

# Validation with the RCSB: Good idea or bad idea?

Paul Hawkins & Gregory Warren



# The non-expert's view

- 10,000 hours required for expertise
- Numbers so far
  - Hours: 50.75
  - Naps: 20
  - Insights: 1



# Validating conformer generators

- It's a solved problem, isn't it?
- The obvious approach
  - Reproducing “the” bioactive conformation.
- How to select a dataset
  - Appropriate structures
  - Appropriate ligands



# It's a solved problem.....

Journal of Molecular Graphics and Modelling 21 (2003) 449–462

www.elsevier

## Assessing the performance of OMEGA with respect to retrieving bioactive conformations

Jonas Boström<sup>a,\*</sup>, Jeremy R. Greenwood<sup>b</sup>, Johan Gottfries<sup>a</sup>

*J. Chem. Inf. Model.* 2005, 45, 461–476

## Comparison of Conformational Analysis Techniques To Generate Pharmacophore Hypotheses Using Catalyst

Rajendra Kristam,<sup>†,§</sup> Valerie J. Gillet,<sup>\*,†</sup> Richard A. Lewis,<sup>‡,||</sup> and David Thorner<sup>‡</sup>

## Comparative Analysis of Protein-Bound Ligand Conformations with Respect to Catalyst's Conformational Space Subsampling Algorithms

Johannes Kirchmair,<sup>†</sup> Christian Laggner,<sup>†</sup> Gerhard Wolber,<sup>‡</sup> and Thierry Langer<sup>\*,†,‡</sup>

## Frog: a FRee Online druG 3D conformation generator

T. Bohme Leite<sup>1</sup>, D. Gomes<sup>1</sup>, M.A. Miteva<sup>2</sup>, J. Chomilier<sup>3</sup>, B.O. Villoutreix<sup>2</sup> and P. Tufféry<sup>1,\*</sup>



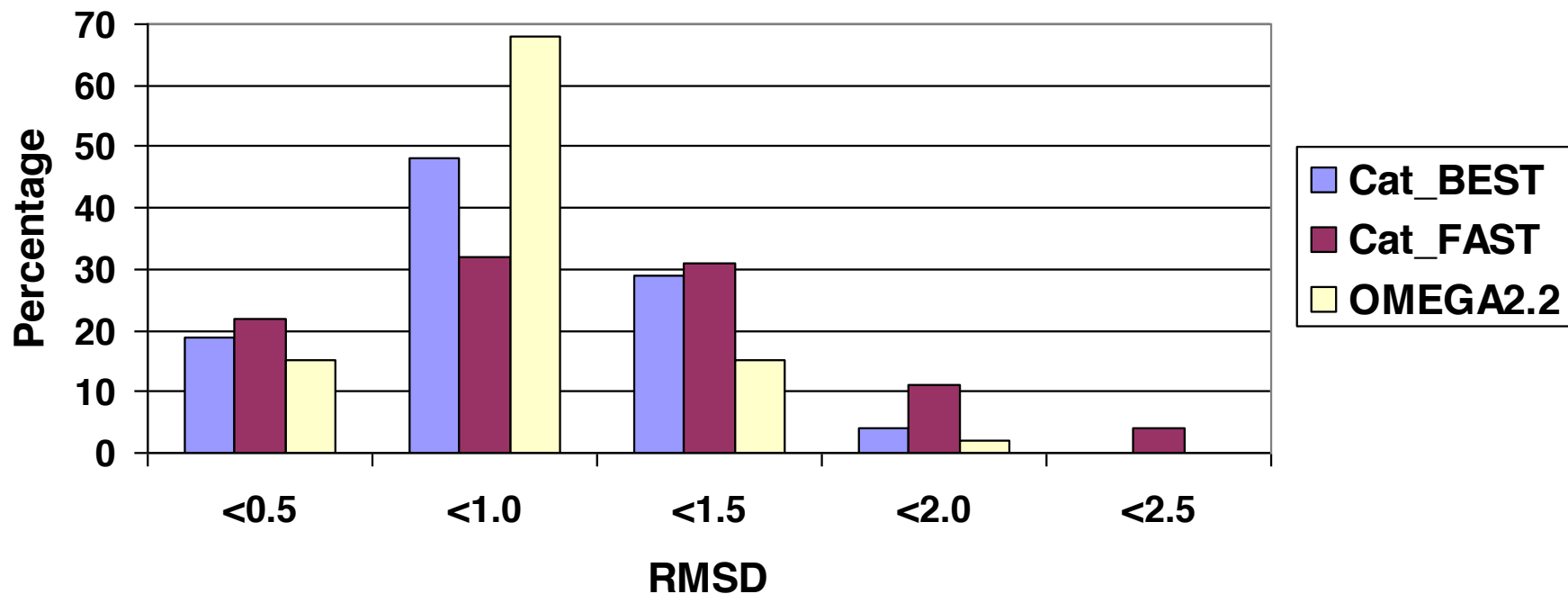
# Testing a conformer generator

- Go to RCSB, get ligands:
  - Rotor count  $\geq 2$  &  $\leq 16$
  - “Drug-like”
  - Diverse
- Generate conformers
- Calculate metric to compare bioactives to computed structures
  - Lowest RMSD to any computed conformer
  - Other geometric
- Get reliable indicator of future predictions
  - Hmm...



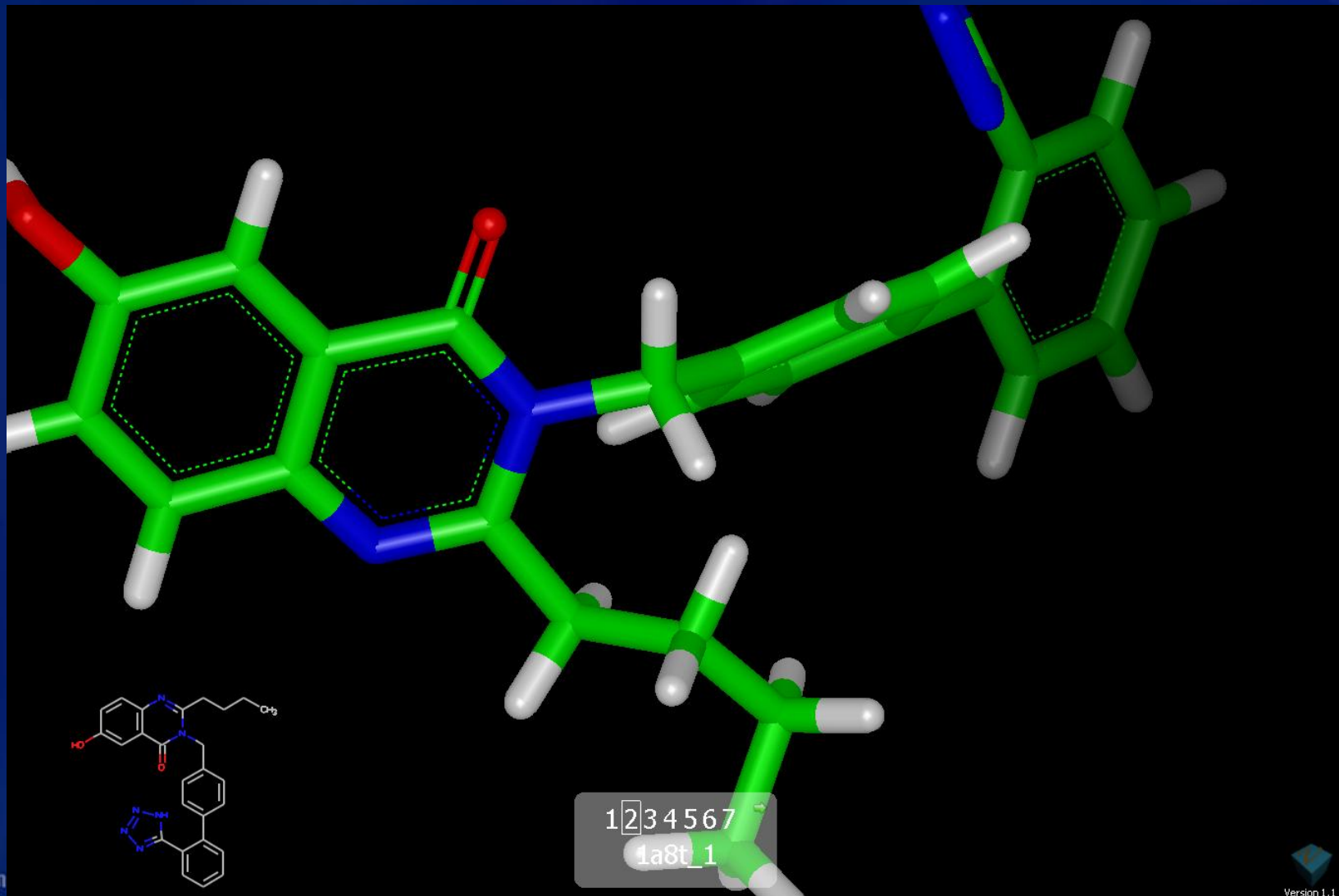
# Reproduction of 100 bioactive conformations

## Catalyst & OMEGA2



*J. Med. Chem.*, 2004, 47, 2499-2510.

# Incorrect ligand conformations



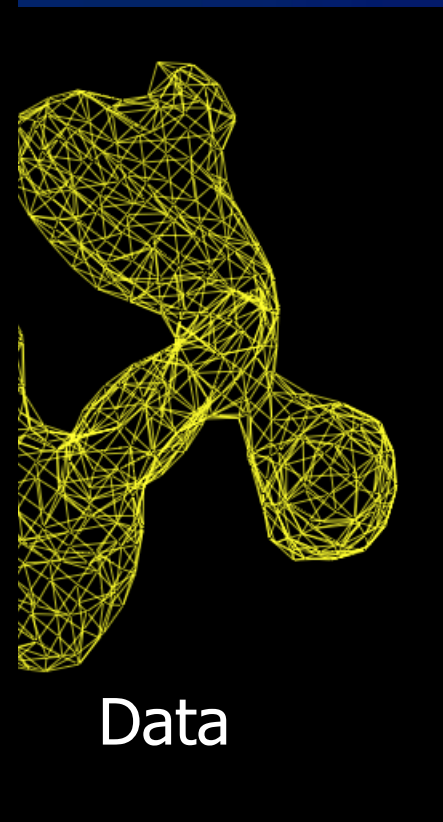
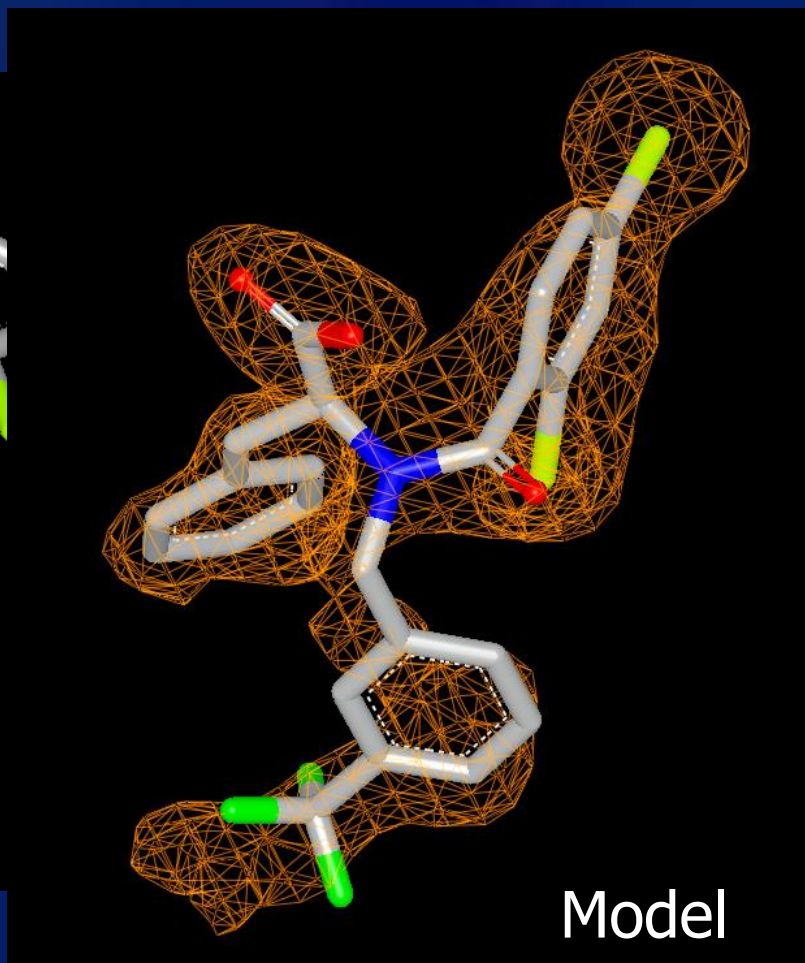
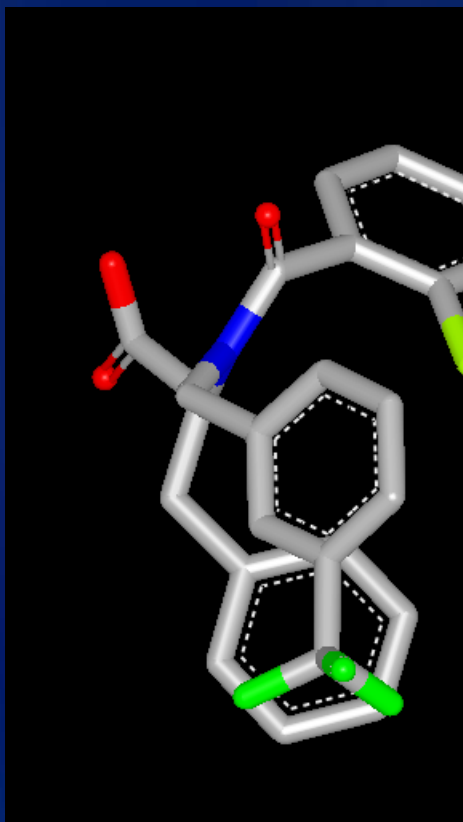
Open

# Problems

- RMSD is a terrible metric
  - Range varies with rotor count
  - Often gives inappropriate measure of fit
    - Shape Tanimoto
- The RCSB is full of rubbish
- Protein structures are not accurate/precise
- Dependency on dataset composition unknown
  - How robust are predictions of performance?



# X-ray Structures are Models



# RMSD - Limitations

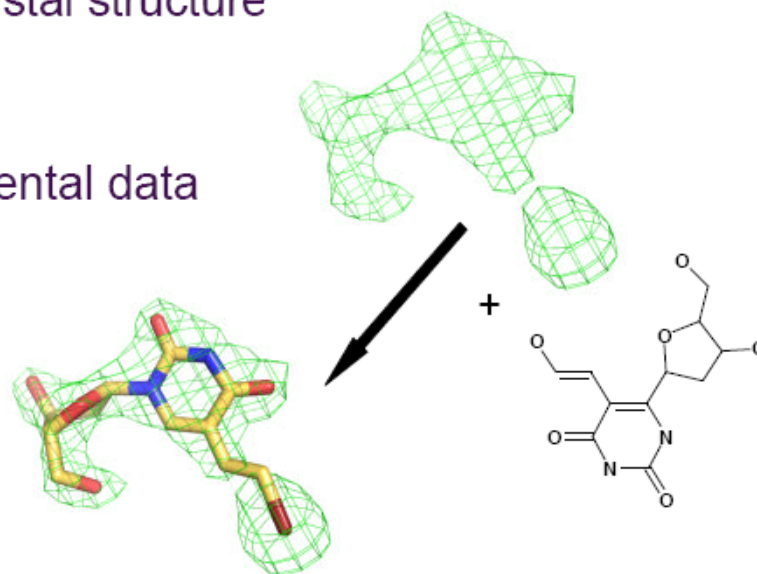
➤ fundamental/scientific:

ligand coordinates from crystal structure:

Better to compare  
calculated density  
to experimental:  
RSR/RSCC/RST

≠ experimental data

= "model"



⇒ RMSD as performance criteria:

= compare "docked\_model" with "cryst\_model" ⚡

# The latest and ....

1848

*J. Chem. Inf. Model.* 2006, 46, 1848–1861

## **Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations**

- **778 co-crystal structures from PDB**
  - Assembled manually
  - High statistical power
- **Compare experimental conformation to computed sets**
  - Report lowest RMSD of set to “the” experimental



# Proteins: Global metrics

- Data/parameter ratio  $> 1$ 
  - Resolution  $\leq 2.7 \text{ \AA}$ 
    - Measure only of quantity of data
- Must have structure factors
  - Independent verification of structure
- Reasonable fit to model
  - $R_{\text{free}} - R \leq 0.05$
- Atomic precision suitable
  - DPI  $< 0.5 \text{ \AA}$



# Atomic coordinate precision

- Goto *et al.*, *J. Med. Chem.*, **2004**, *47*, 6804.

Volume of asymmetric unit cell



$$\sigma(r, B_{\text{avg}}) = 2.2 N_{\text{atoms}}^{1/2} V_{\text{a}}^{1/3} n_{\text{obs}}^{-5/6} R_{\text{free}}$$

*Acta Cryst.* **2002**, D58, 792.

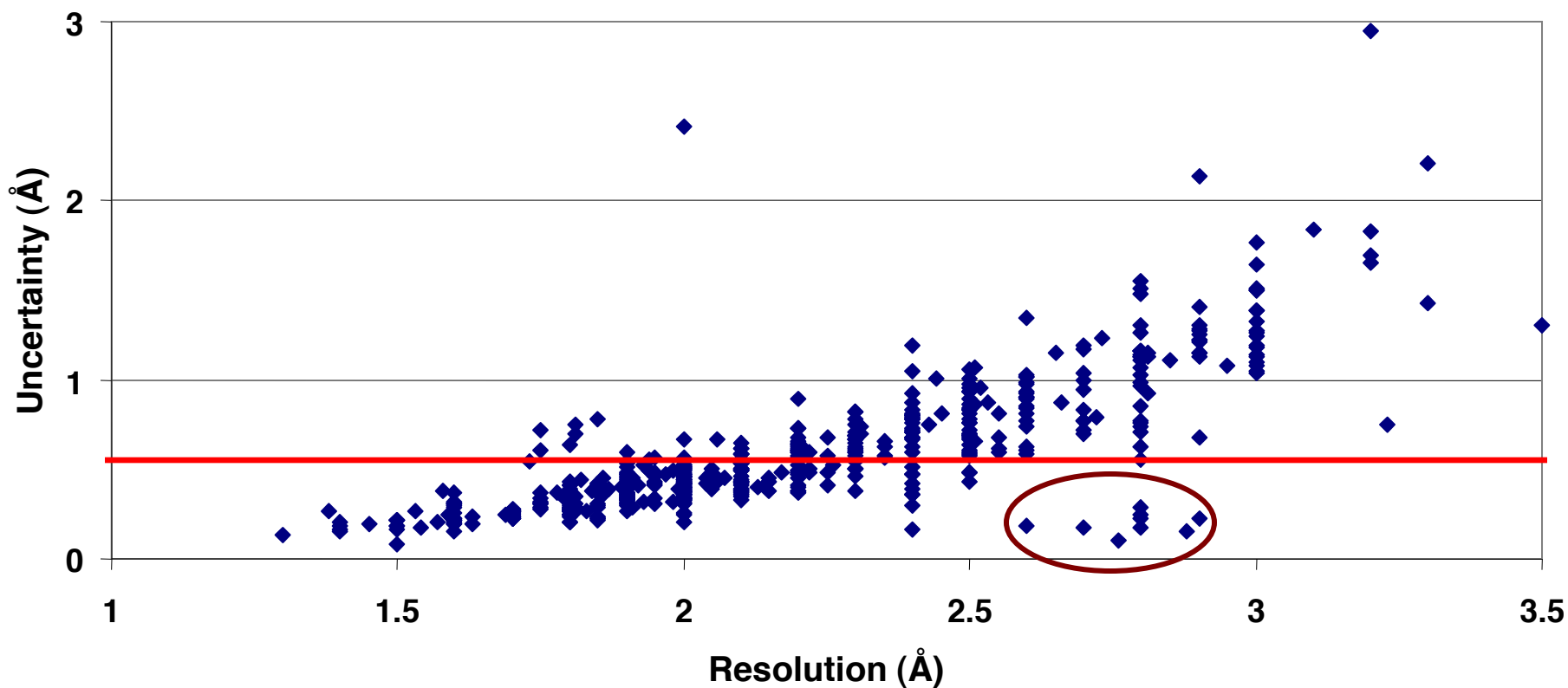


No. of observations

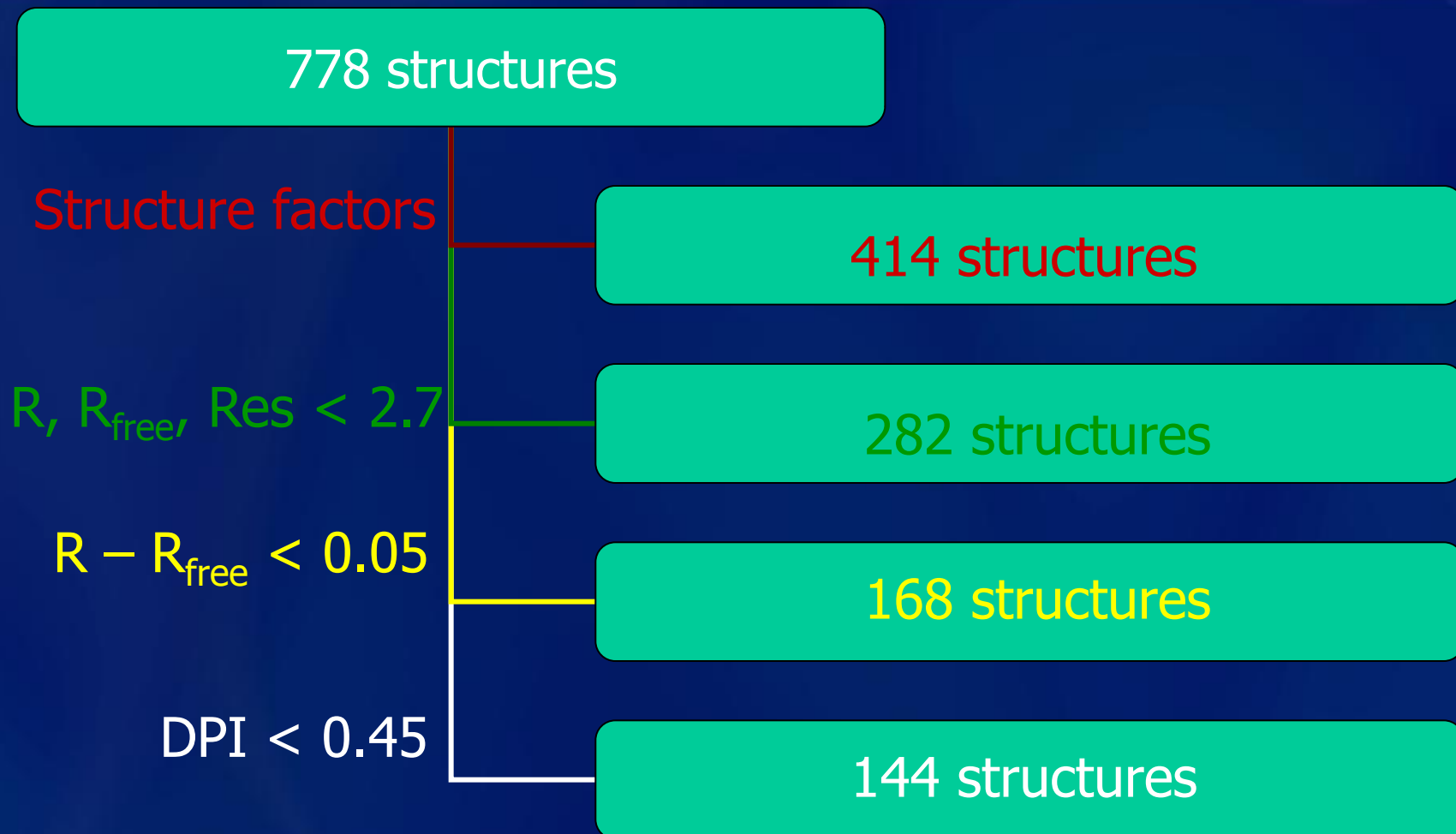
$$\text{Uncertainty} = \sqrt{2} * \sigma$$

# Quality $\propto 1/(\text{Nominal Resolution})$ ?

571 structures from Kirchmair *et al.*



# Kirchmair: Global measures

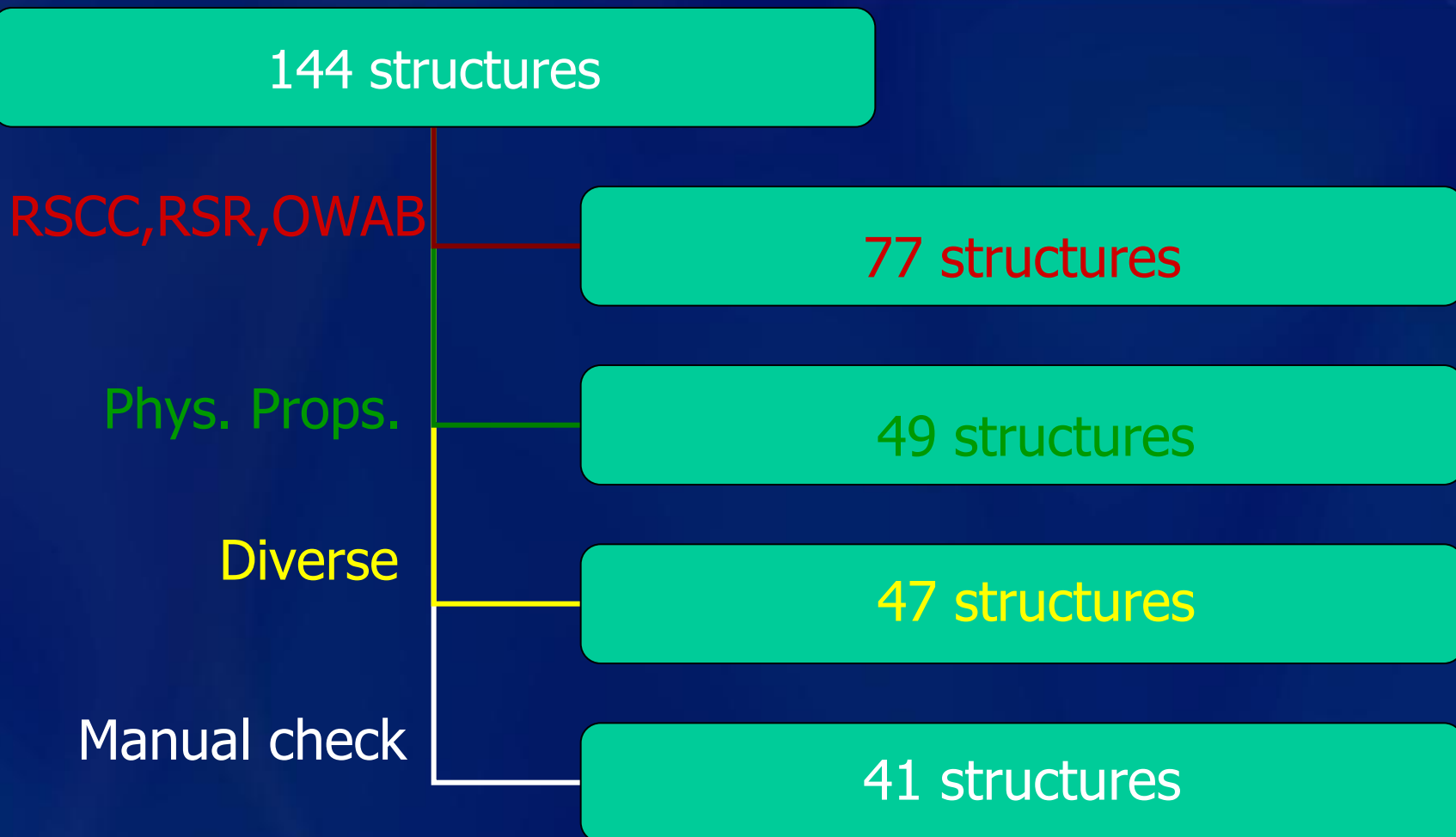


# Proteins: Local metrics

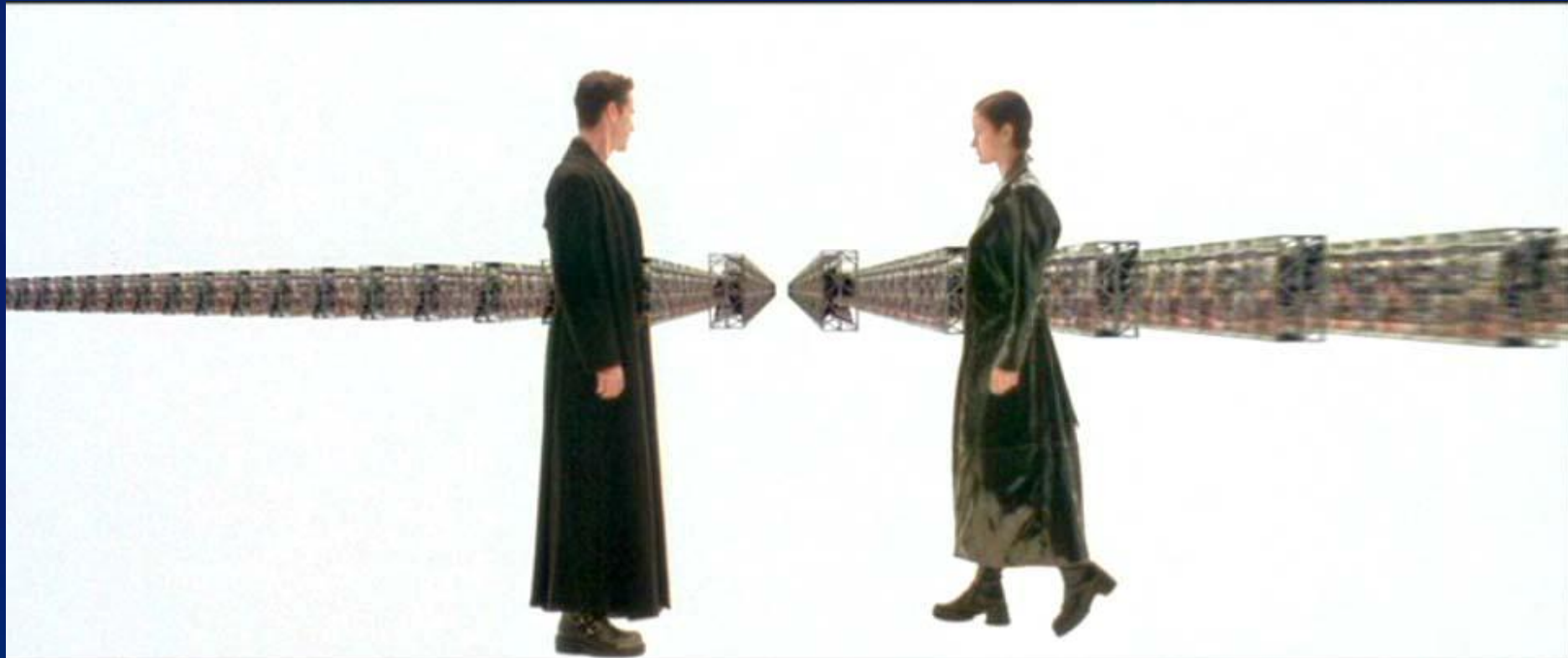
- **RSCC > 0.9**
  - Correlation between model map and experimental map
    - $-1.0 < \text{RSCC} < +1.0$
- **RSR < 0.2**
  - Similar to a normal R-factor
- **Occupancy-weighted B-factor < 50**
- **Ligand properties**
  - Rotor count, number of heavy atoms, diversity



# Kirchmair: Local measures



**“Data. Lots of data.”**



# PDBbind to the rescue...

- <http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp>
  - Measured binding affinities
- 3214 protein-ligand pairs
- Remove covalent and incomplete ligand
  - 2519 remain



# PDBbind: Global measures

2519 structures

Structure factors

1562 structures

$R, R_{\text{free}}, \text{Res} < 2.7$

1245 structures

$R - R_{\text{free}} < 0.05$

888 structures

DPI < 0.45

766 structures



# PDBbind: Local measures

766 structures

RSCC, RSR, OWAB

457 structures

Phys. props.

268 structures

Diverse, functs.

198 structures

Manual check

161 structures



# MIMUMBA set

*J. Chem. Inf. Model.* 2006, 46, 2305–2309

2305

## MIMUMBA Revisited: Torsion Angle Rules for Conformer Generation Derived from X-ray Structures<sup>†</sup>

Jens Sadowski\* and Jonas Boström

- 1267 co-crystal structures from PDB
  - Assembled manually
  - High statistical power
- Compare experimental conformation to computed sets
  - Report lowest RMSD of set to “the” experimental



# MIMUMBA: Global measures

1223 structures

Structure factors

766 structures

$R, R_{\text{free}}, R_{\text{es}} < 2.7$

634 structures

$R - R_{\text{free}} < 0.05$

485 structures

DPI < 0.45

421 structures



# MIMUMBA: Local measures

421 structures

RSCC, RSR, OWAB

216 structures

Phys. props.

125 structures

Diverse, functs.

117 structures

Manual check

99 structures



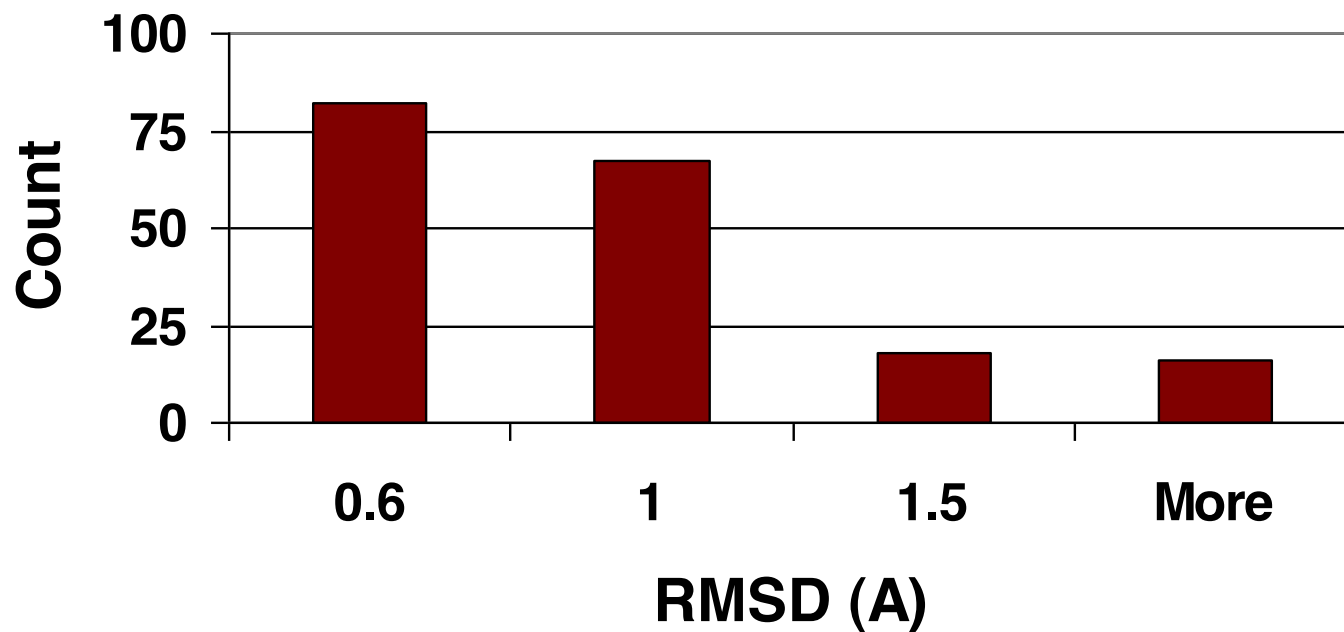
# At last!

- 41 Langer structures, 161 from PDBbind
  - Only 19 duplicates, 183 in total
- Run OMEGA2 with defaults
- Compute lowest RMSD of any conformer to experimental
  - Should also look at % “close” to experimental



# Results from OMEGA

Combined results



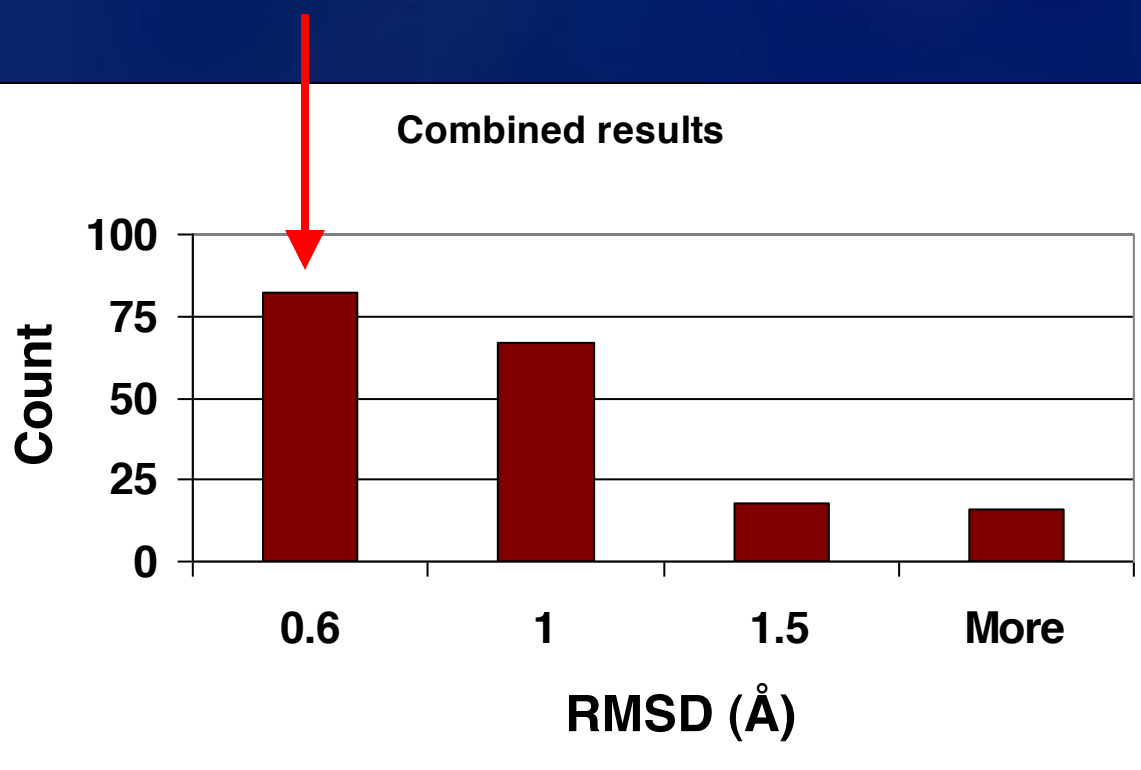
Mean RMSD: 0.75Å  
Median RMSD: 0.63Å

% < 1.25Å RMSD = 7.8



# Adjusting for error

Upper limit for atomic uncertainty.



**MAX(RMSD,Uncert)**  
Mean RMSD: 0.76Å  
Median RMSD: 0.63Å

**RMSD-Uncert**  
Mean RMSD: 0.42Å  
Median RMSD: 0.31Å

# Reliability

- **Bootstrapping**

- Remove subset of A from total of B, recalculate mean RMSD, repeat 10,000 times

- **183 from RCSB**

- Mean = 0.76Å
- A = 20; stddev of means = 0.011

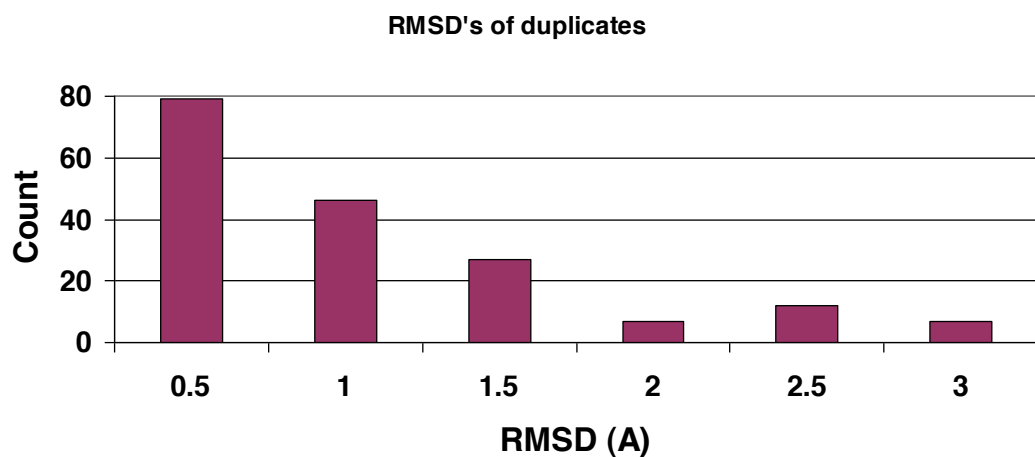


# Well that's that then....

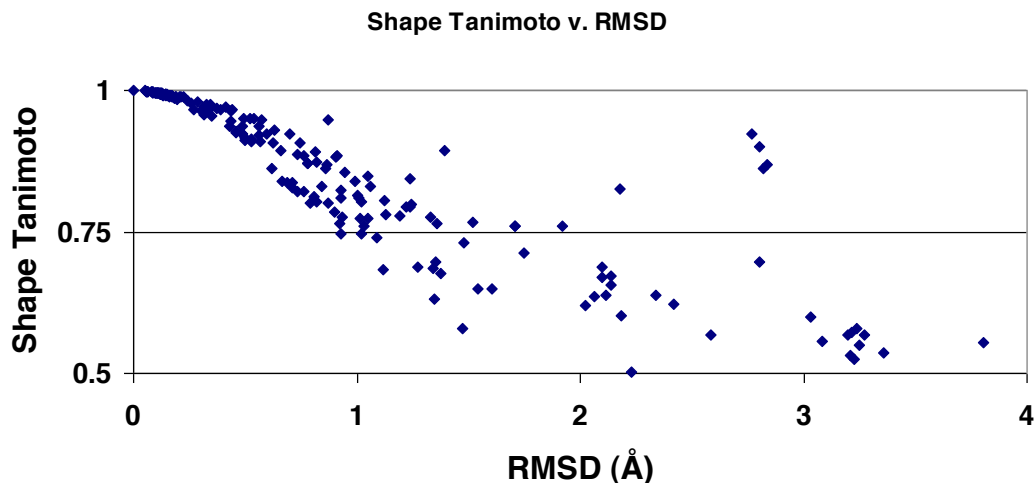
- There are 100 duplicate compounds in Kirchmair set.
- Can there be different conformations?
  - By RMSD
  - By shape



# Differences in duplicates

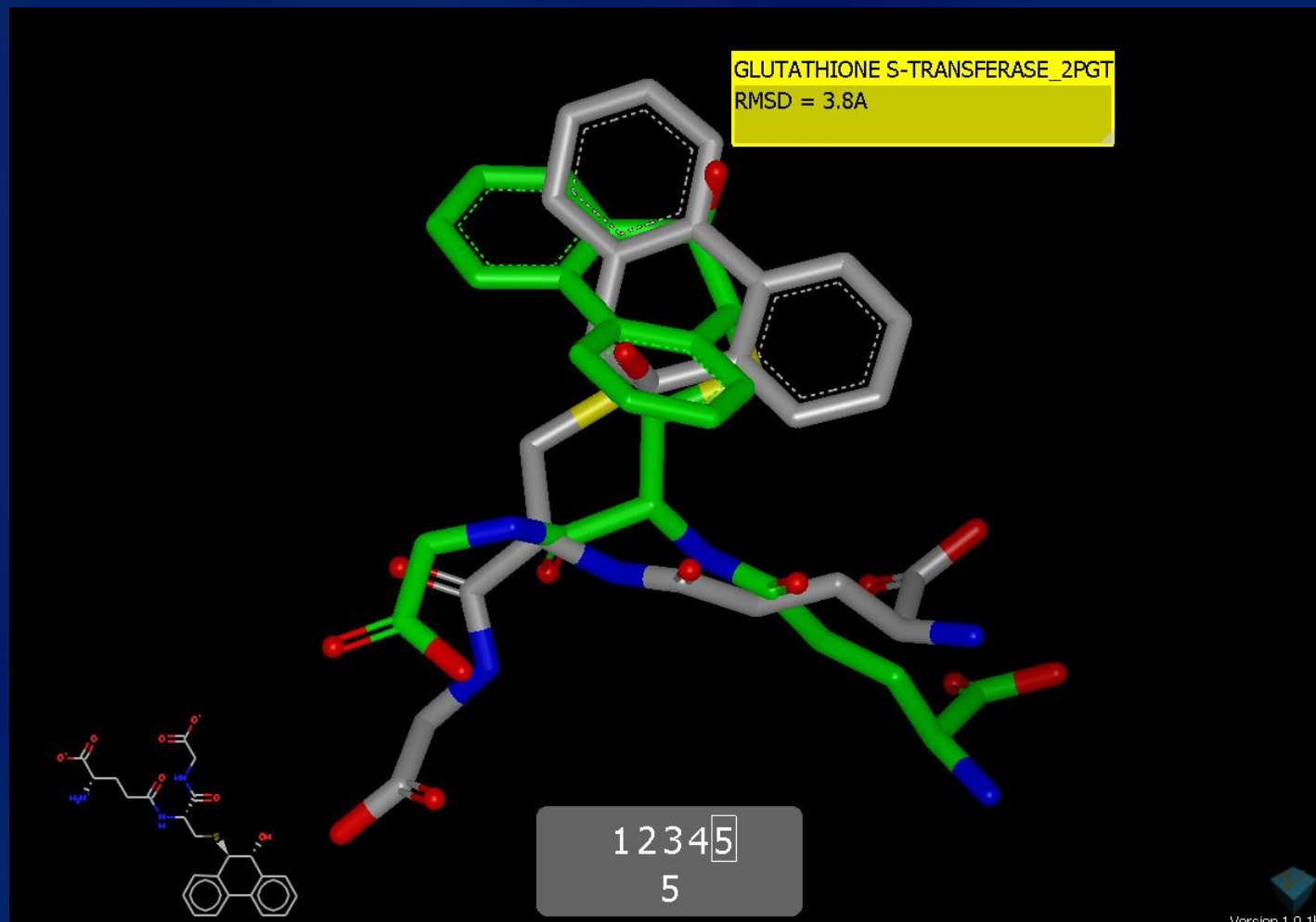


34% > 1Å RMSD



30% < 0.8 shape Tanimoto

# An example



2PGT  
6GSU

C121

RMSD:3.8Å  
S.T.:0.56



# Conclusions

- PDB is a good source of data
  - Most of which is really bad
- Resolution does not indicate quality!
  - Use DPI
- RMSD is a terrible metric
  - Especially if you use it to compare conformers
- There is no such thing as “the” bioactive conformation



# Acknowledgments

- Bob “Train Man” Tolbert
- Bruce “Brian” Cole
- Geoff “Speeling Bee” Skillman
  
- Gerard “DVD” Kleywegt and EDS team
  - <http://eds.bmc.uu.se/eds/>



