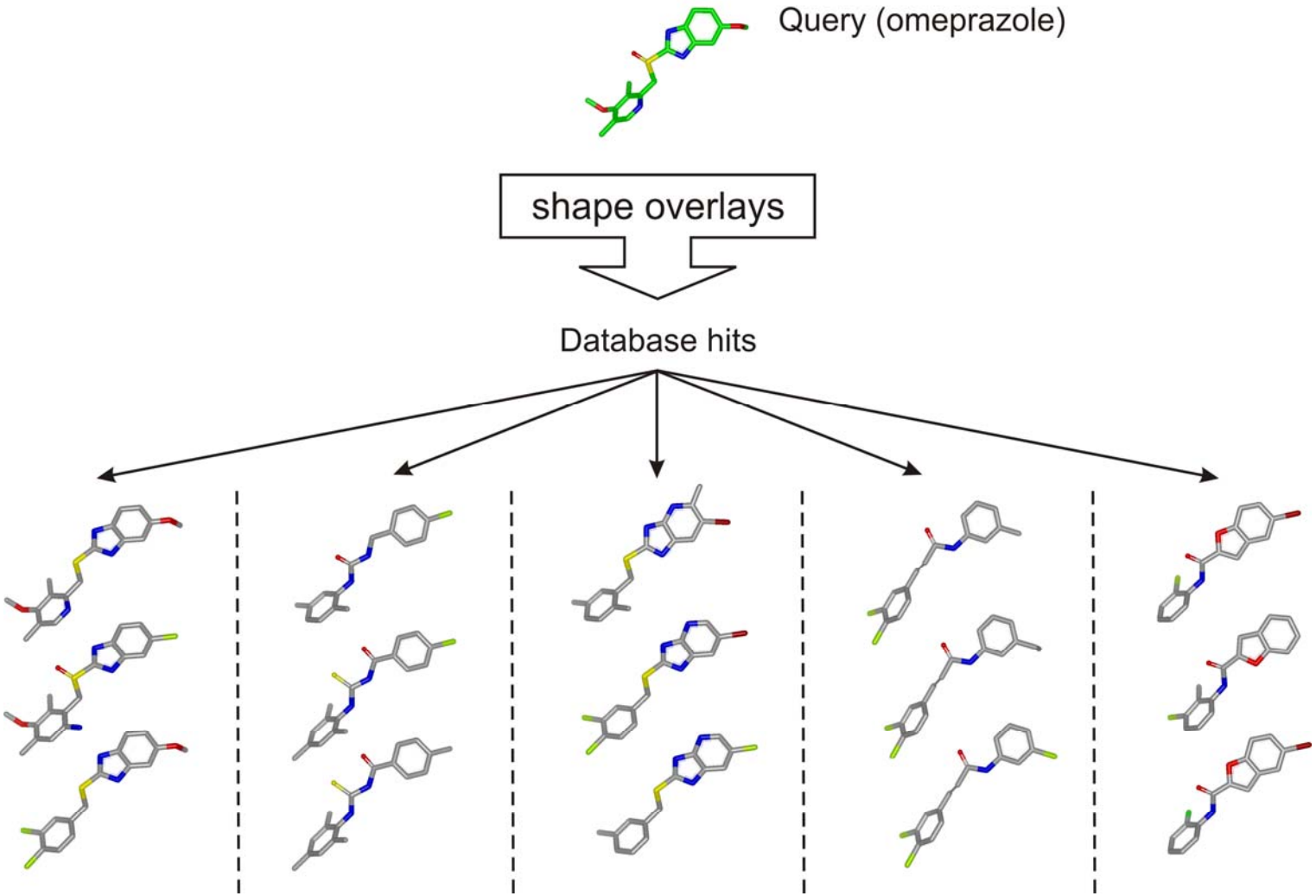


… is a lack of variety

just similar “analogs” …



3-D Similarity can be different from 2-D



Similarity of a different kind …

I have a biological screening hit or lead…

How do I find something <u>not tested</u> that will provide a similar biological result?

Search a database …

…**2-D** Similarity (Analogs)

…**3-D** Feature Similarity (Pharmacophore)

…**3-D** Shape Similarity (Fits Binding Pocket)

…**3-D** Electrostatic Similarity (Binds Same Way)

…

Why do 3-D?

- 2-D 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 …

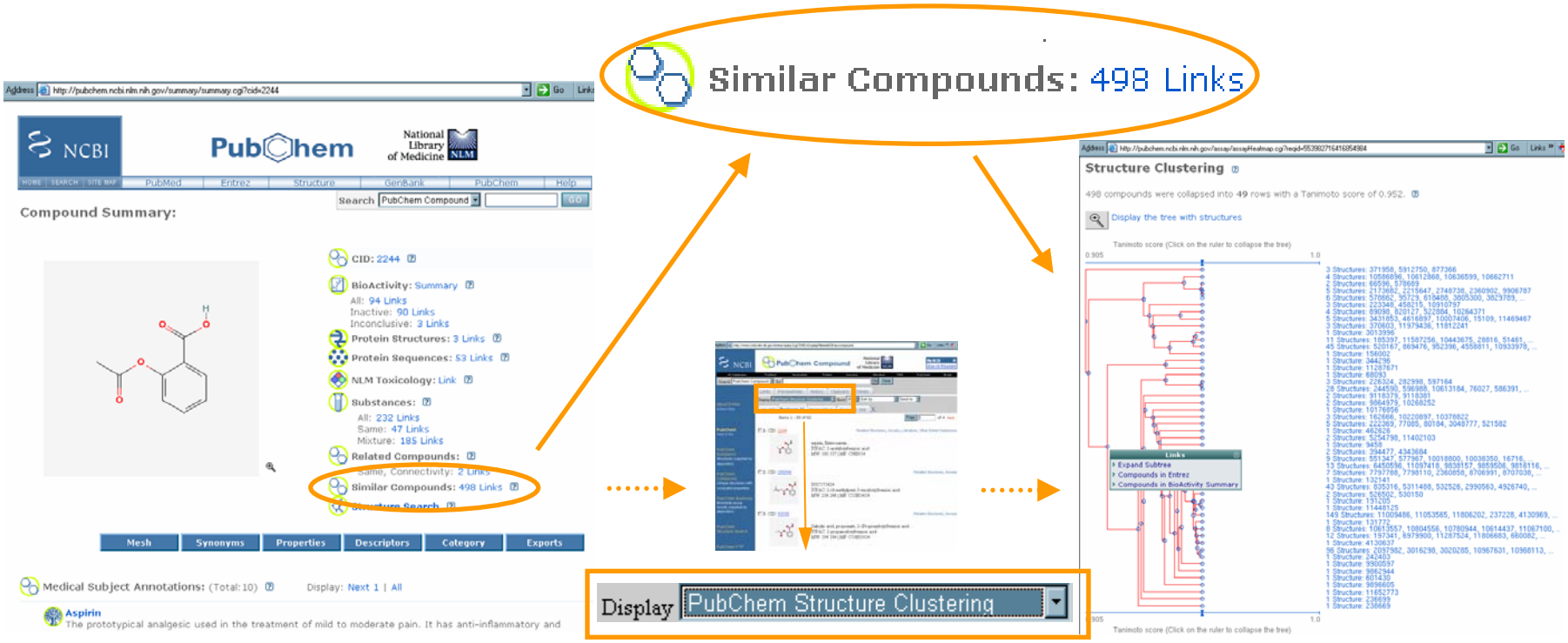
  … better integrate with other NCBI projects/data

  … consider advanced Virtual Screening methods

  … etc.

What might we consider doing with 3-D?

- Generate conformer ensemble (when possible)
- (Pre)Compute Shape Neighbors



"Effective Rotors" Concept

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

R^2 = 0.92 and RSE = 0.57, with N = 6 766

where:

- ***RMSD*** is the desired root-mean-square-deviation
- ***NC*** is the number of conformers estimated at ***RMSD***

- ***$RMSD_0$*** is 2.0 Å
- ***NC_0*** 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" SP_3-hybridized ring atoms

“Molecular Flexibility” Measure

“Effective Rotors” … takes into account ring flexibility

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

- ***nr*** is the number of rotatable bonds
- ***nnara*** is the number of “non-aromatic” SP_3-hybridized ring atoms

How to do Conformer Generation?

- How “should” Omega be (ab)used?
- How “accurate” is Omega?
- What are Omega’s “limitations”?

Omega Conformer Generation is a Two Stage Process

- Fragment Library

Choose…

… Sampling Rate
… Force-Field
… Energy Window

- Generate Conformers

Choose…

… Fragment Library
… Force-Field
… Energy Window

Omega 2.1 Parameter Variation

- Fragment Sampling (-startfact)
  … 20 (Frag Default), 50
- Force Field (MMFF94s)
  … Full, NoEstat (Frag/Conf Default), Trunc
- Energy Window
  … 1-7, 8 (Frag Default), 9, 10, 15, 20, 25 (Conf Default), 30
- Non-varied, non-default Parameters
  - MaxConfsGen = 250,000 (Default: 30,000)
  - MaxConfs = **100,000** (Default: 20,000)
  - MaxRotors = 15 (Default: 12)
  - MaxSearchTime = 180 (Default: 30)
  - DefaultRMSD = 0.0 (Default: 0.8)

Largest possible value…

Omega 2.1 Validation Study

- Data…
  - 3-D Ligand Structures
  - NCBI's MMDB
    - … only experimental data from PDB
    - … ligand bond orders "perceived" (OEChem 1.3.4)
  - Limited to…
    - … 2 - 50 Non-Hydrogen Atoms
    - … 0 - 15 Rotatable Bonds (OEChem 1.4.2)

25,972 non-unique organic **3-D** experimental reference structures

Analysis

- "Molecular Flexibility"

  … count of effective rotors (binned by whole numbers)
- "Molecular Size"

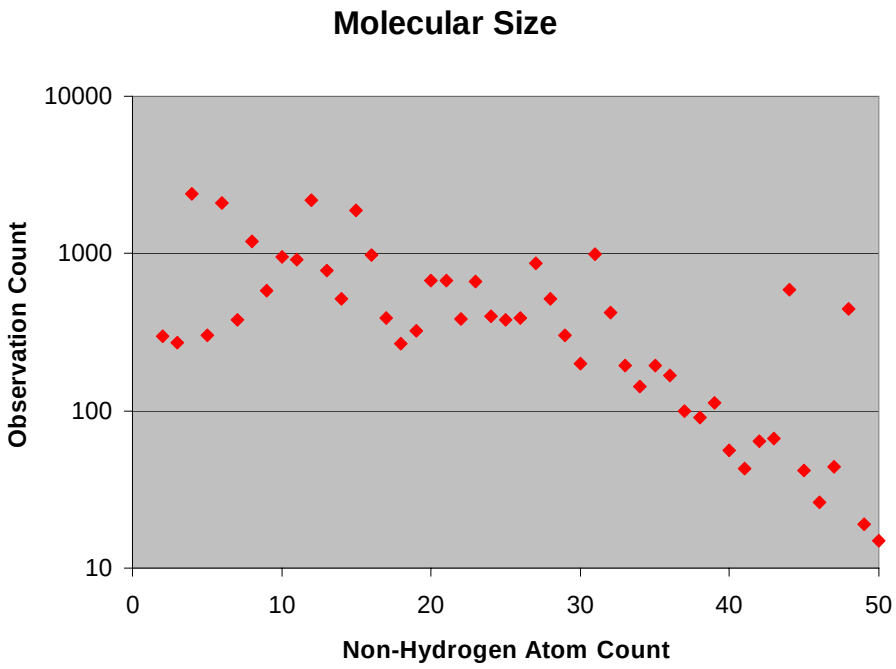
  … count of non-hydrogen atoms
- "Accuracy"

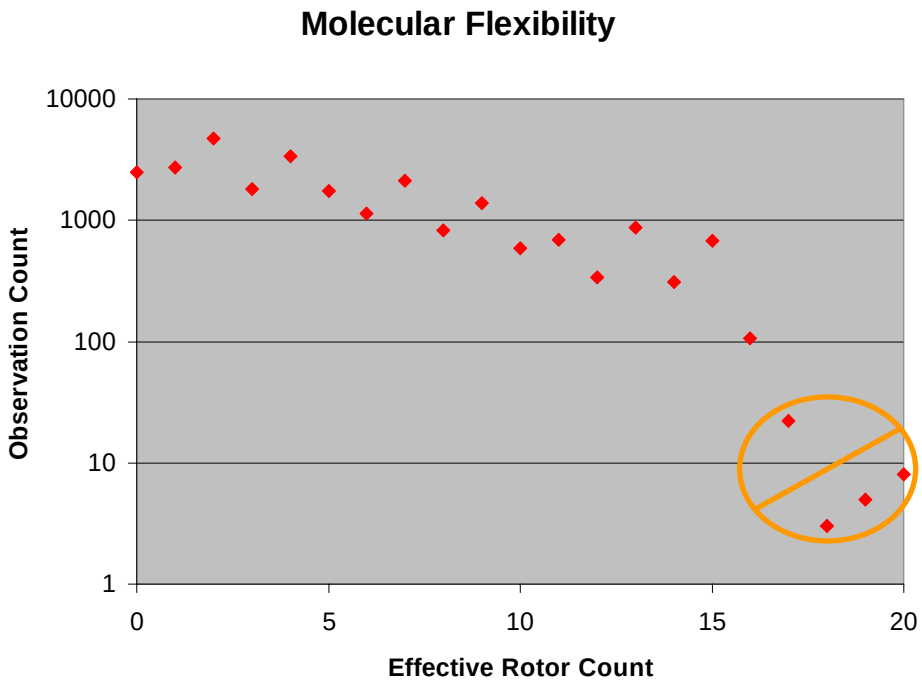
  … RMSD and ST …

  minimum root mean square distance (OEChem 1.4.2) and maximum shape tanimoto (OEShape 1.6.1) from "experimental reference" conformation to a conformation produced by Omega 2.1, respectively
- "Limitation"

  … count of structures with 100k conformations

Data Set Distribution





Observation count for all 25,972 3-D reference structures as a function of count type

Omega Conformer Generation is a Two Stage Process

- Fragment Library

Choose…

<u>… Sampling Rate</u>

… Force-Field

… Energy Window

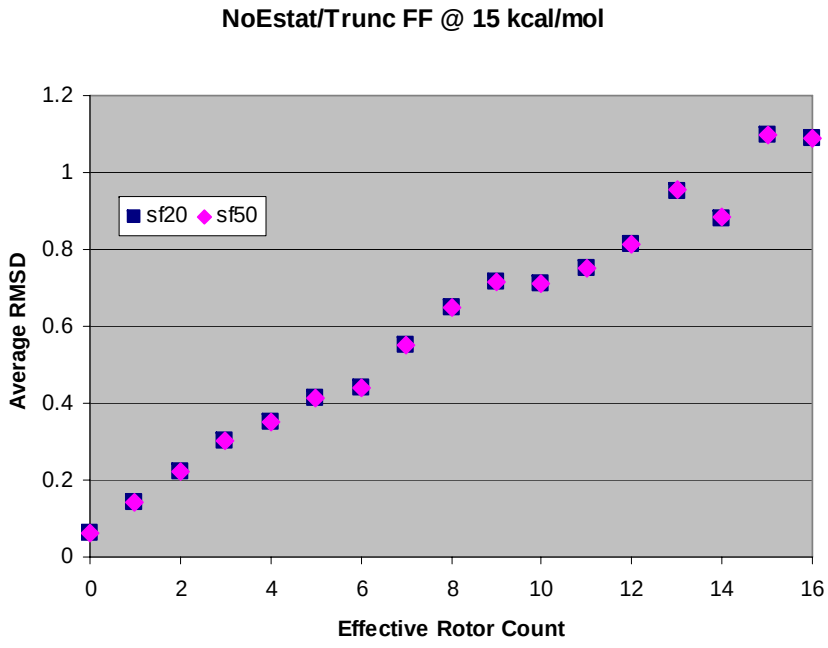
- Generate Conformers

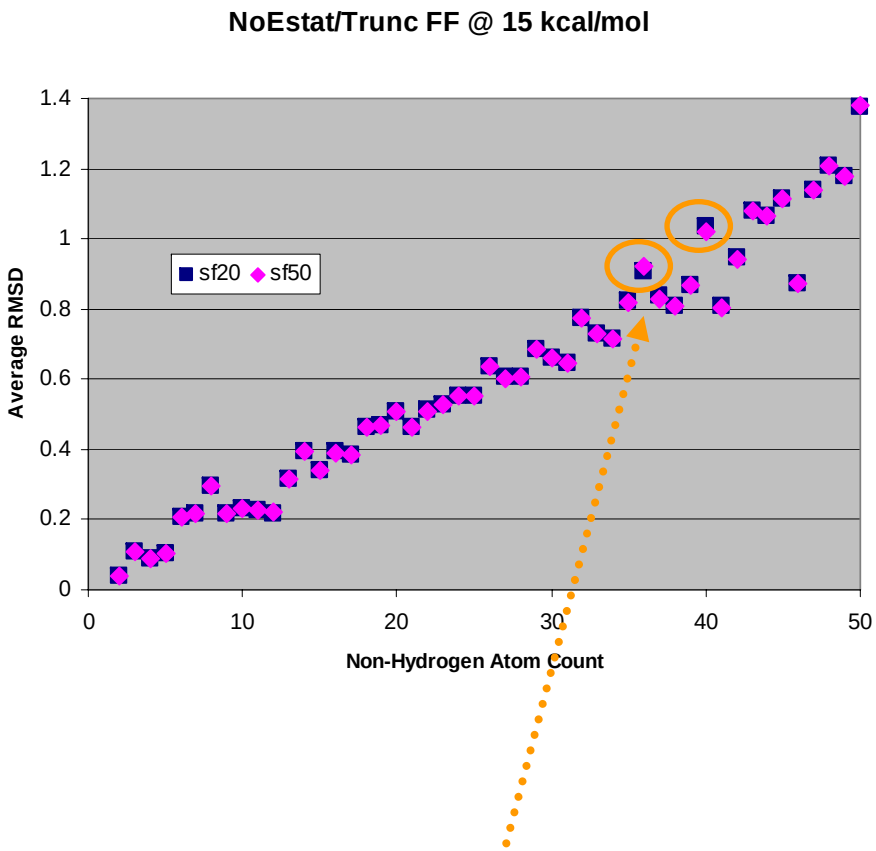
Choose…

… Force-Field

… Energy Window

Fragment Sampling Affect?





Statistically insignificant differences can be found

Average RMSD values for all 25,972 3-D reference structures as a function of count type

Default Fragment Sampling is sufficient ...

Energy Window	5/5 (Fragment/Conformer)																	
Conformer FF	Full						NoEstat						Trunc					
Fragment FF	Full		NoEstat		Trunc		Full		NoEstat		Trunc		Full		NoEstat		Trunc	
Sampling	20	50	20	50	20	50	20	50	20	50	20	50	20	50	20	50	20	50
RMSD	0.69	0.69	0.69	0.69	0.71	0.71	0.48	0.48	0.48	0.48	0.48	0.48	0.467	0.466	0.467	0.467	0.462	0.462
Shape Tanimoto	0.89	0.89	0.89	0.89	0.89	0.89	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Conformers	698	696	698	697	645	646	2827	2834	2814	2818	2870	2875	5369	5377	5352	5358	5852	5857
100k Limit Count	23	21	18	18	36	36	94	96	93	96	118	119	216	217	211	211	239	239

Energy Window	10/10		5/15				10/15		15/15	
Conformer FF	NoEstat		Trunc				NoEstat		NoEstat	
Fragment FF	Trunc		Full		Trunc		Trunc		Trunc	
Sampling	20	50	20	50	20	50	20	50	20	50
RMSD	0.42	0.42	0.41	0.41	0.41	0.41	0.40	0.40	0.42	0.42
Shape Tanimoto	0.94	0.94	0.95	0.95	0.95	0.95	0.943	0.946	0.94	0.94
Conformers	10794	10796	12957	12981	13012	13025	13118	13124	10794	10796
100k Limit Count	1842	1841	2203	2203	2209	2210	2198	2196	1842	1841

Statistically insignificant differences can be found

Overall count or average values for all 25,972 3-D reference structures as a function of fragment sampling

Omega Conformer Generation is a Two Stage Process

- Fragment Library

Choose…

… Sampling Rate [Use default]

<u>… Force-Field</u>

… Energy Window

- Generate Conformers

Choose…

… Force-Field

… Energy Window

Fragment Force Field Affect?

5 kcal/mol	Conformer Generation Force Field		
Fragment FF	Full	NoEstat	Trunc
Full	0.69 / 0.89	0.48 / 0.93	0.47 / 0.93
NoEstat	0.69 / 0.89	0.48 / 0.93	0.47 / 0.93
Trunc	0.71 / 0.89	0.48 / 0.93	0.46 / 0.93

10 kcal/mol	Conformer Generation Force Field		
Fragment FF	Full	NoEstat	Trunc
Full	0.52 / 0.92	0.42 / 0.94	0.42 / 0.94
NoEstat	0.52 / 0.92	0.42 / 0.94	0.42 / 0.94
Trunc	0.52 / 0.92	0.41 / 0.94	0.42 / 0.94

15 kcal/mol	Conformer Generation Force Field		
Fragment FF	Full	NoEstat	Trunc
Full	0.46 / 0.93	0.40 / 0.94	0.40 / 0.95
NoEstat	0.46 / 0.93	0.40 / 0.94	0.40 / 0.95
Trunc	0.46 / 0.93	0.40 / 0.94	0.40 / 0.95

Fragment library force field type has negligible impact on accuracy …

Minor variations that disappear as energy is increased …

Overall average RMSD / ST as a function of force field combination and energy window for all 25,972 reference structures

Omega Conformer Generation is a Two Stage Process

- Fragment Library

Choose…

… Sampling Rate [Use default]

… Force-Field [Doesn't matter]

… Energy Window

- Generate Conformers

Choose…

… Force-Field

… Energy Window

Conformer Generation Force Field

Energy Window (kcal/mol)	Fragment/Conformer Force Field		
	Full	NoEstat	Trunc
5	0.69 / 0.89	0.48 / 0.93	0.46 / 0.93
10	0.52 / 0.92	0.42 / 0.94	0.42 / 0.94
15	0.46 / 0.93	0.40 / 0.943	0.40 / 0.946

RMSD / ST

No significant difference between NoEstat and Trunc at higher energy values

Energy Window (kcal/mol)	Fragment/Conformer Force Field		
	Full	NoEstat	Trunc
5	698	2 818	5 857
10	2 817	10 852	10 889
15	5 599	13 219	13 178

Conformer Count

Larger conformer count gives better accuracy …

Energy Window (kcal/mol)	Fragment/Conformer Force Field		
	Full	NoEstat	Trunc
5	21	96	239
10	250	1 788	1 844
15	491	2 189	2 220

100k Limit Count

… but 100k limit is reached more often

Trunc/5 is as good or better than Full/15

Overall count or average values as a function of force field and energy for all 25,972 3-D structures

Omega Conformer Generation is a Two Stage Process

- Fragment Library

Choose…

… Sampling Rate [Use default]

… Force-Field [Doesn't matter]

… Energy Window

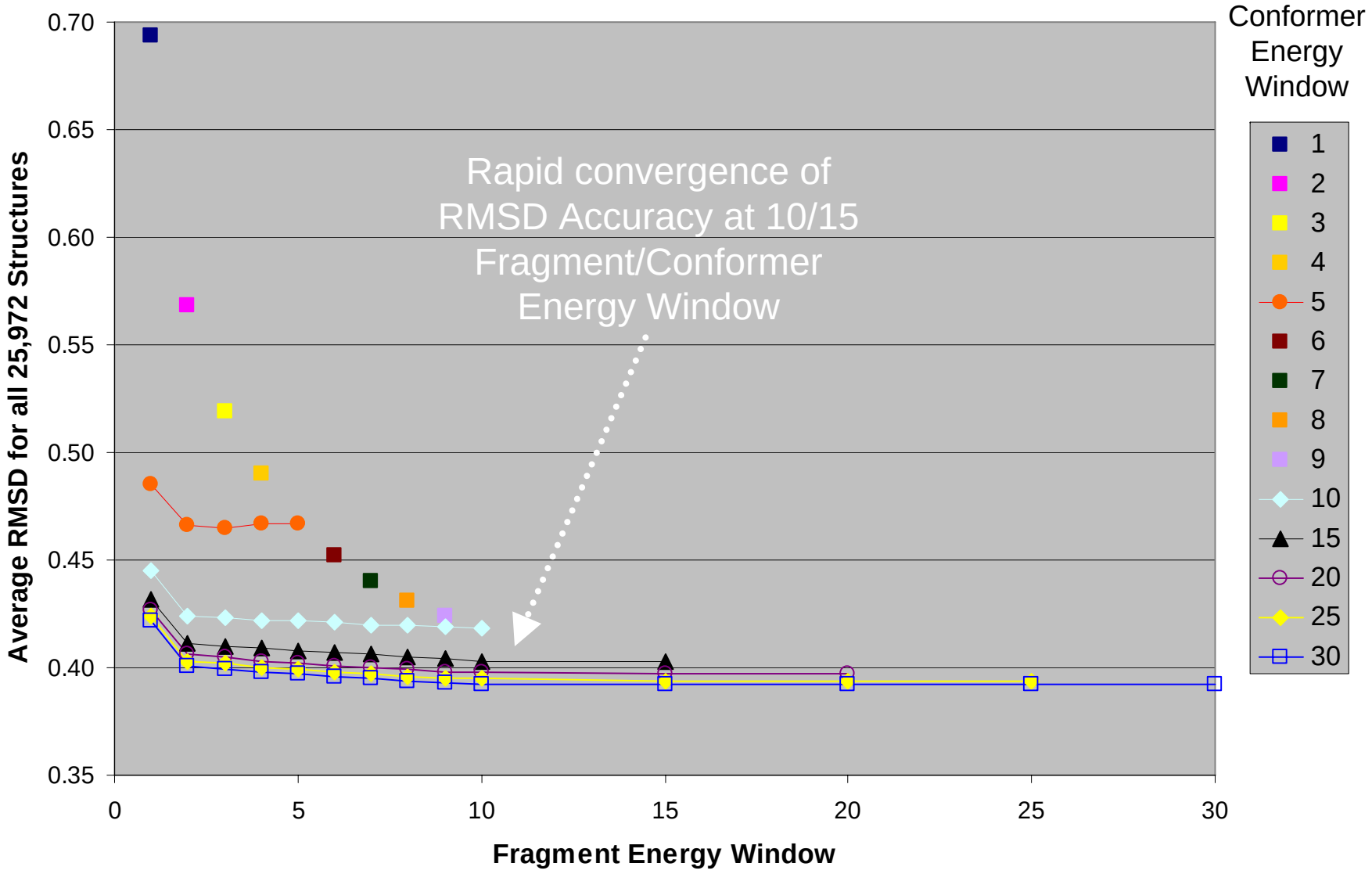
- Generate Conformers

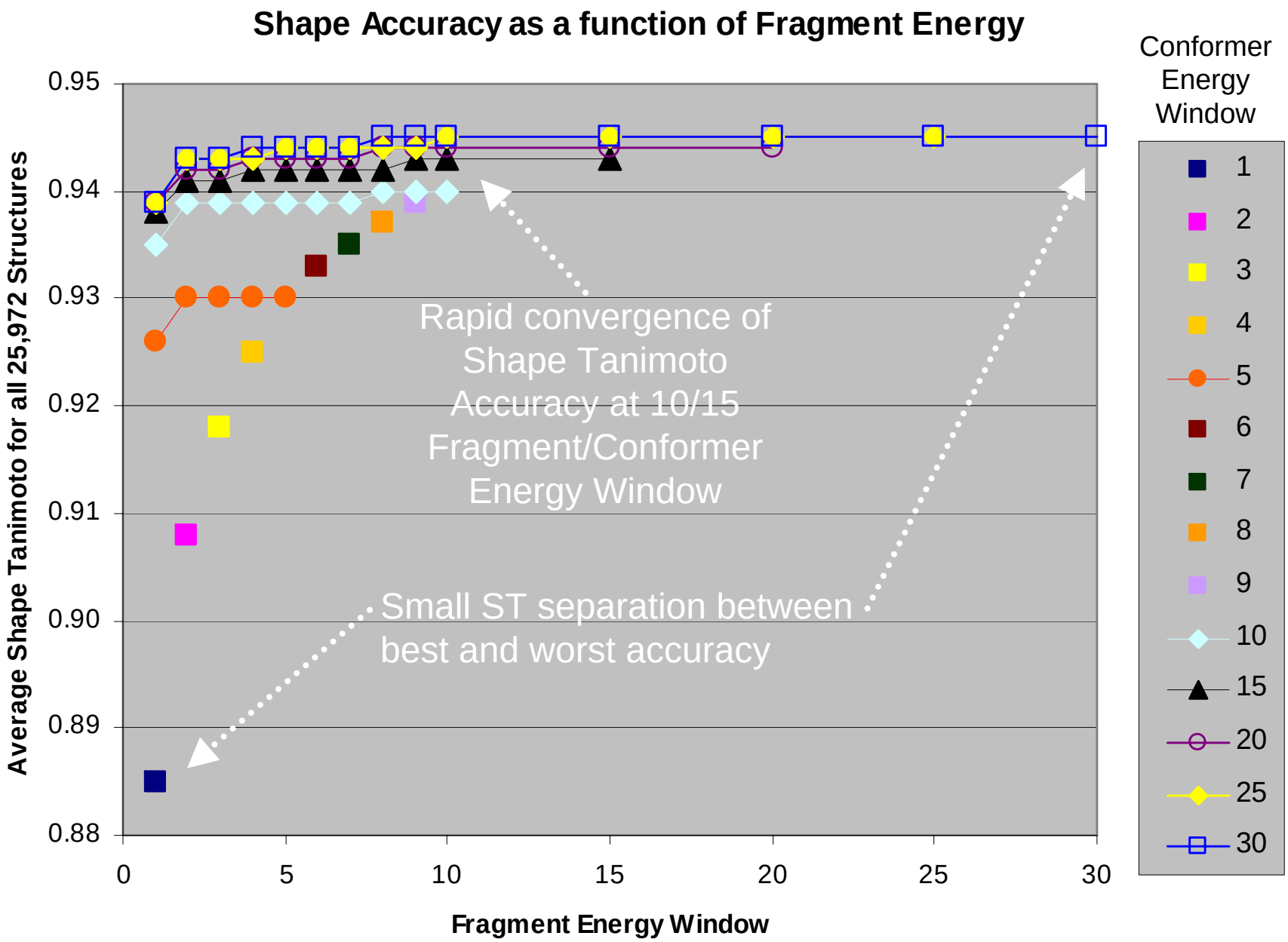
Choose…

… Force-Field [Use NoEstat or Trunc]

… Energy Window

Accuracy as a function of Energy




Shape Accuracy as a function of Fragment Energy
Average Shape Tanimoto for all 25,972 Structures
0.95
0.94
0.93
0.92
0.91
0.90
0.89
0.88
0
5
10
15
20
25
30
Fragment Energy Window
Rapid convergence of Shape Tanimoto Accuracy at 10/15 Fragment/Conformer Energy Window
Small ST separation between best and worst accuracy
Conformer Energy Window
1
2
3
4
5
6
7
8
9
10
15
20
25
30

Omega Conformer Generation is a Two Stage Process

- Fragment Library

Choose…

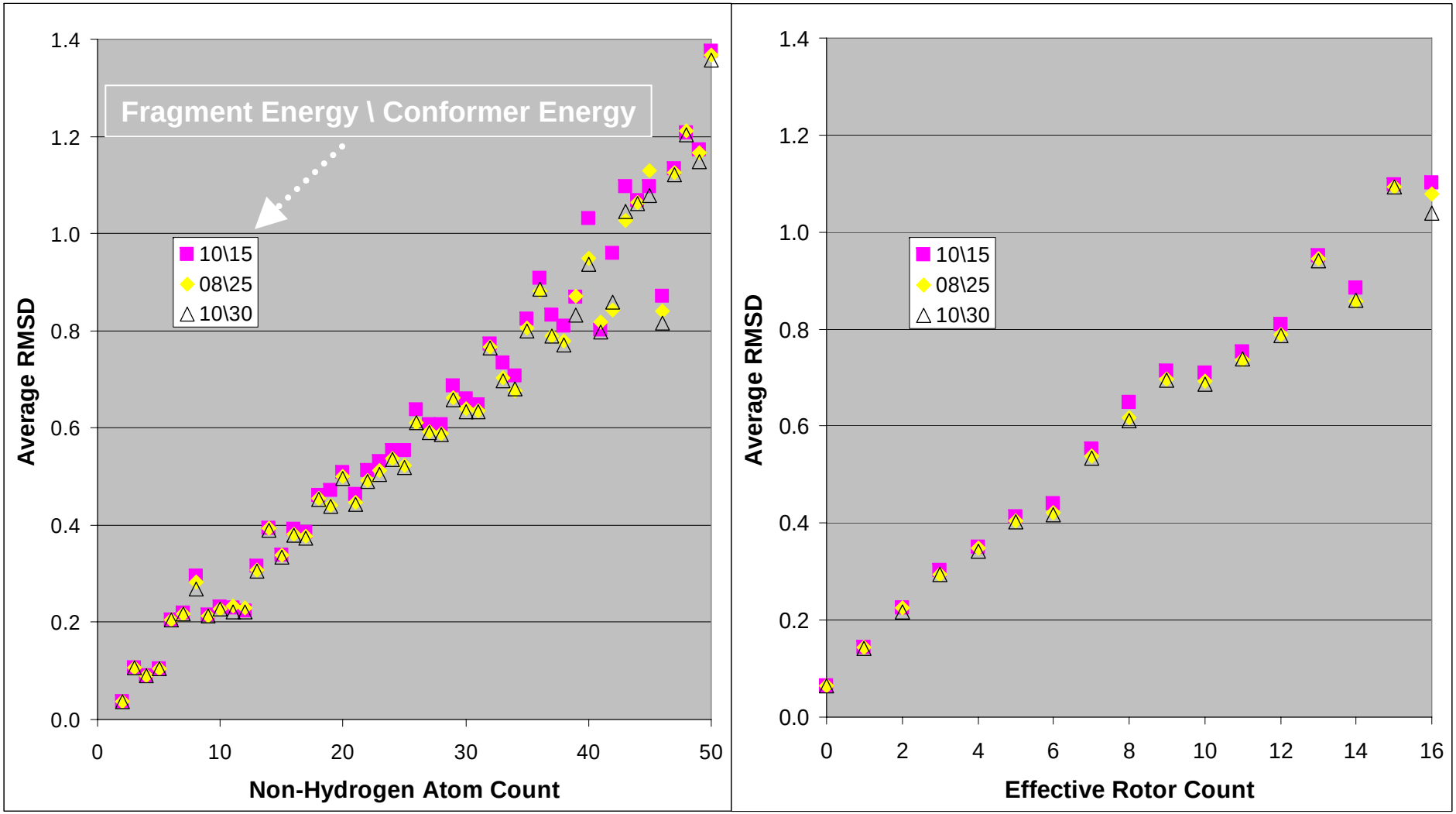
… Sampling Rate [Use default]
… Force-Field [Doesn't matter]
… Energy Window [Use 10 kcal/mol]

- Generate Conformers

Choose…

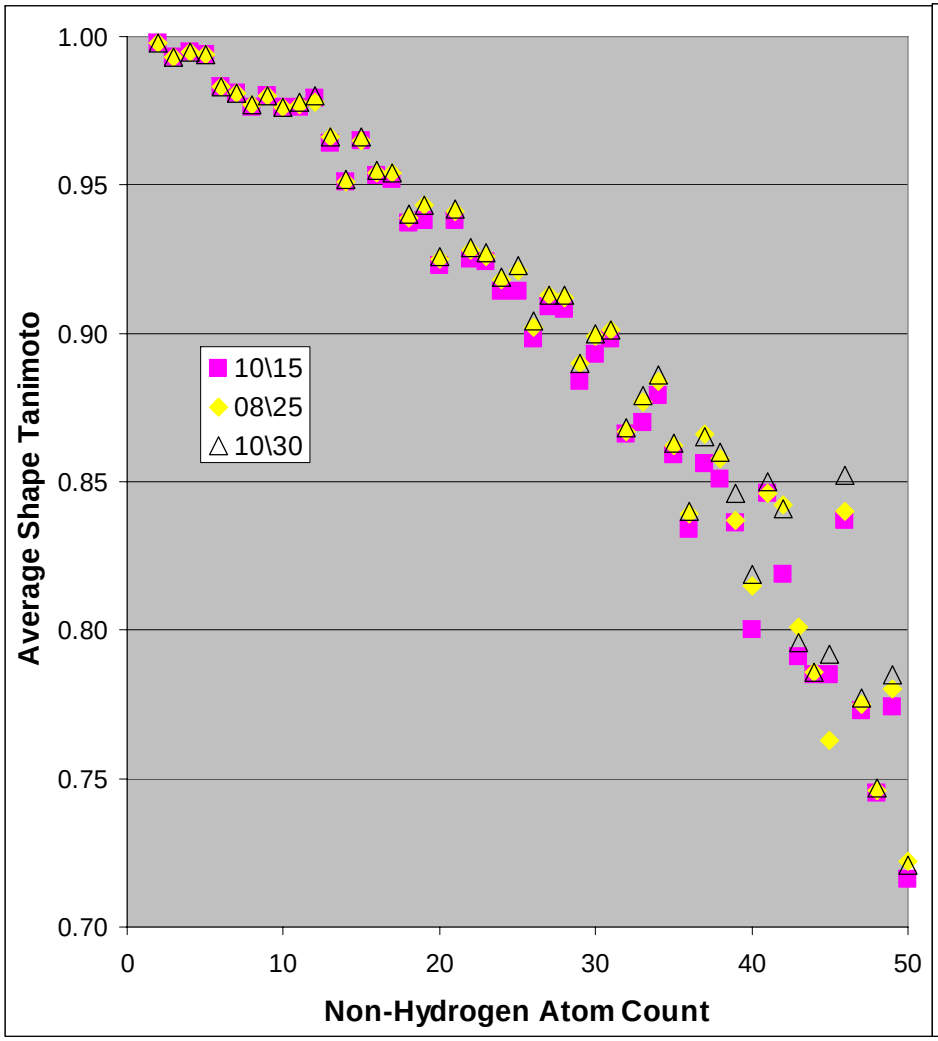
… Force-Field [Use NoEstat or Trunc]
… Energy Window [Use 15 kcal/mol]

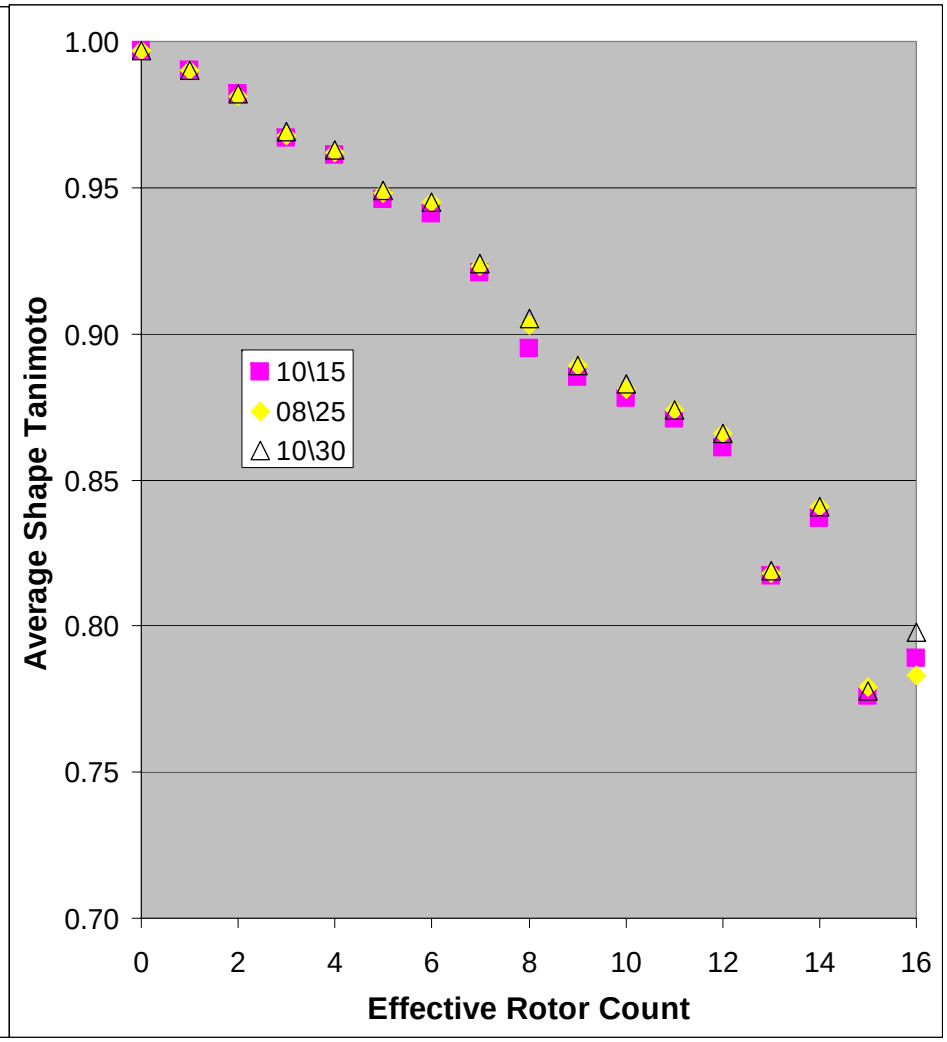
RMSD Accuracy



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

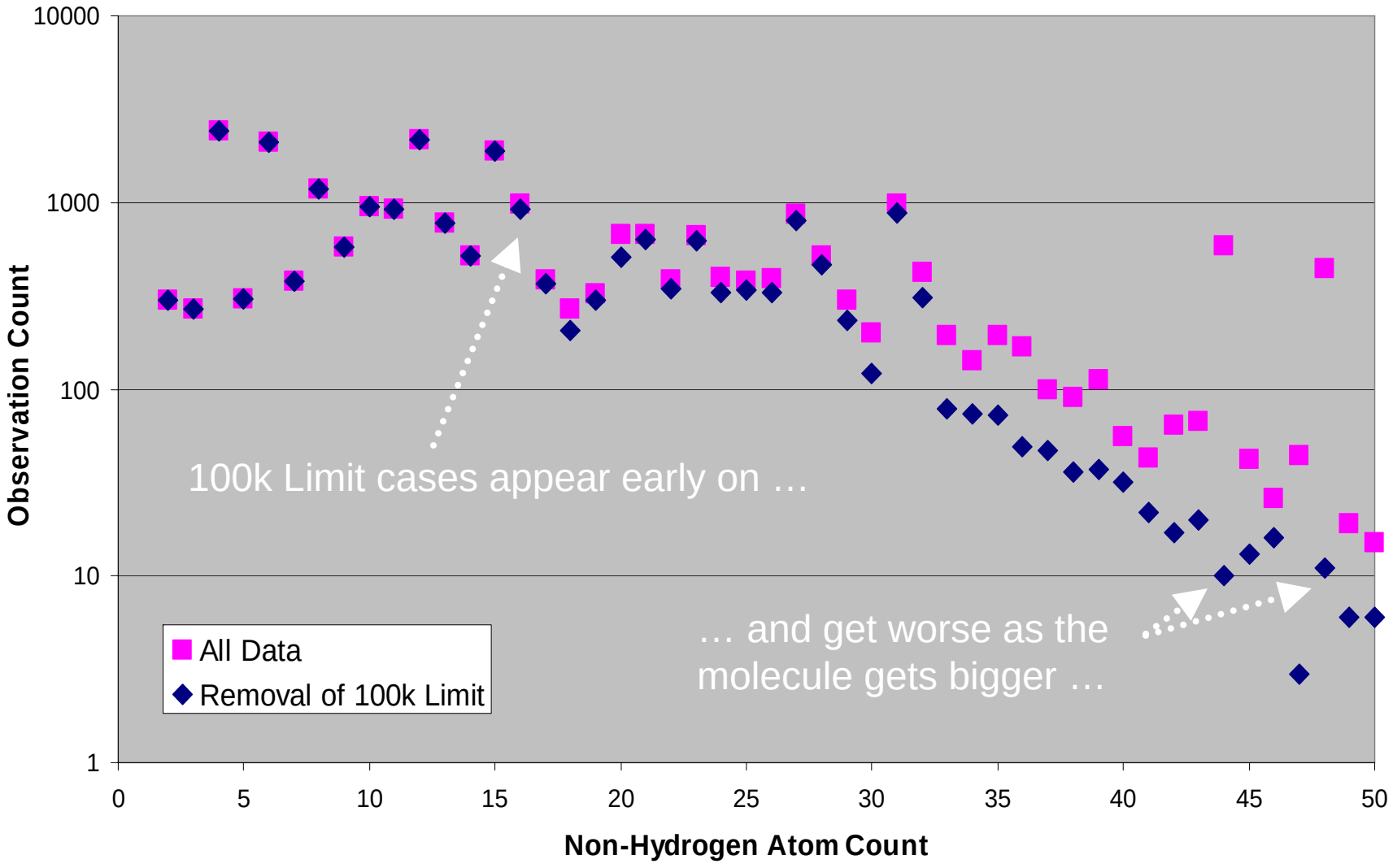
Shape Tanimoto Accuracy



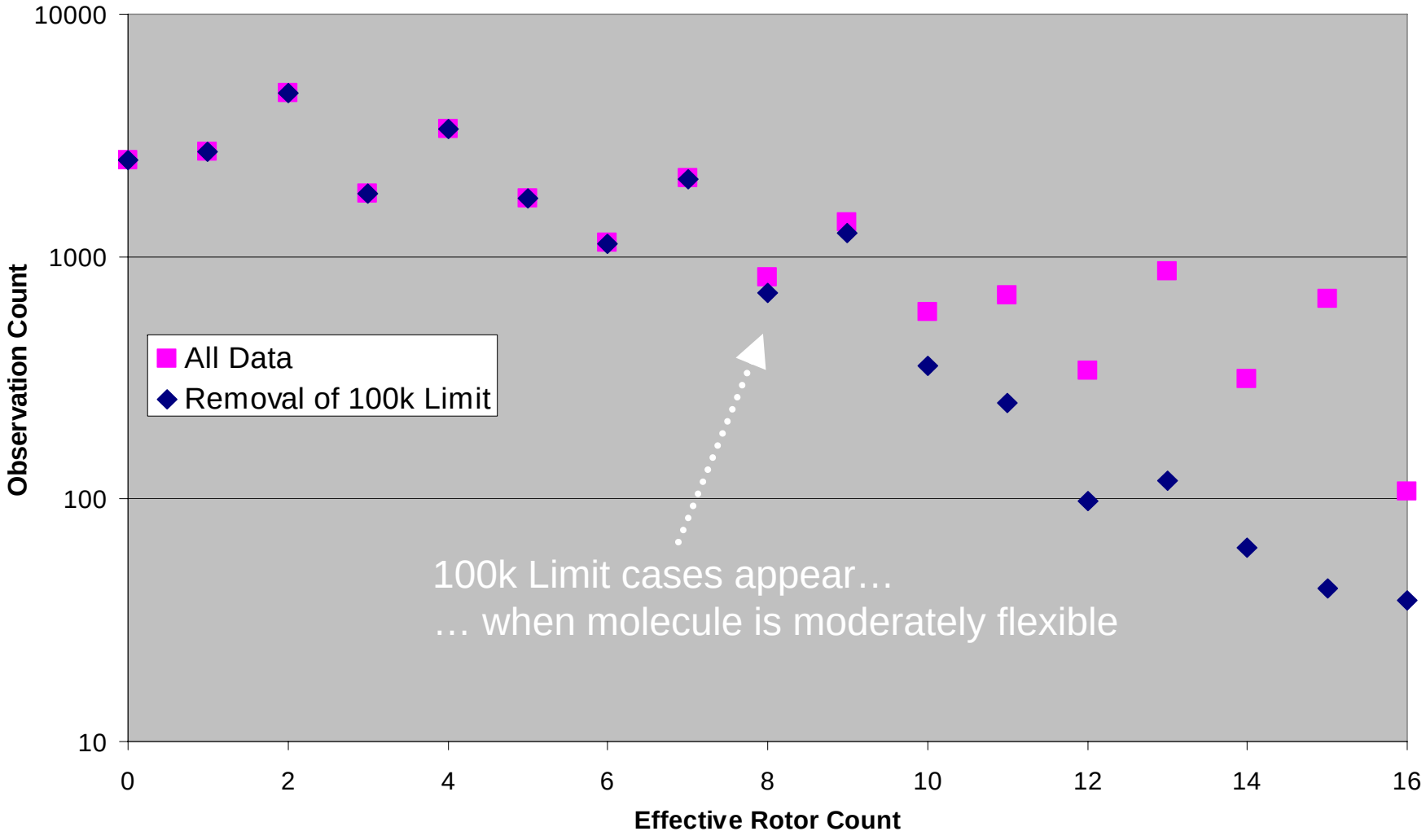


Average Shape Tanimoto values for all 25,972 3-D reference structures as a function of Size and Flexibility

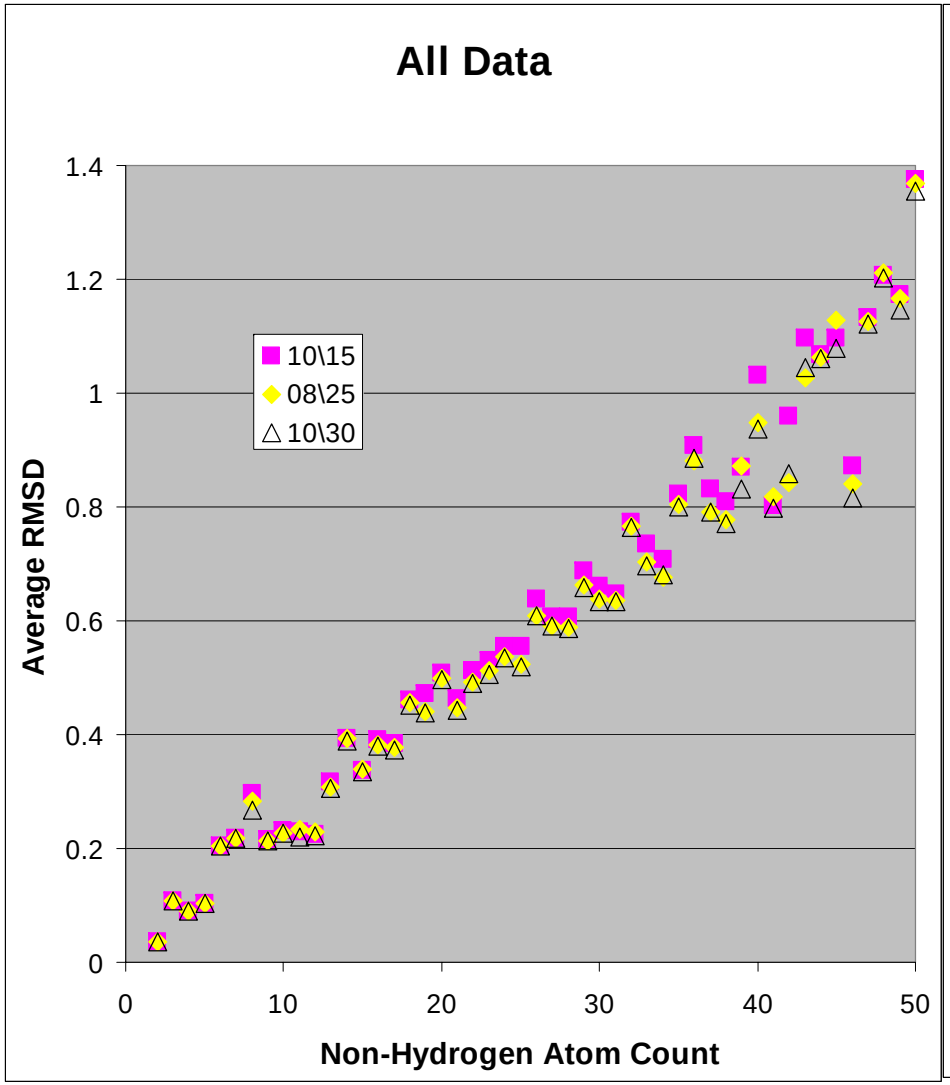
100k Conformer Limitation

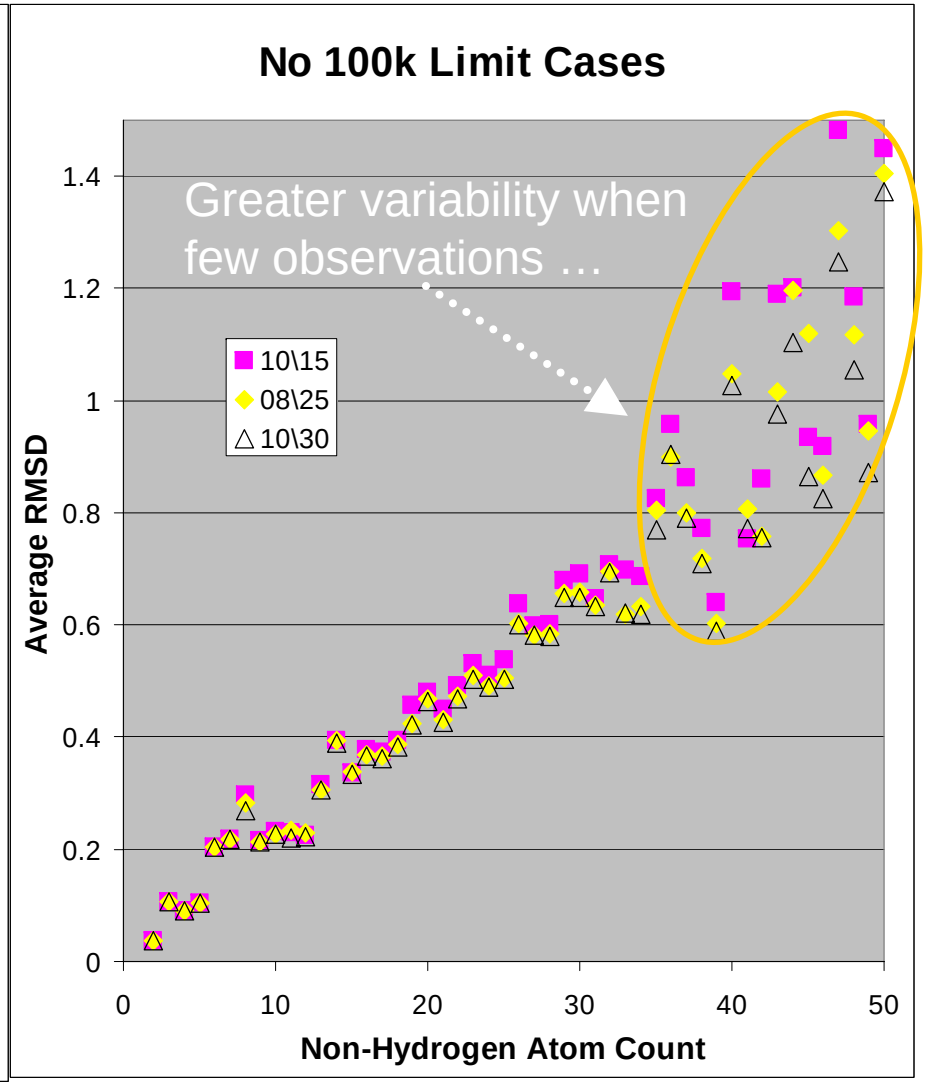


100k Conformer Limitation

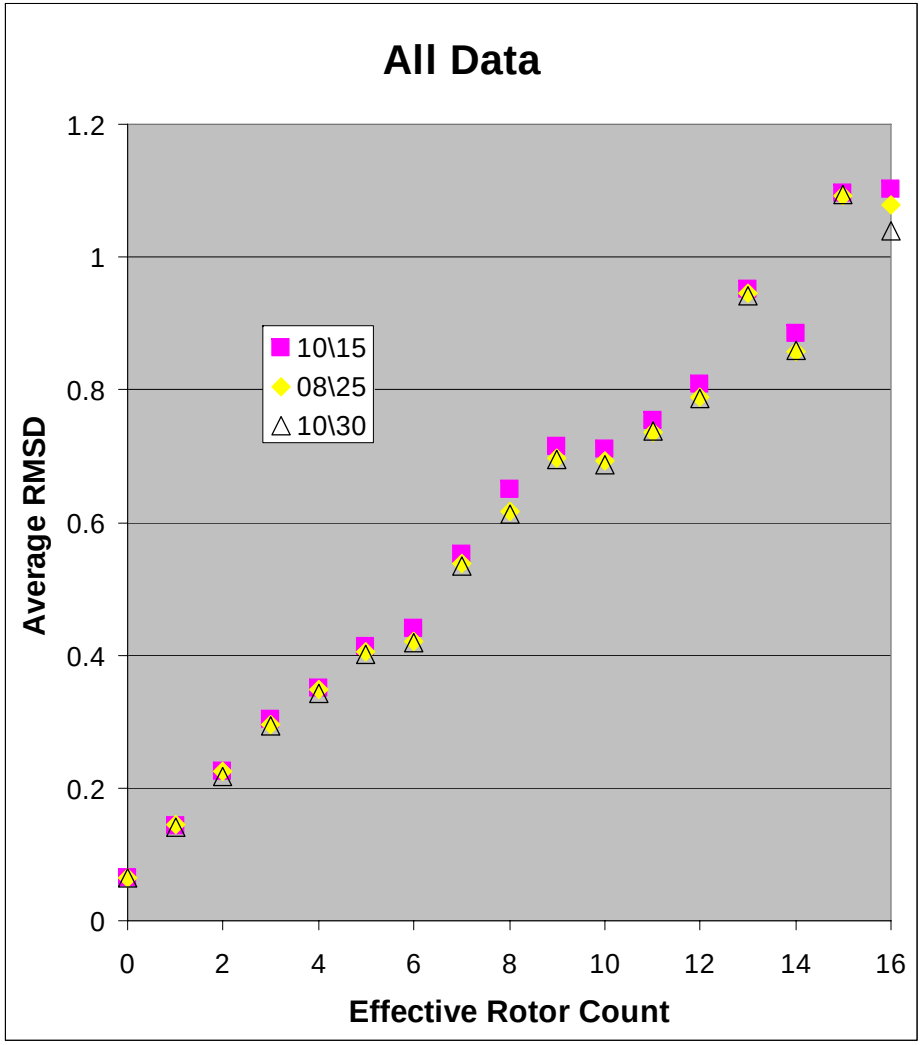


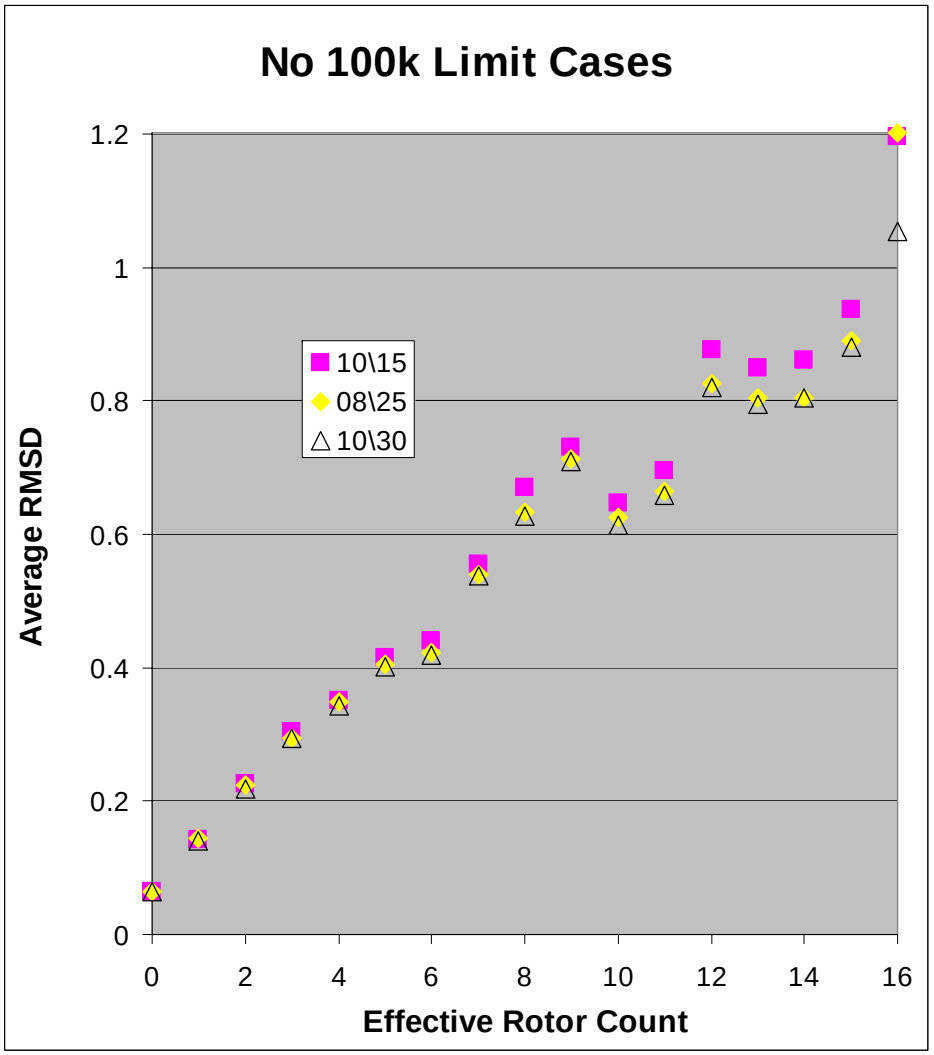
Affect of 100k Limit and Accuracy



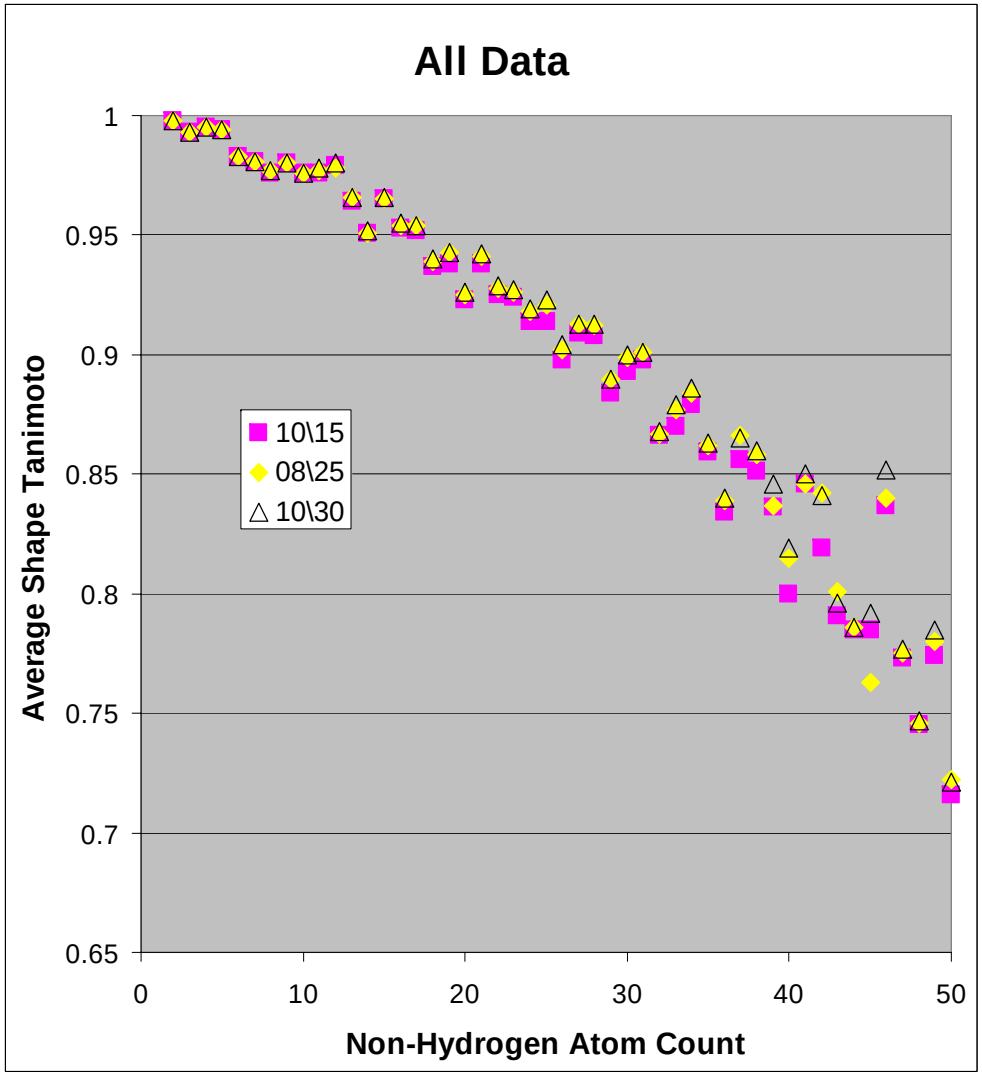


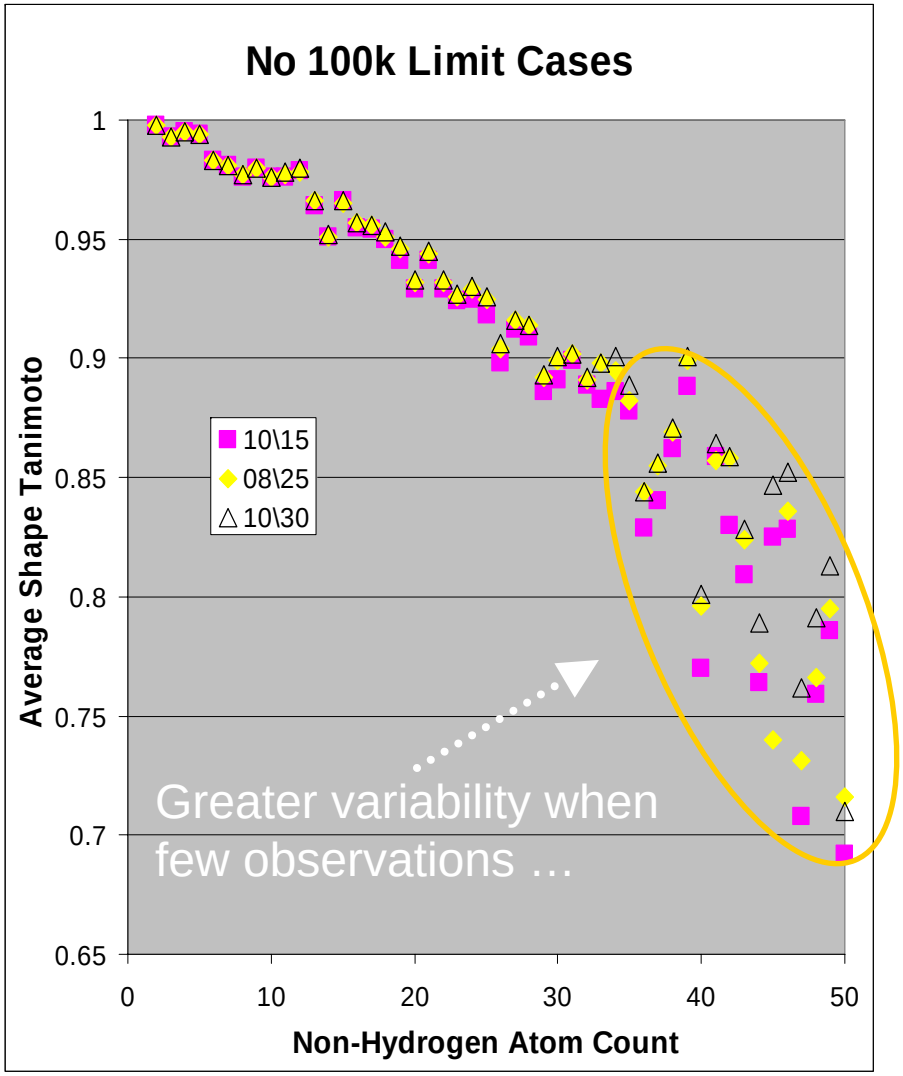
Affect of 100k Limit and Accuracy



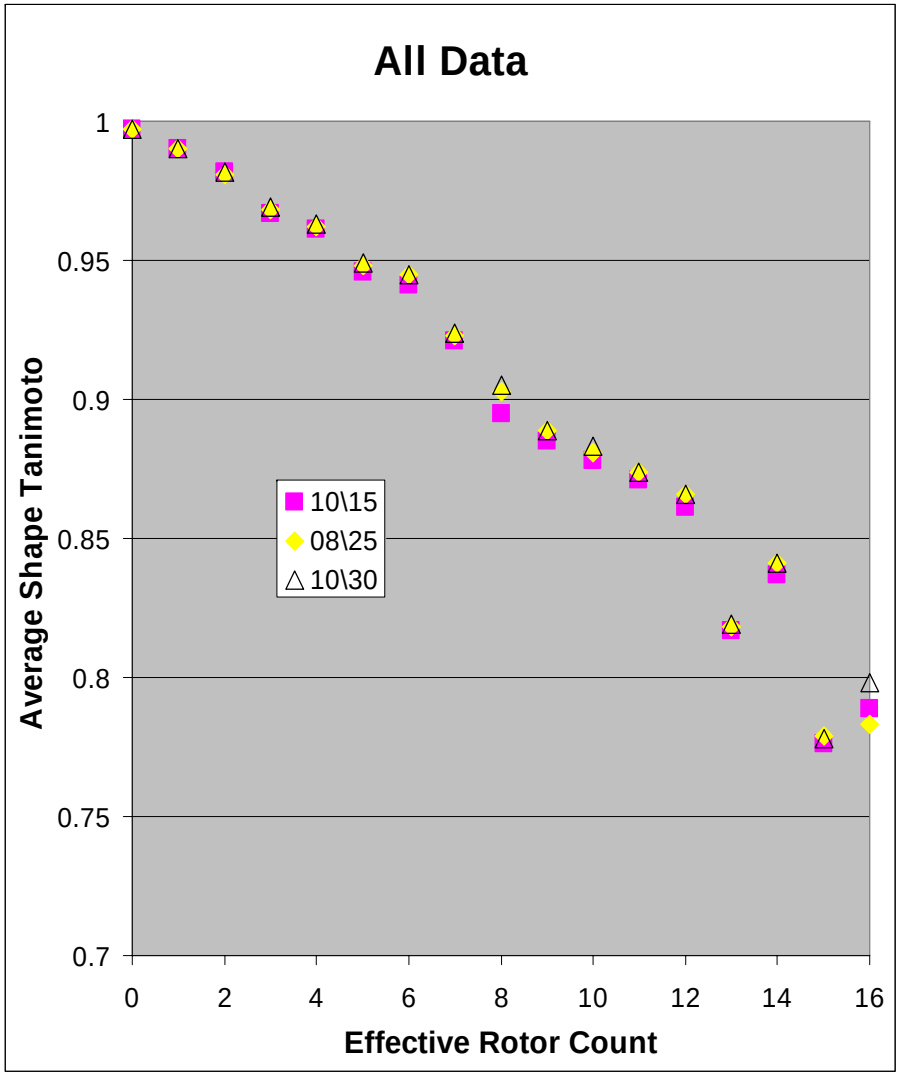


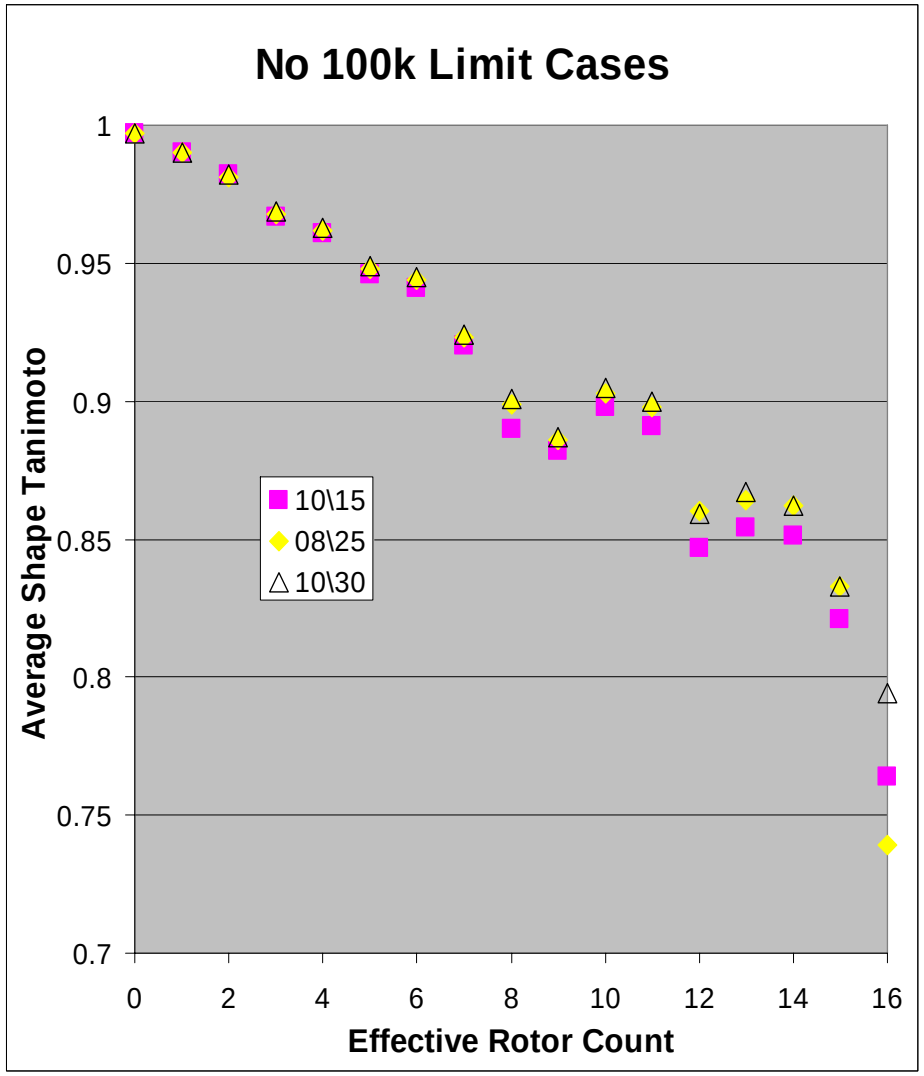
Affect of 100k Limit and Accuracy



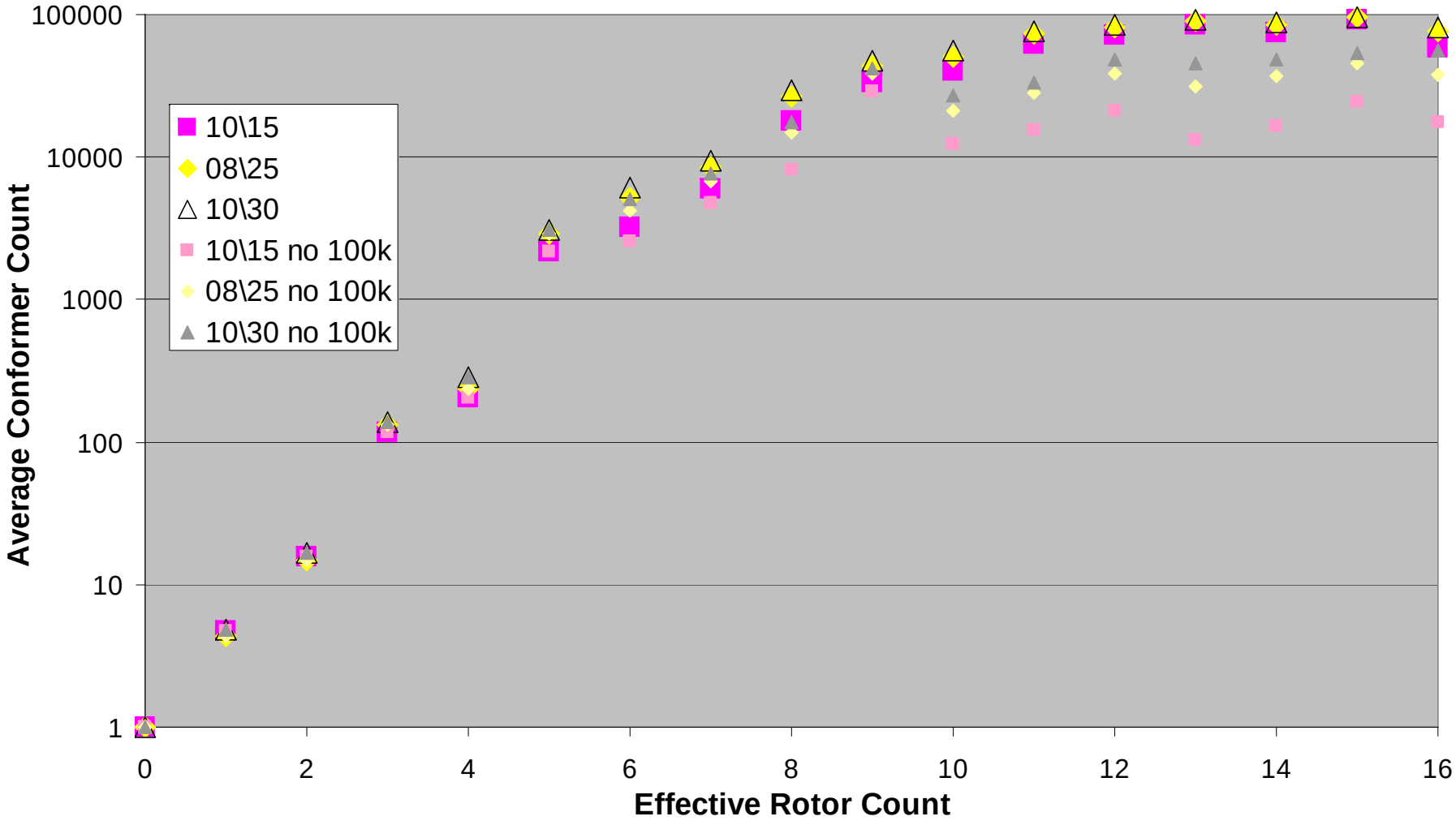


Affect of 100k Limit and Accuracy





100k Limit and Conformer Count



Omega: What was learned?

- 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 MM94s
  … Reasonable default energy windows for Fragment/Conformer

  However …
  … Total Conformer Count may limit overall accuracy

- Use of full MMFF94s force field produces fewer conformers and 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

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

Steve Bryant

Yulia Borodina

Jie Chen

Svetlana Dracheva

Fabien Fontaine

Lewis Geer

Lianyi Han

Jane He

Siqian He

Wolf-Dietrich Ihlenfeldt

Karen Karapetian

Wenyao Shi

Ben Shoemaker

Vahan Simonyan

Tugba Suzek

Paul Thiessen

Valery Tkachenko

Jiyao Wang

Yanli Wang

Jewen Xiao

Jian Zhang