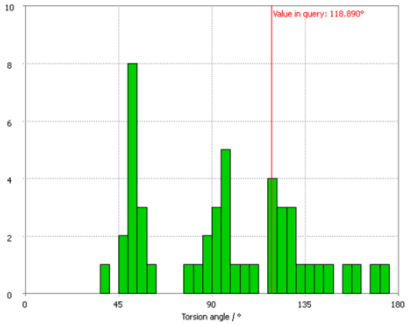
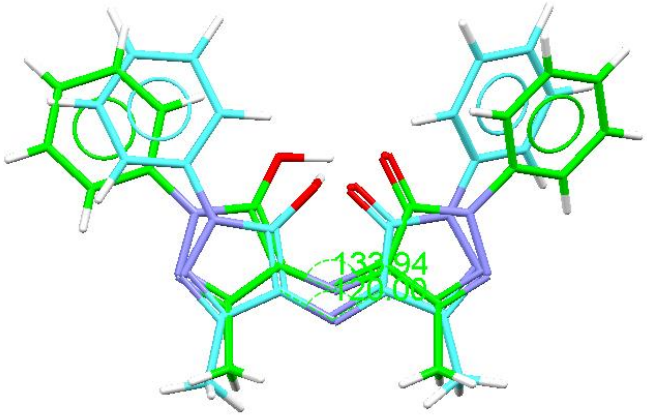


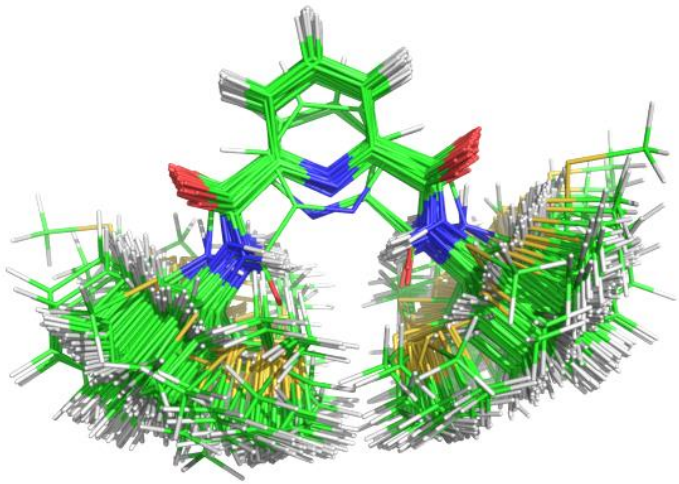
Generating Small Molecule Conformations from Structural Data

Jason Cole

cole@ccdc.cam.ac.uk

Cambridge Crystallographic Data Centre





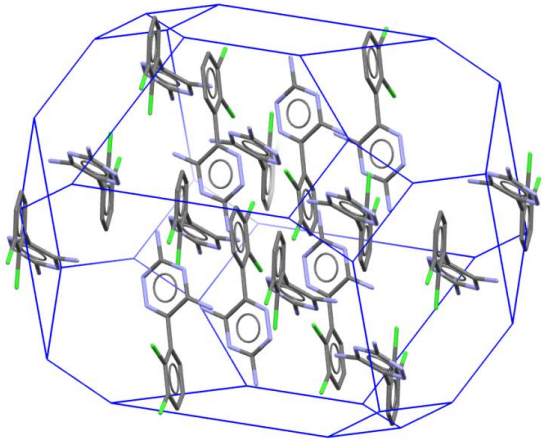
The Cambridge Crystallographic Data Centre

About us

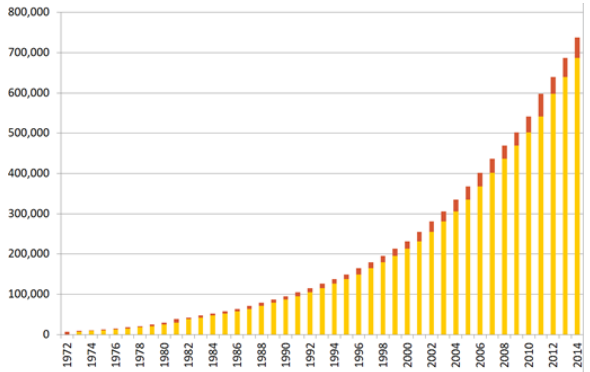
- A not-for-profit, charitable institution, est. 1965
- Self-financing and self-administering since 1989
  - No investors, no shareholders
  - No national, EU or international grant support
  - Funded entirely by contributions

- A University of Cambridge Partner Institute,
  - recognized for postgraduate degrees of the University of Cambridge

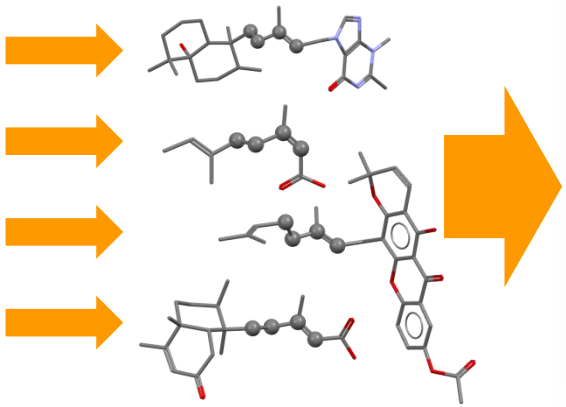
Cambridge Structural Database System

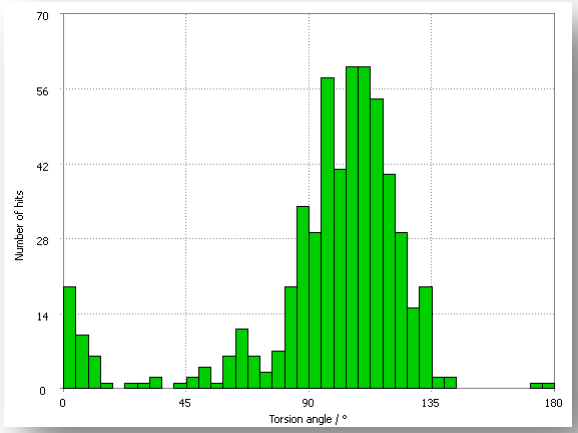


The world's repository of small molecule crystal structures

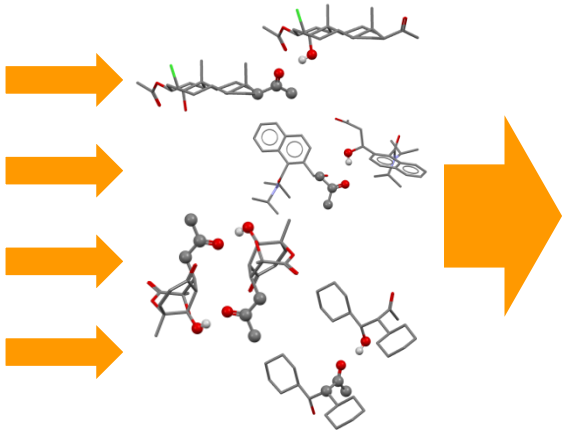


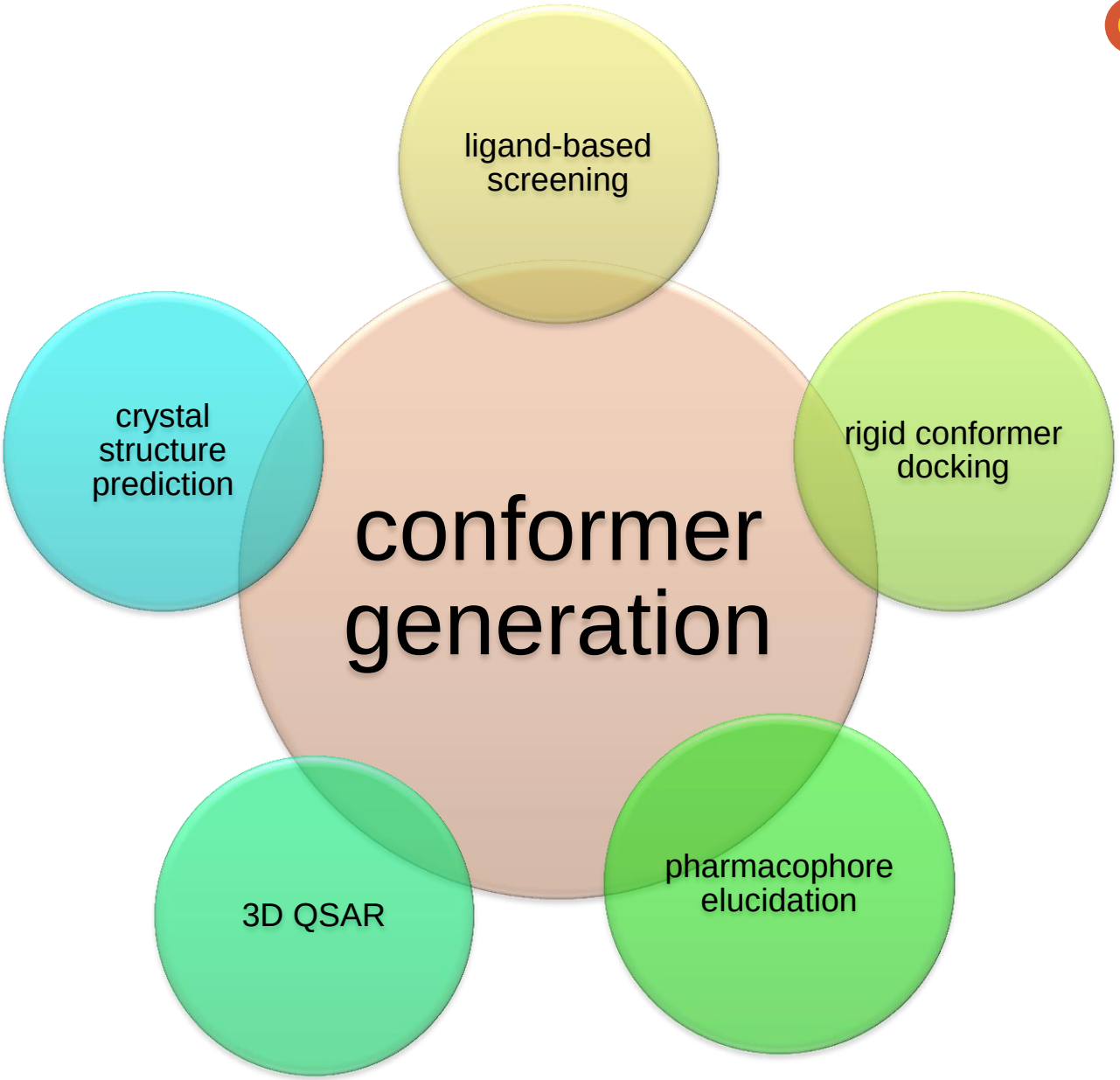
Molecular geometry





Molecular interactions

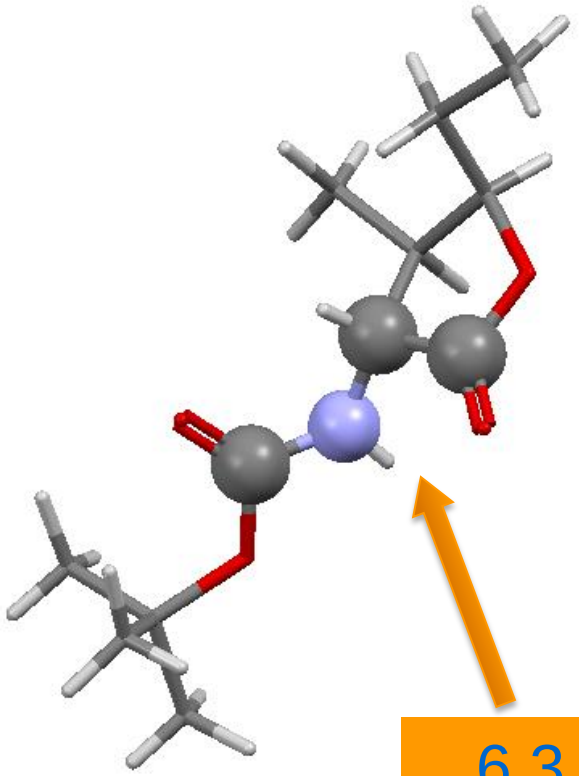



ligand-based screening
crystal structure prediction
rigid conformer docking
conformer generation
3D QSAR
pharmacophore elucidation

Why another conformer generator?

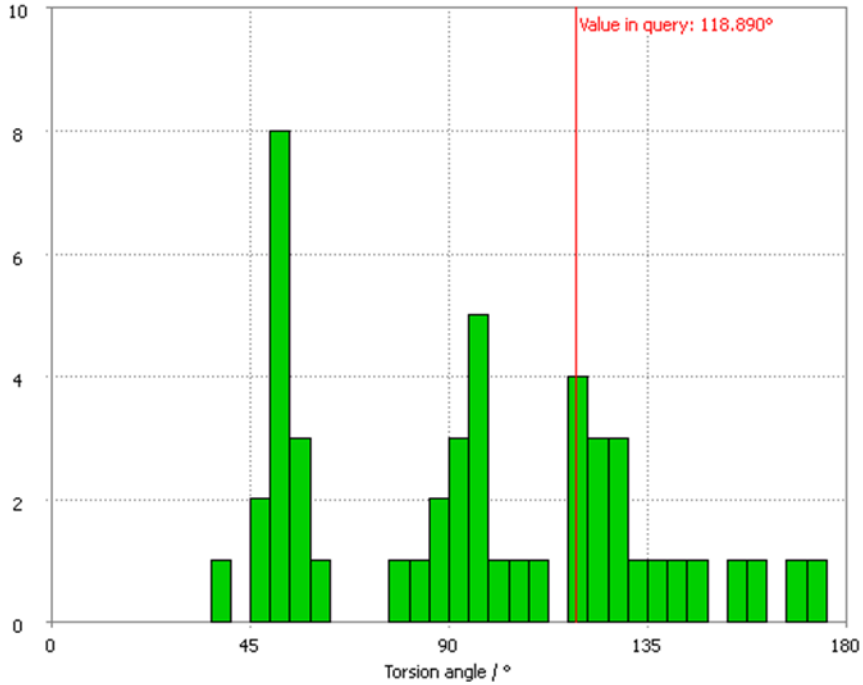
- existing methods rely on systematic or stochastic sampling techniques mainly guided by force-field calculations or empirical rules
- no guarantee, **but** 700,000 high-resolution crystal structures may be a good basis for predicting low-energy conformations
- Will improve as the CSD increases in size and chemical specificity increases
- Could also allow a user to take direct advantage of proprietary crystallographic information

Mogul - A Knowledge Base of Molecular Geometries



6.3.0.1.1.2

keys describe the chemical environment



- rotamer distributions
- torsion angles
- bond lengths
- bond angles
- ring geometries

Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R.; Cooper, R. I.; Harris, S. E.; Orpen, A. G. *J. Chem. Inf. Com. Sci.* 2004, *44*, 2133-2144

Extended Mogul

New Rotamer & Ring Distributions:

- A complete representation of the conformational preferences of a rotatable bond or ring
- One distribution per rotatable bond (unlike a torsion representation)
- Chirality is implicit in the distribution

Faster Searching:

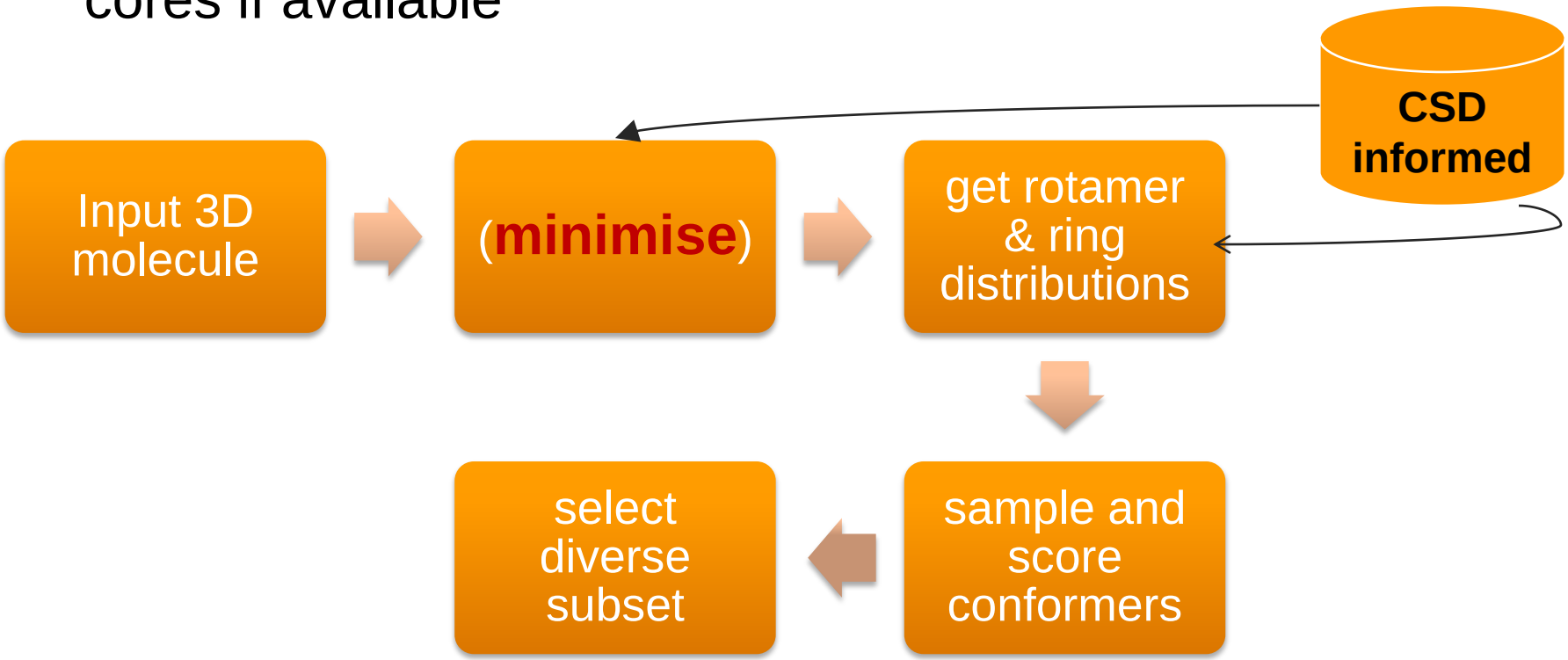
- Flat multiple cascading library structure
- Smaller summary distributions

Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules

R. Taylor, J. Cole, O. Korb, P. McCabe, *J. Chem. Inf. Model.*, **54**, 2500-2514, 2014
10.1021/ci500358p

Conformer Generation Workflow

- embedded in a multi-processing workflow framework
- molecules automatically processed on multiple CPU cores if available



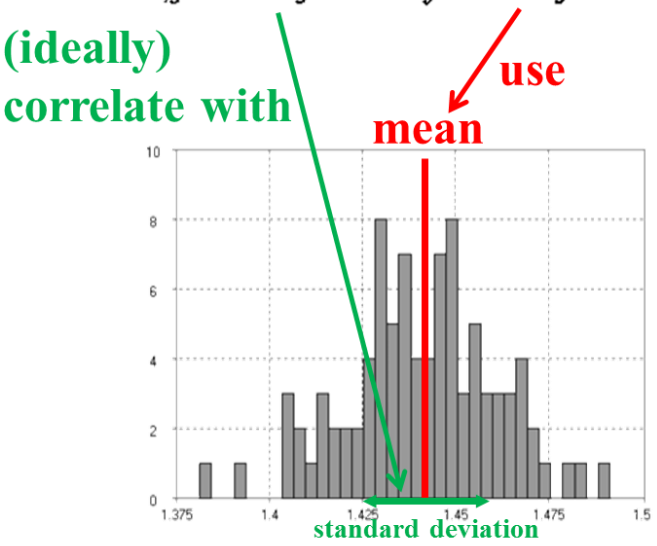
Minimisation

Modified force field approach

- get bond length and bond angle distributions from Mogul
- use distribution means as equilibrium values in Tripos force field
- perform gradient-based Cartesian minimisation

Tripos bond stretching term

$$E_{i,j} = k_{ij} * (d_{ij} - d_{ij}^{0})^2$$



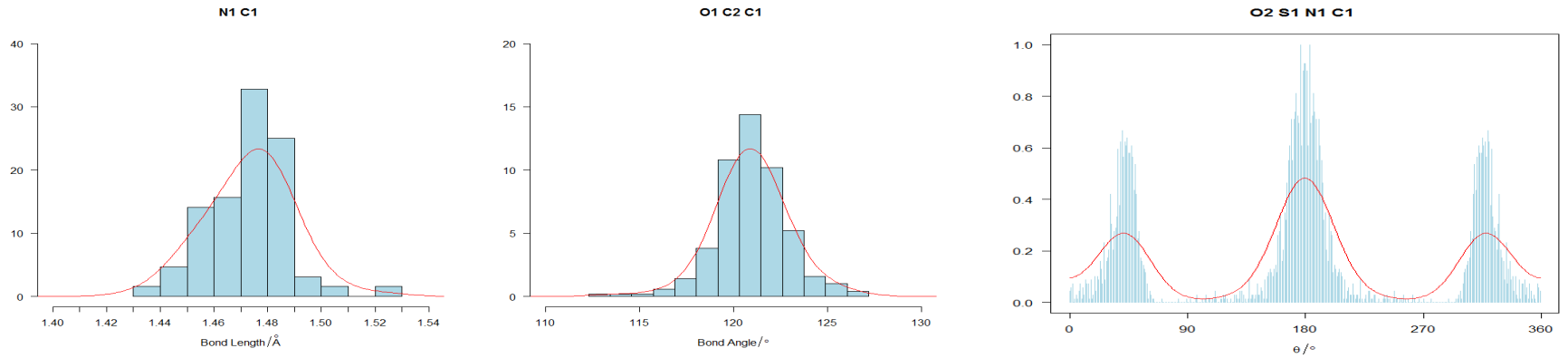
Minimisation

Toward a Fully CSD-Based Approach

- probability density functions (PDF) from histograms using kernel density estimation: $f(x) = \frac{1}{n}\sum_{i=1}^{n} K_i\,(x)$

- kernel functions for bond length and angle data: $K_i(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{h}e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2}$

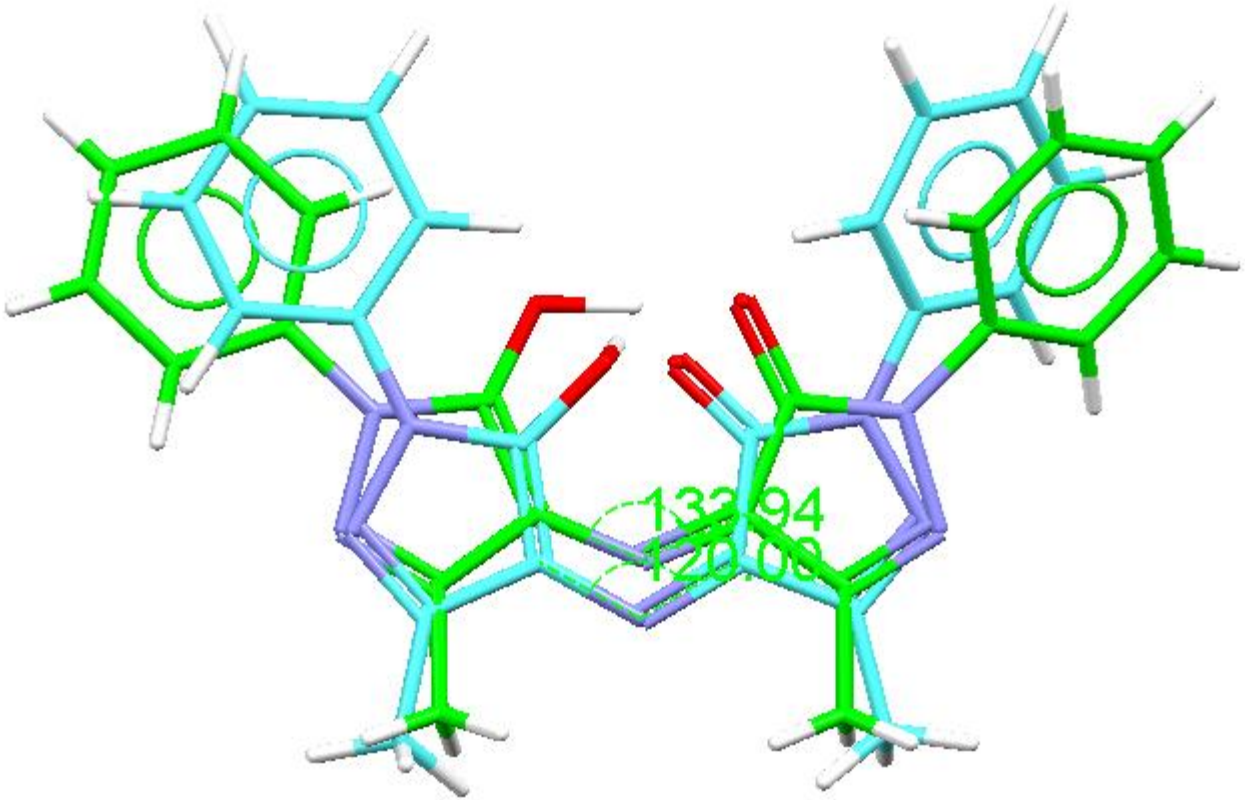
- kernel functions for circular data (torsions): $K_i(\theta) = \frac{1}{2\pi I_0(\nu)}e^{\{\nu(\cos(\theta-\theta_i))\}}$



Kernel Density Estimation Applied to Bond Length, Bond Angle and Torsion Angle Distributions
P. McCabe, O. Korb, J. C. Cole, *J. Chem. Inf. Model.*, **54**, 1284-1288, 2014 10.1021/ci500156d

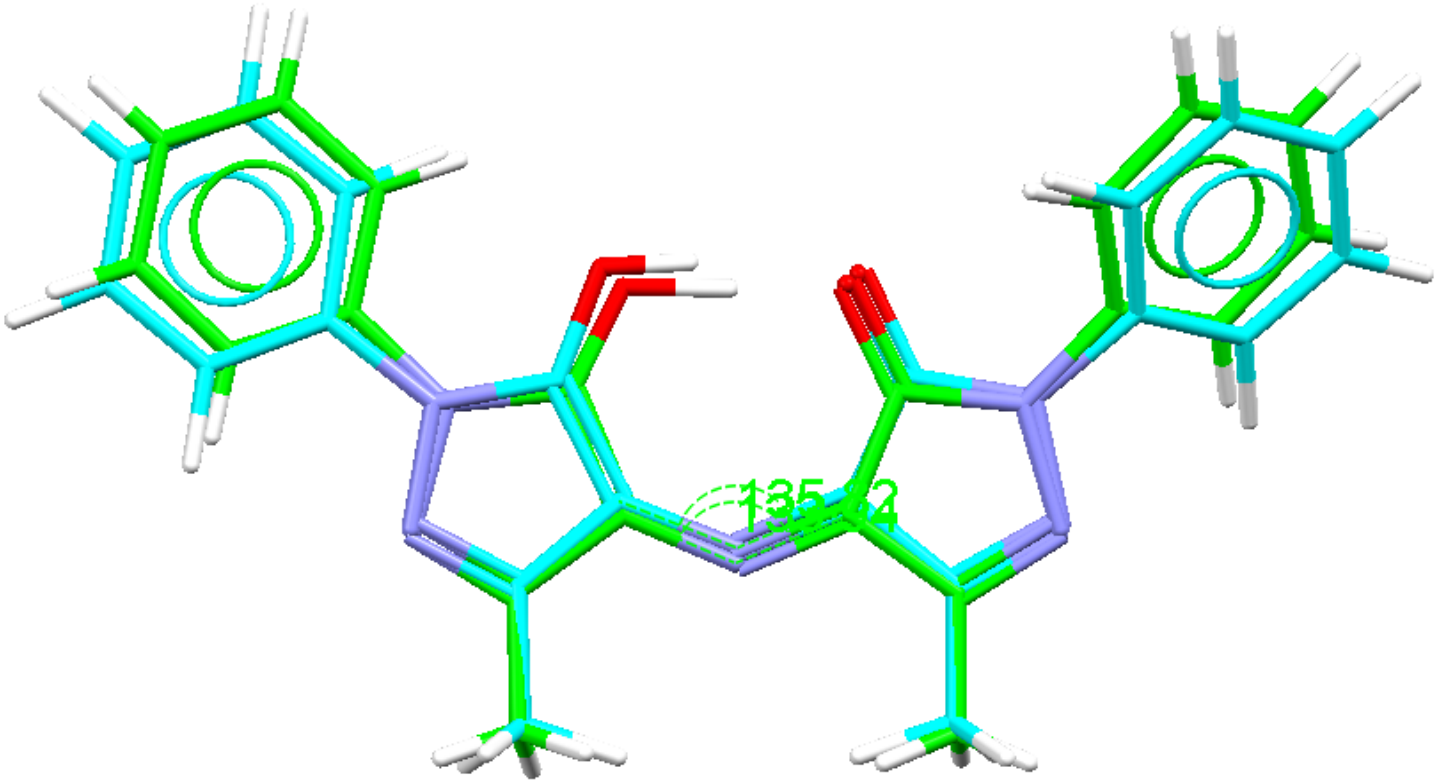
Minimisation

CEPMOA: CSD (green) vs Corina (light blue)



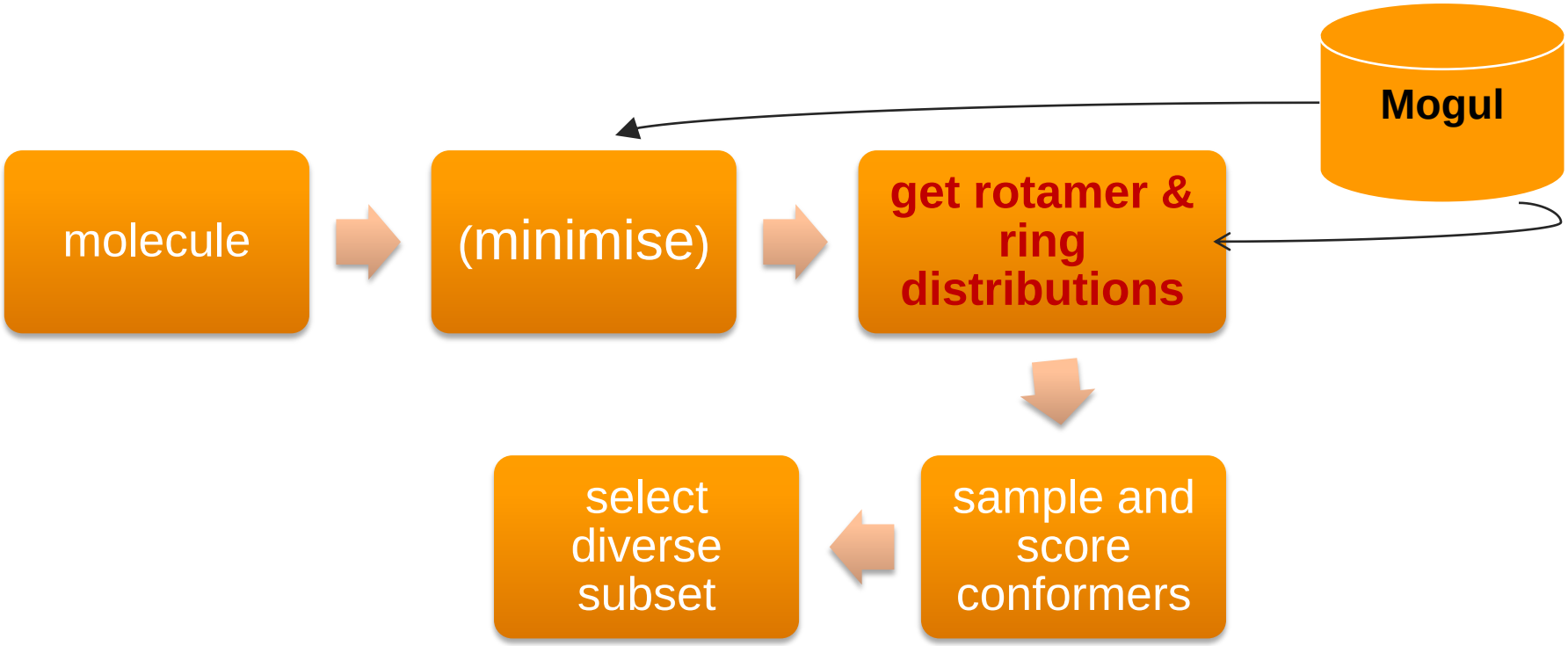
Minimisation

CEPMOA: CSD (green) vs Corina minimised (light blue)

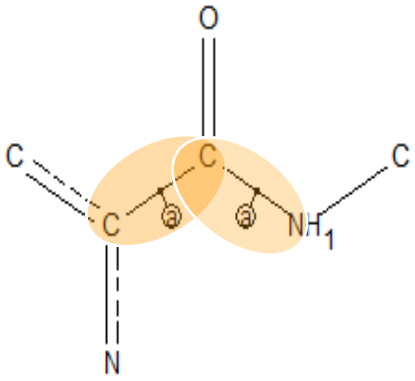


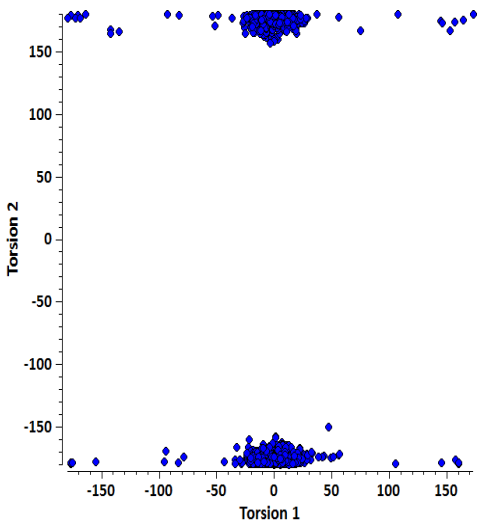
Minimisation to CSD-observed values in specific chemistry

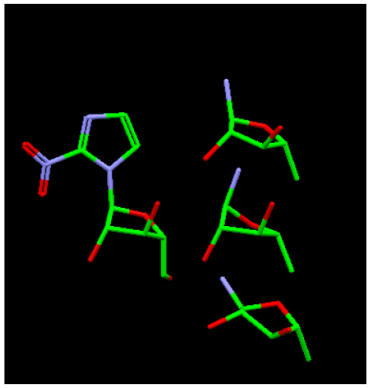
Conformer Generation Workflow

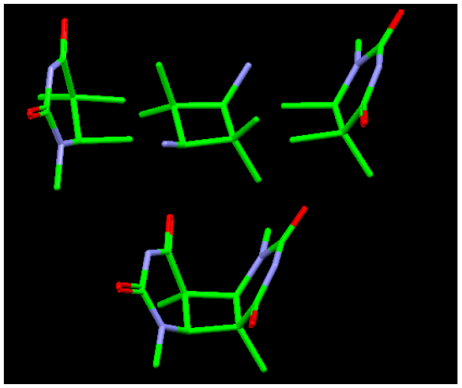


Structural Data Sources



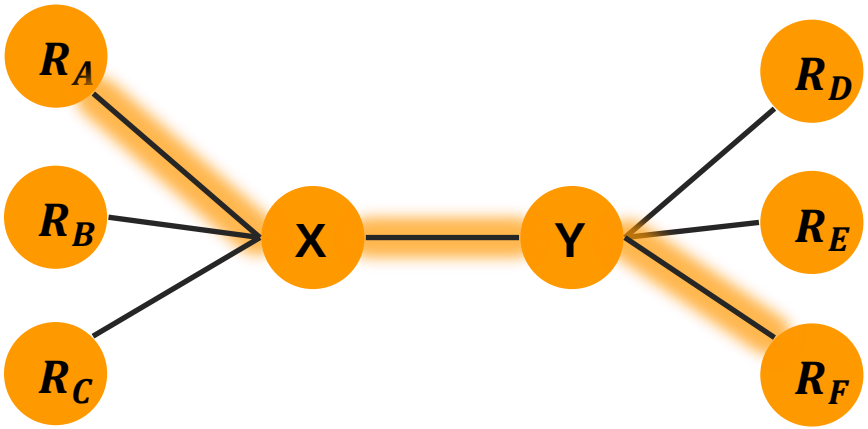






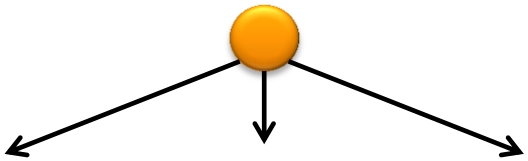
Ring geometries pre-clustered, taking symmetry and substituent atoms into account, and assigned probabilities

- Multi-dimensional user fragment libraries
- CSD-defined ring templates
- CSD-defined rotamers

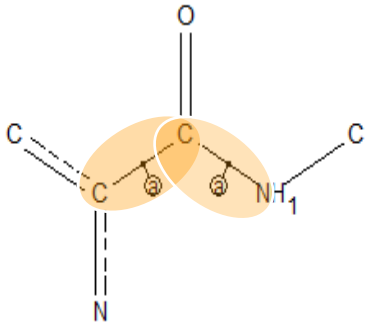


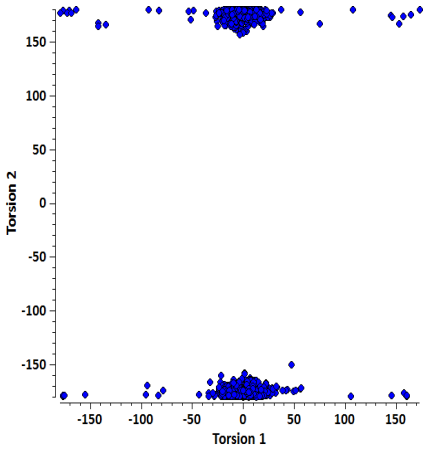
Applying Distributions

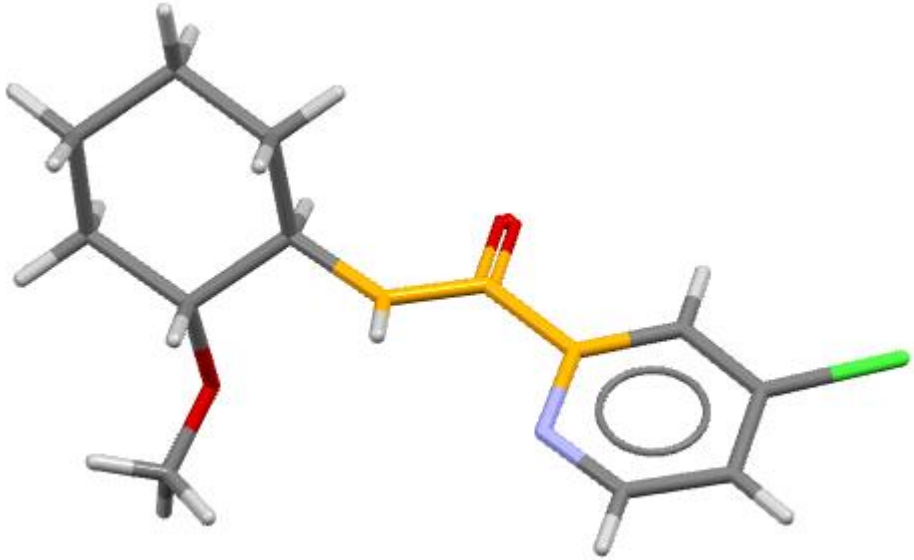
- building the tree



1. Match fragment library entries
2. Match CSD ring templates
3. Match remaining acyclic rotamers



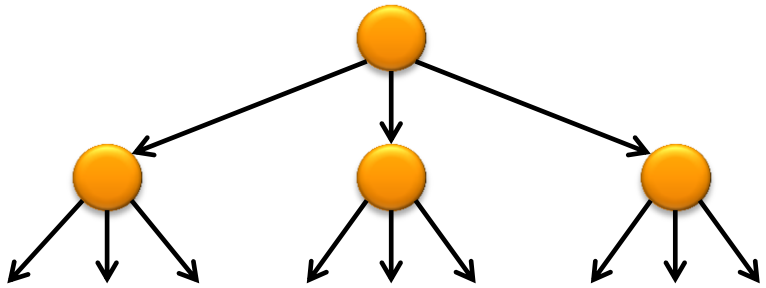


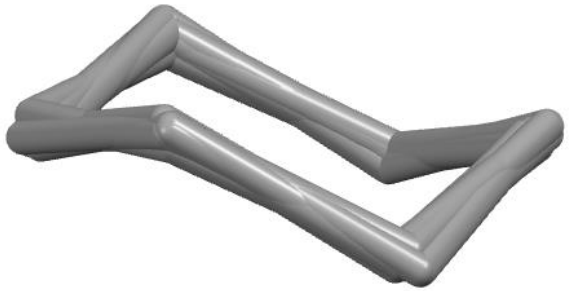


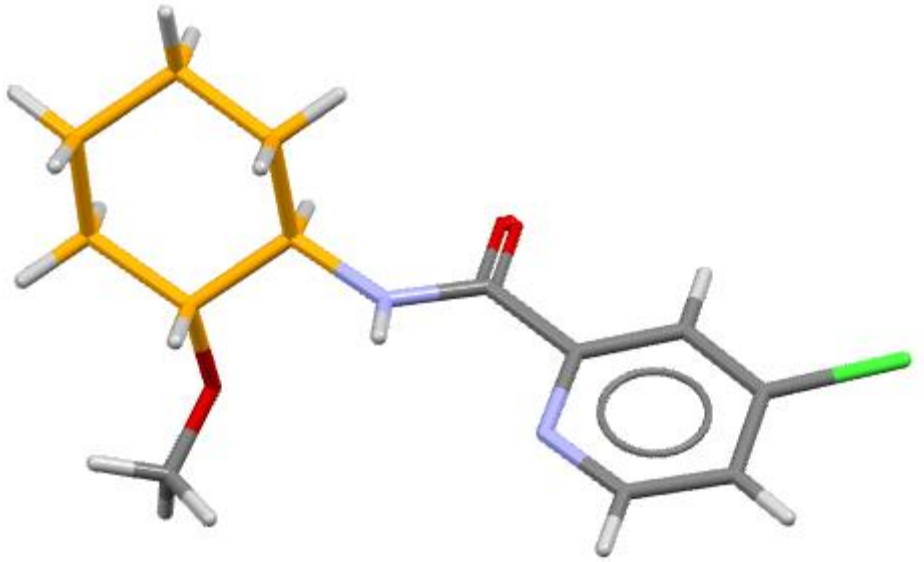
Applying Distributions

- building the tree

1. Match fragment library entries
2. Match CSD ring templates
3. Match remaining acyclic rotamers



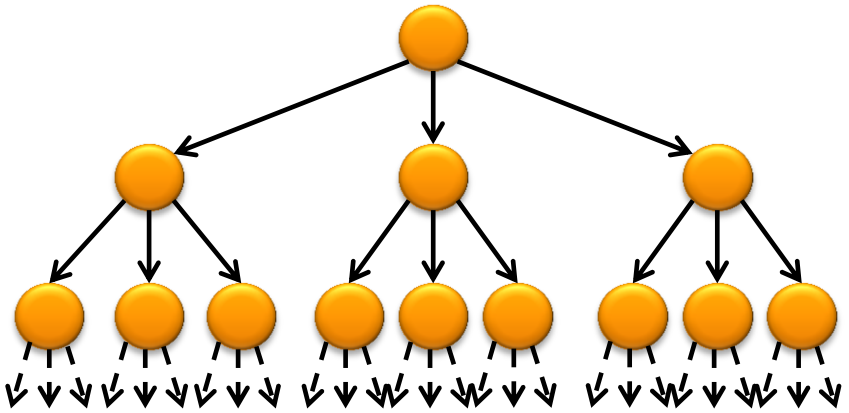


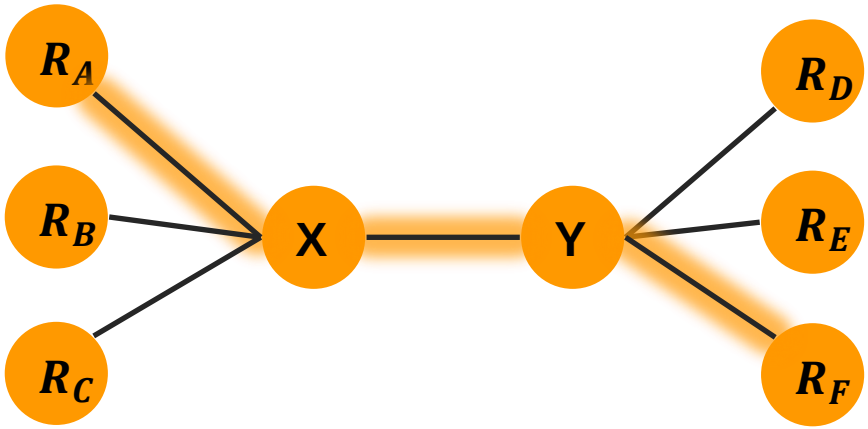


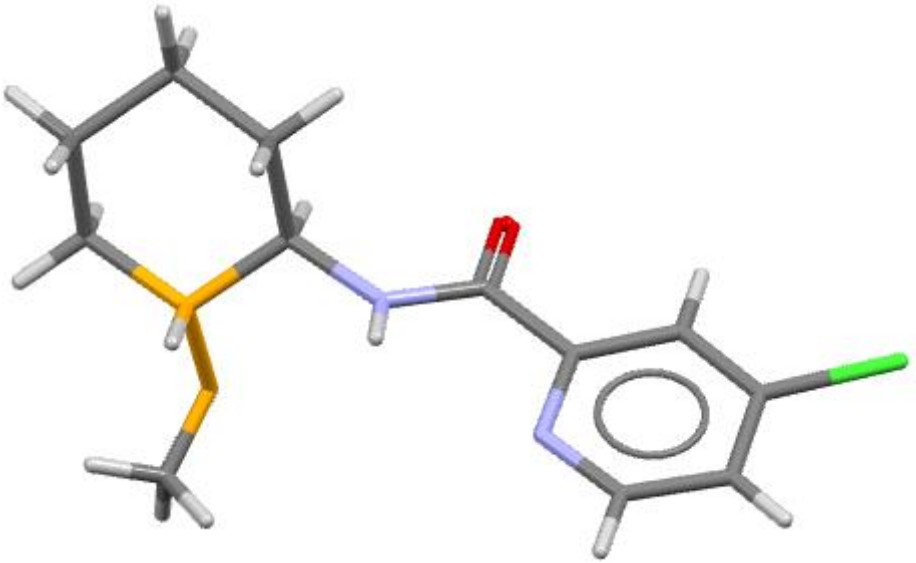
Applying Distributions

- building the tree

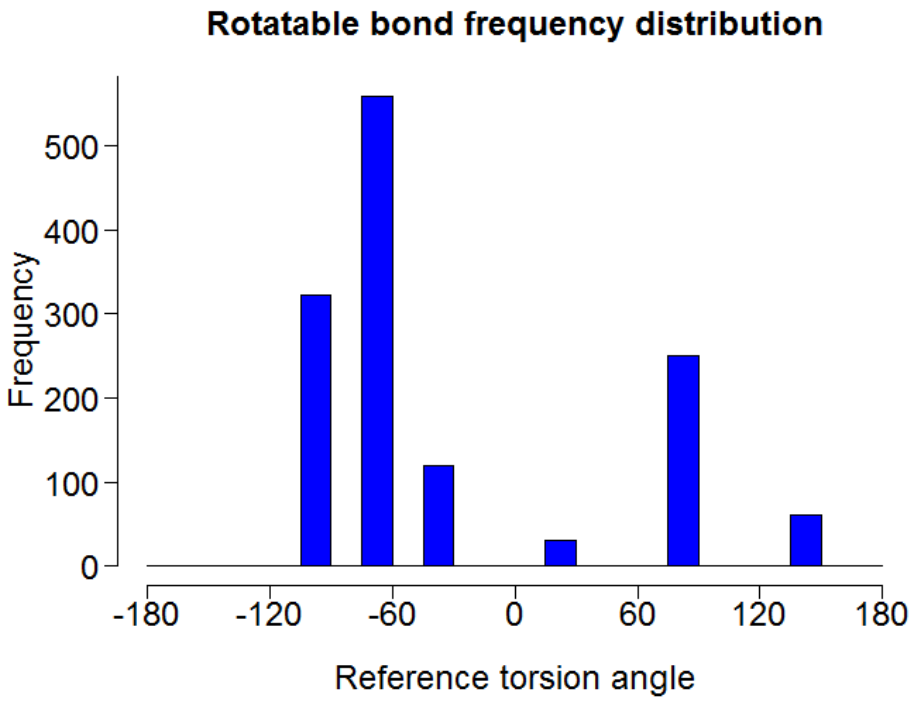
1. Match fragment library entries
2. Match CSD ring templates
3. Match remaining acyclic rotamers



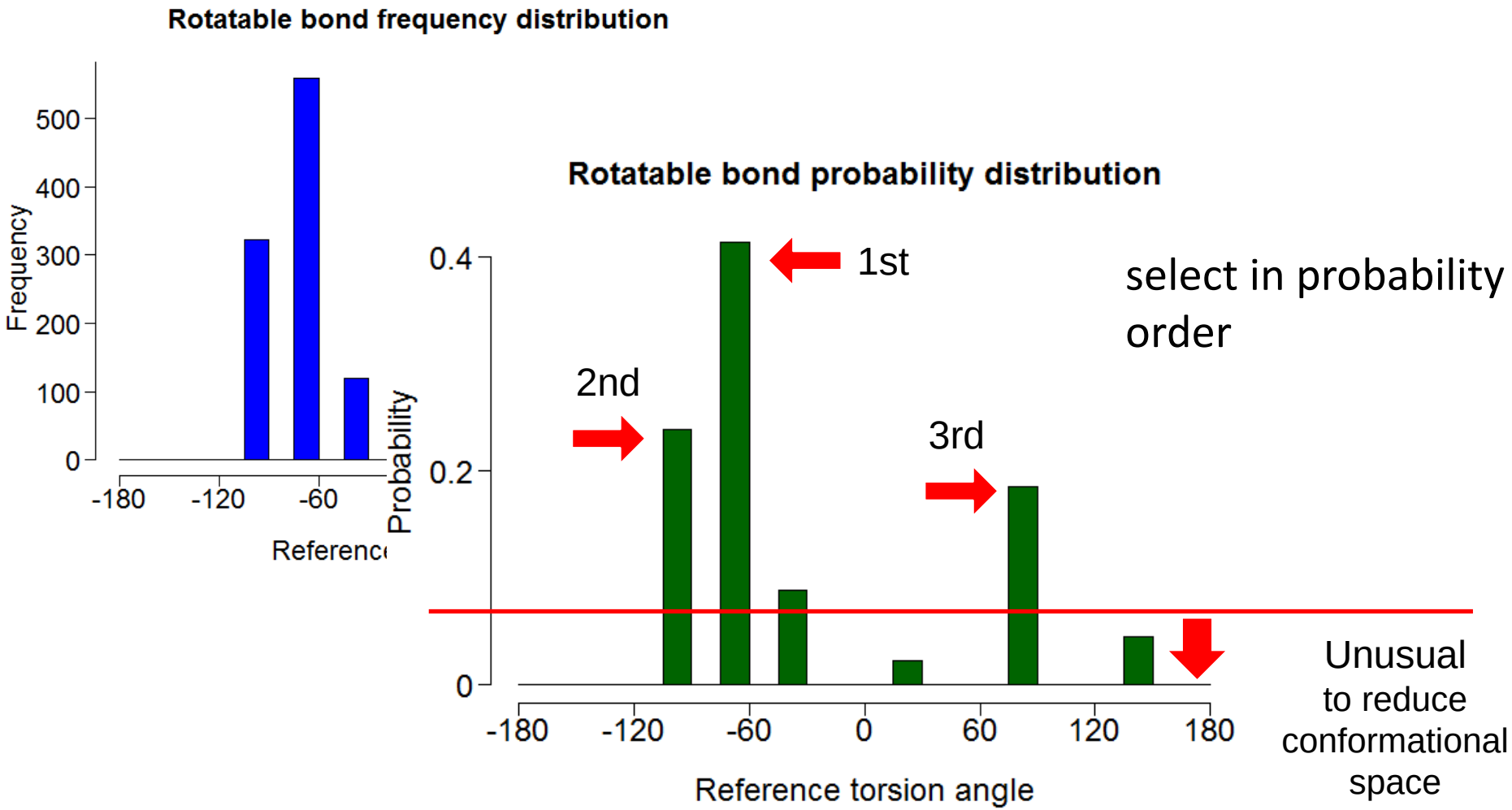




Applying Distributions – Rotamer Example



Applying Distributions – Rotamer Example



Scoring Conformations

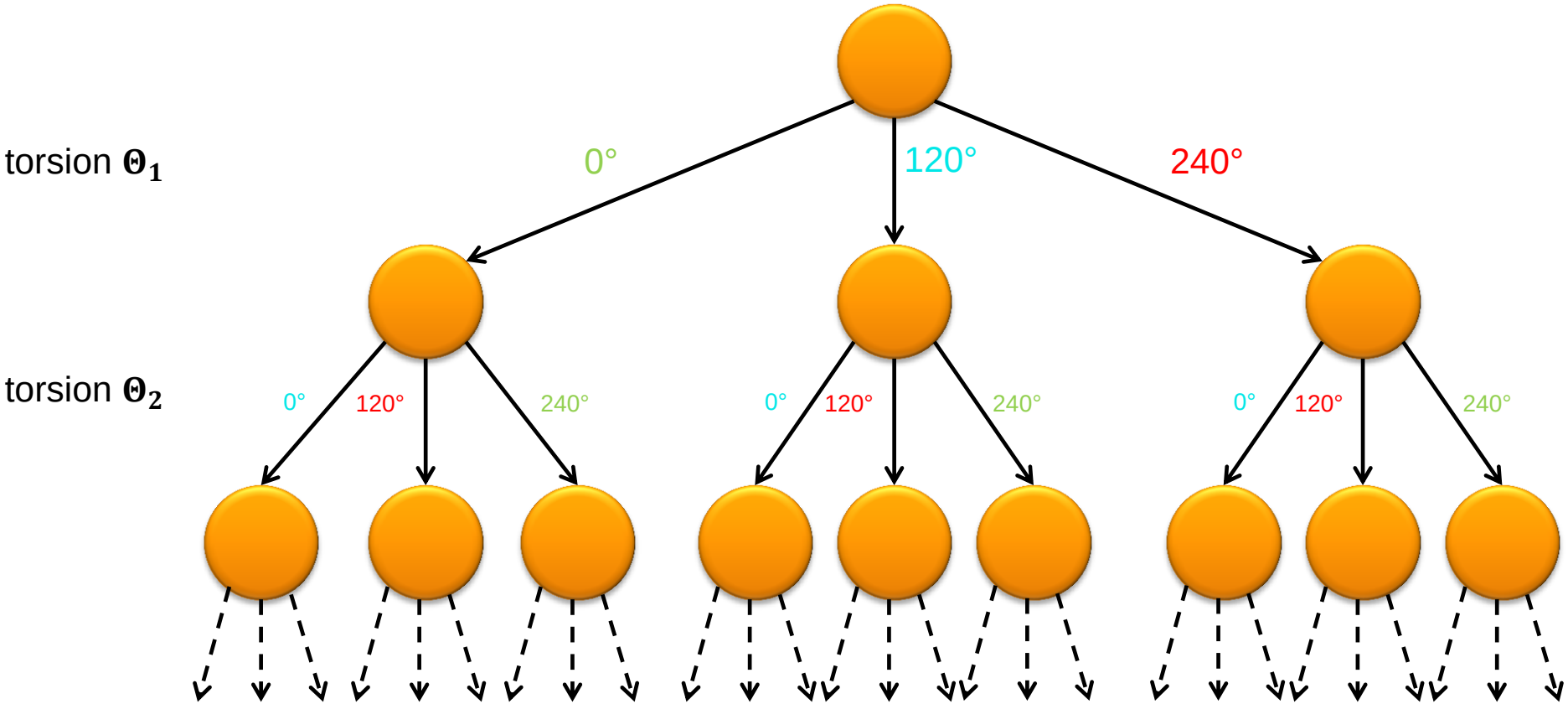
- Each rotamer angle, ring template instance or fragment configuration has an associated probability p_i
- calculate overall conformer score as

$$P = \prod_i p_i$$

$$lnP = \sum_i ln\ p_i$$

Searching

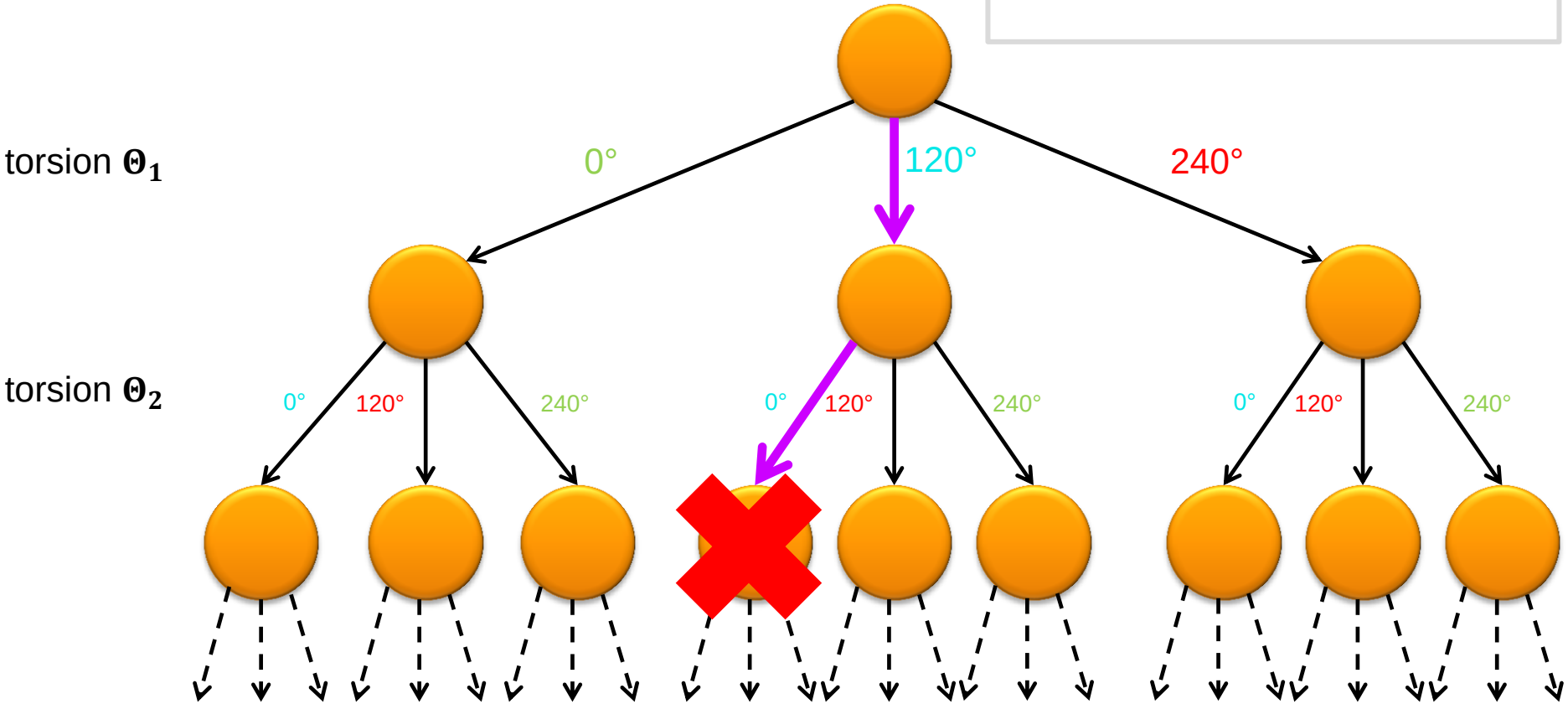
p=0.7
p=0.25
p=0.05 (unusual)



Limiting Conformational Space

unusual torsions

p=0.7
p=0.25
p=0.05 (unusual)

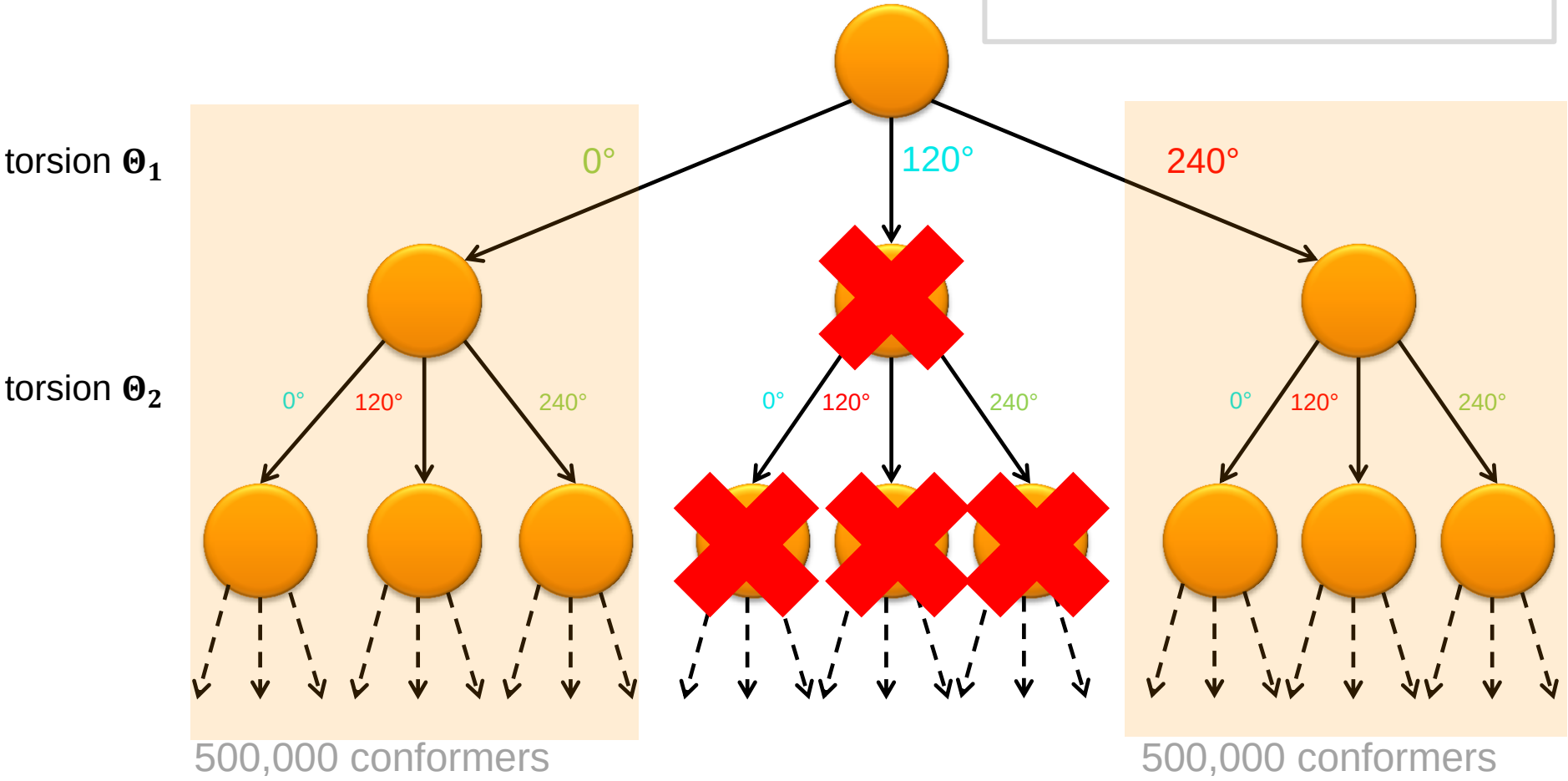


assuming one unusual torsion is allowed

Limiting Conformational Space

number of conformations

p=0.7
p=0.25
p=0.05 (unusual)



Limiting Conformational Space

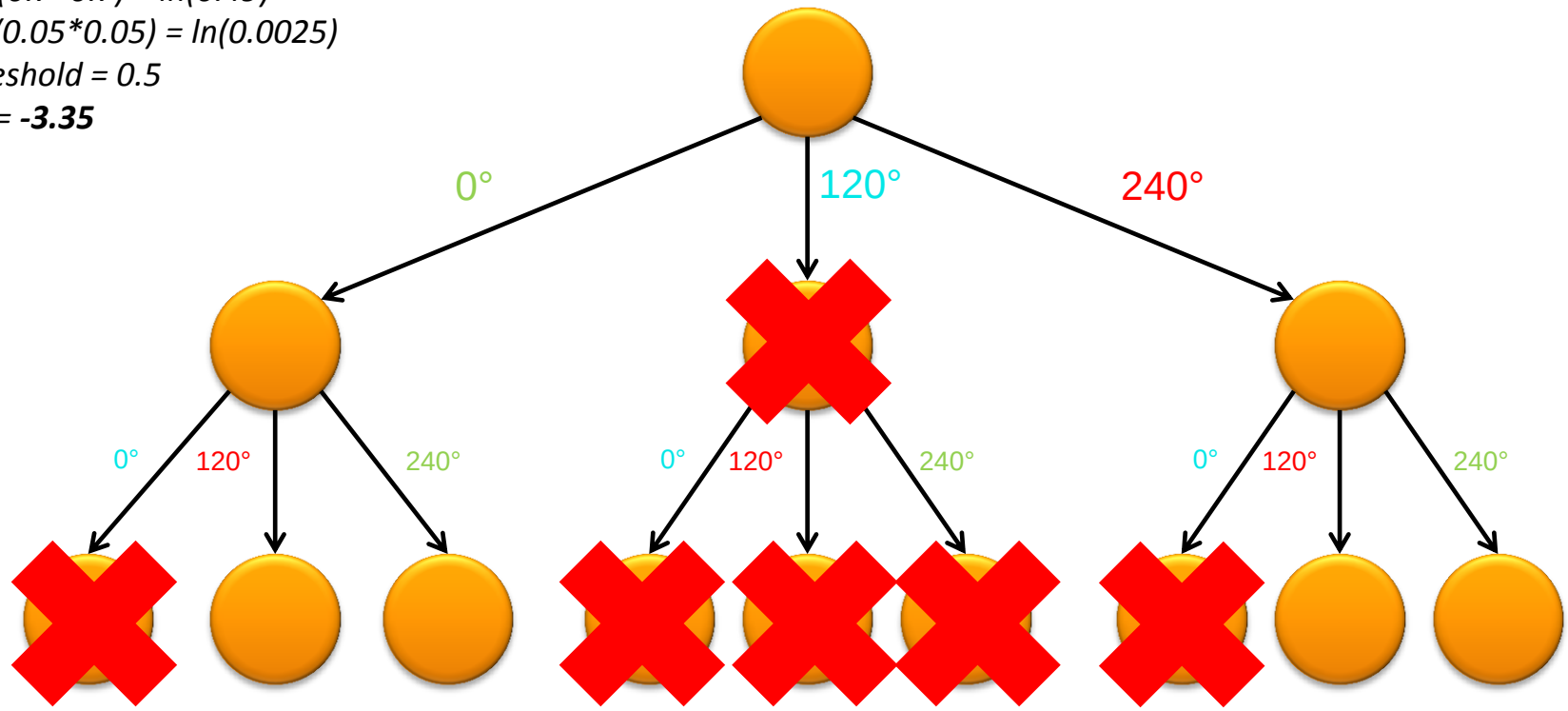
probability threshold

p=0.7	*ln(0.7)*	*= -0.357*
p=0.25	*ln(0.25)*	*= -1.386*
p=0.05	*ln(0.05)*	*= -3*

$ln(P_{threshold}) = ln(P_{max})\text{-}threshold*(ln(P_{max})\text{-}ln(P_{min}))$
$ln(P_{max}) = ln(0.7*0.7) = ln(0.49)$
$ln(P_{min}) = ln(0.05*0.05) = ln(0.0025)$
assume threshold = 0.5
$ln(P_{threshold}) =$ ***-3.35***

torsion Θ_1

torsion Θ_2

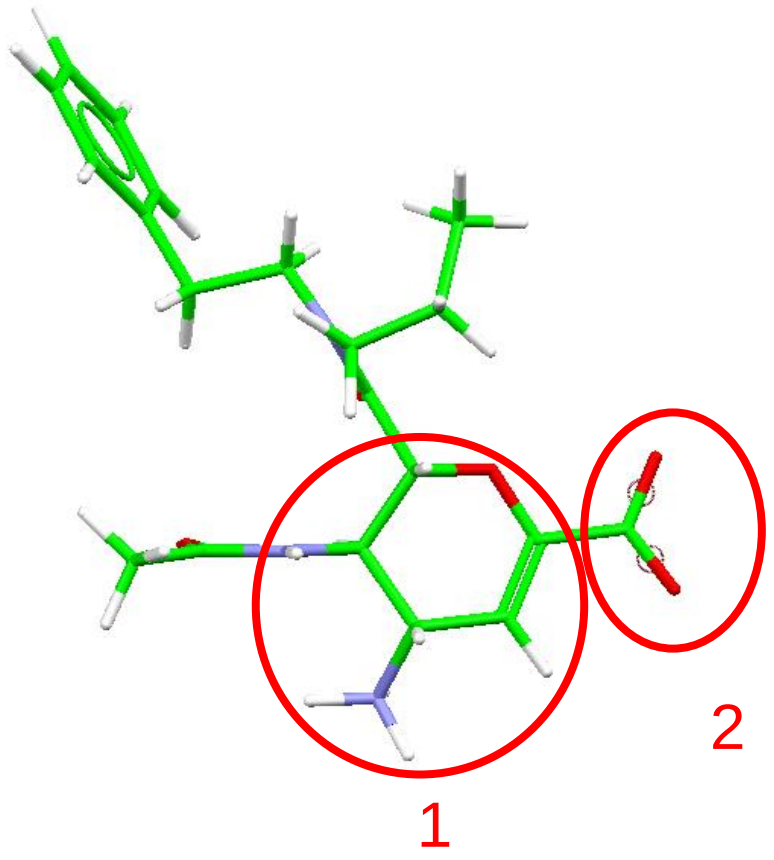


any conformers with ln-scores lower than ***-3.35*** *will be eliminated*

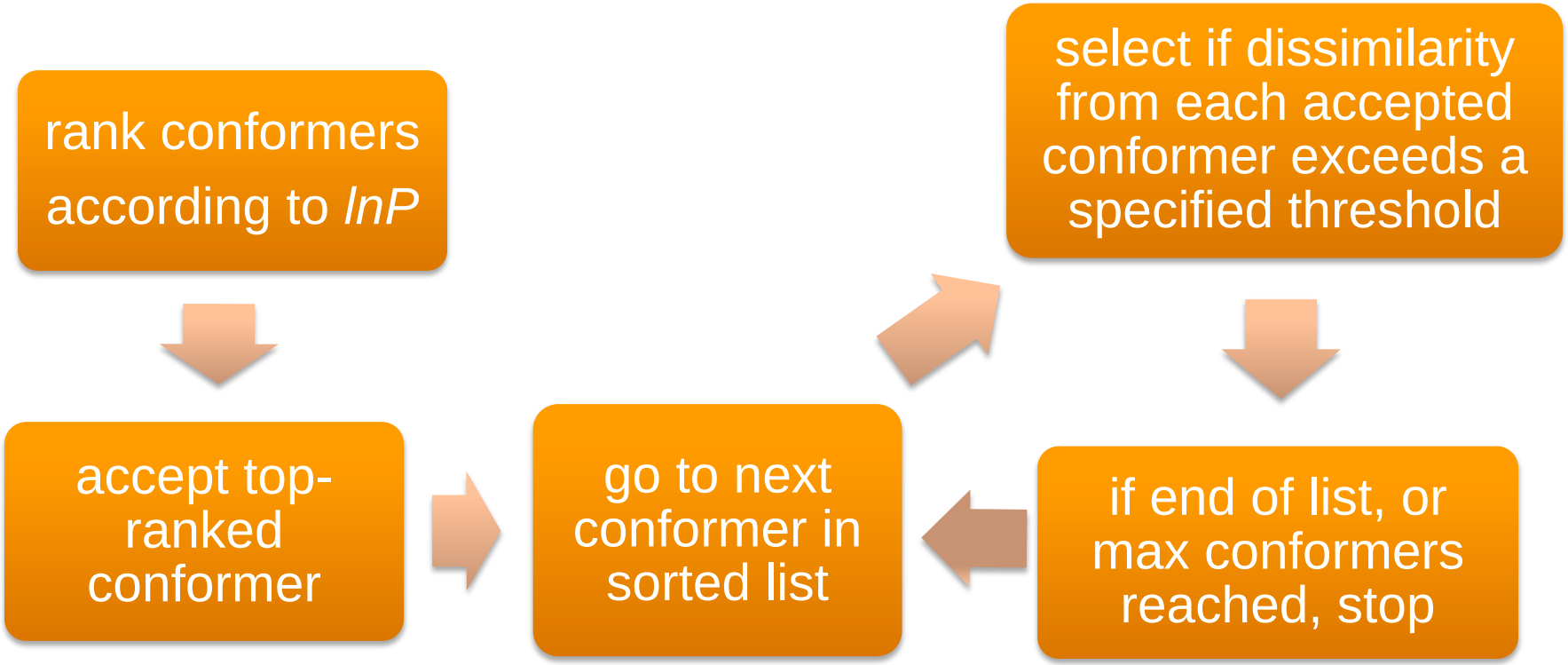
Limiting Conformational Space

close contacts

- uses a simple clash function
- discard if sum of contributions greater than user-defined limit
- use bounding volume approach
  1. if bounding volumes don't clash, atoms won't either
  2. if bounding volumes clash, check atoms in bounding volume 2 against bounding volume 1
  3. if atom clashes with bounding volume, check against all atoms



Selecting a Diverse Subset



Conformer dissimilarity

- standard measure is non-hydrogen atom *rmsd* but this is too slow to calculate
- torsion dissimilarity (similar to[1]) used as pre-screen:

$$\sqrt{\sum \#atoms_i(\tau_{i1} - \tau_{i2})^2}$$

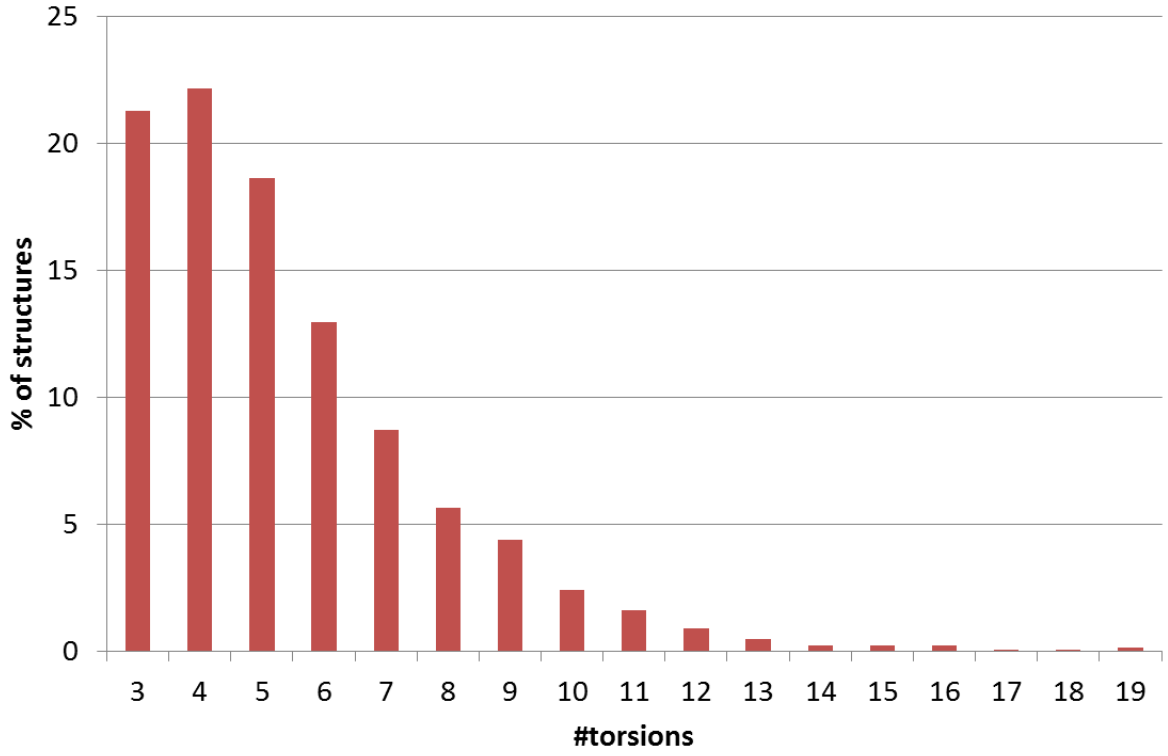
- if torsion dissimilarity >100, conformers deemed dissimilar and both accepted for final solution set
- if torsion dissimilarity <100, decision based on atom *rmsd*
- < 4% conformer pairs have torsion *rmsd* > 100, atom *rmsd* < 0.5

[1]Schulz-Gasch, T.; Schärfer, C.; Guba, W.; Rarey, M. TFD: Torsion Fingerprints As a New Measure To Compare Small Molecule Conformations. *J. Chem. Inf. Model.* 2012, *52*, 1499–1512

VALIDATION OF THE GENERATOR

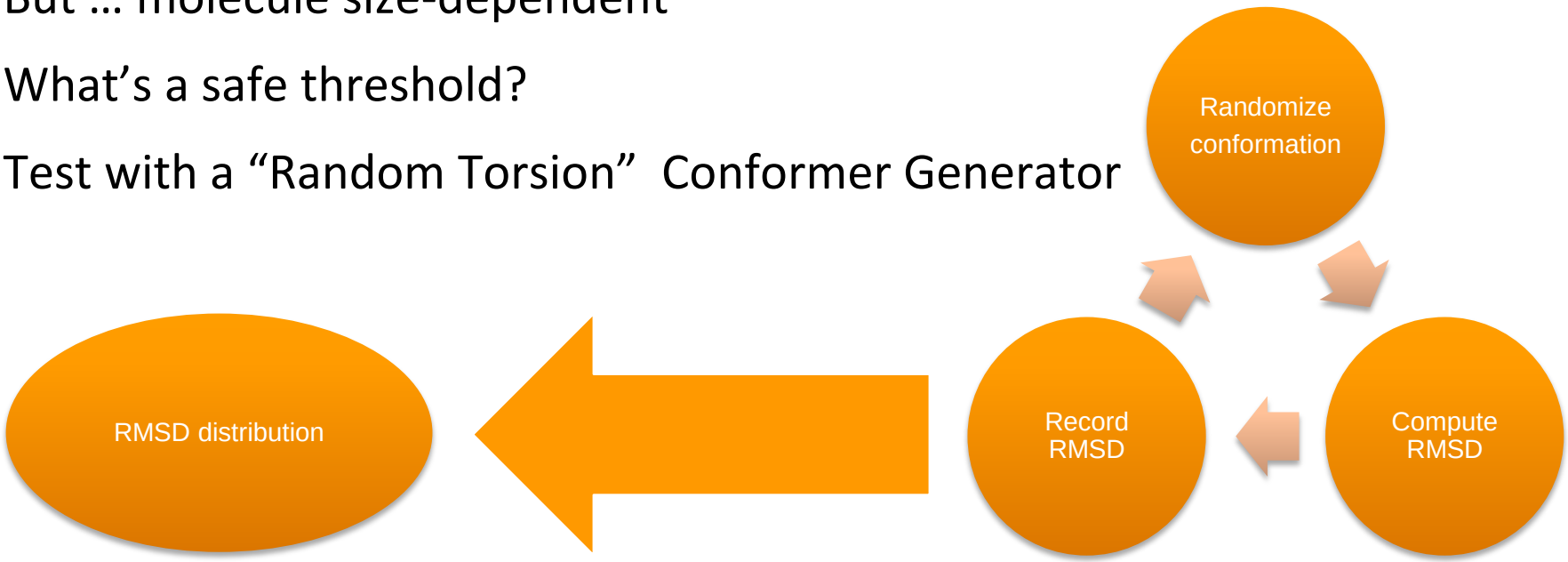
Test Set Description

- 3291 Corina structures based on CSD structures
- All excluded from geometry data libraries
- Average of 5-6 rotatable bonds per structure



Problem: What's a good definition of success?

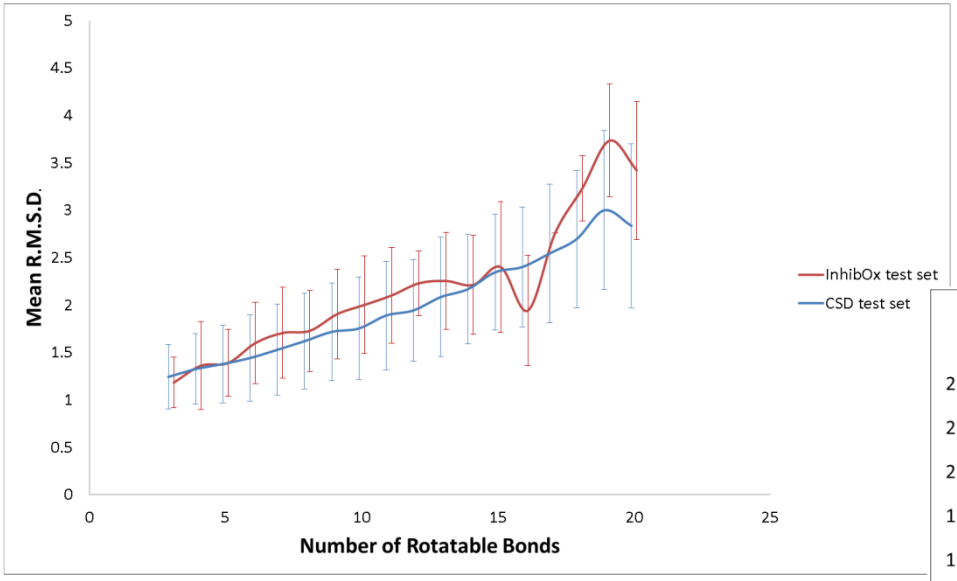
- Root Mean Square Deviation (RMSD) of atomic coordinates is convenient measure of conformer similarity
- But … molecule size-dependent[1]
- What's a safe threshold?
- Test with a "Random Torsion" Conformer Generator

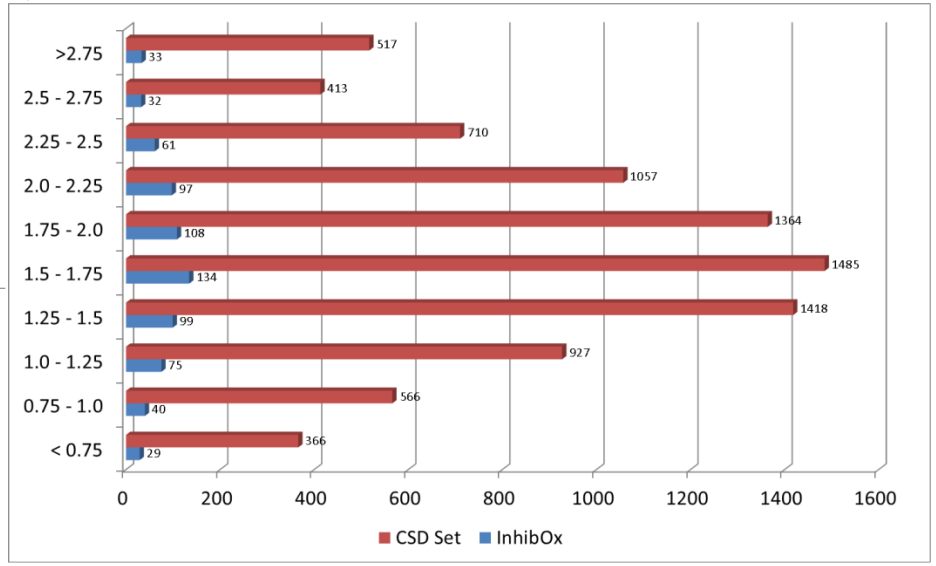


1. See e.g. Paul C.D. Hawkins *et al J. Chem. Inf. Model.*, **2012**, *52* (11), pp 2919–2936
DOI: 10.1021/ci300314k

Random Sampling RMSDs

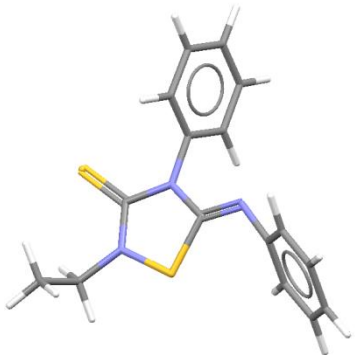
Random sampling can achieve RMSDs of 1.25-1.5 Å for more rigid molecules



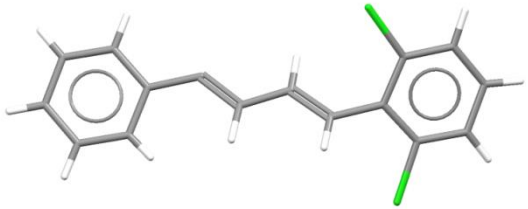


Extreme Examples

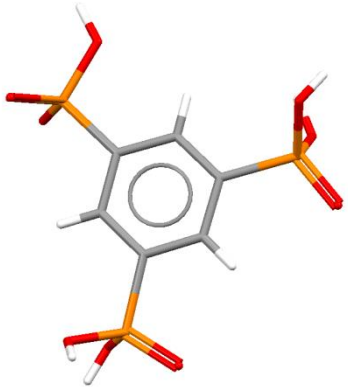
- Mean RMSD < 1.0



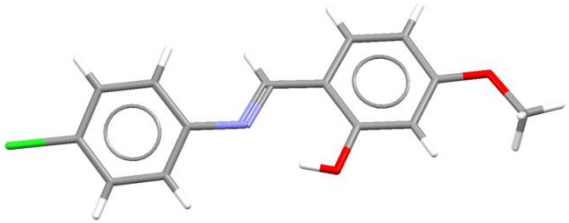
DUDPAT



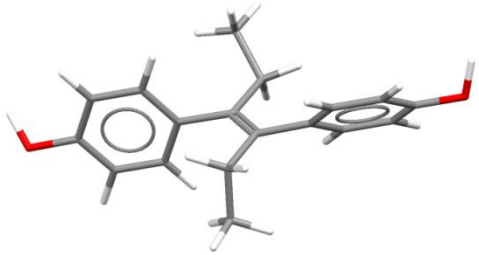
CLPTBU



LACVUH



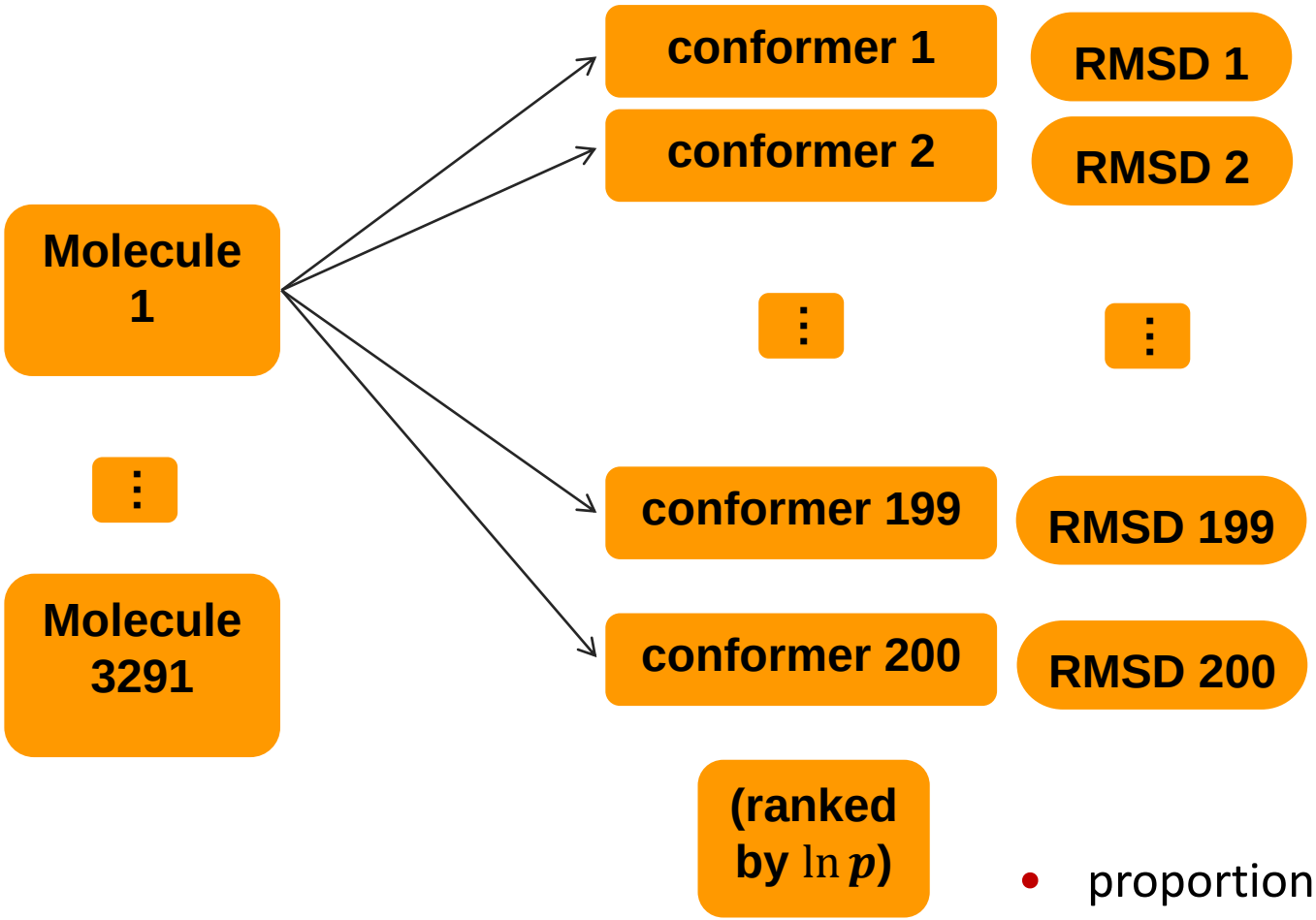
AJITOD



DES in PDB entry 1tz8

- Conclusion: play it safe: use 0.5Å RMSD as cutoff for right/wrong

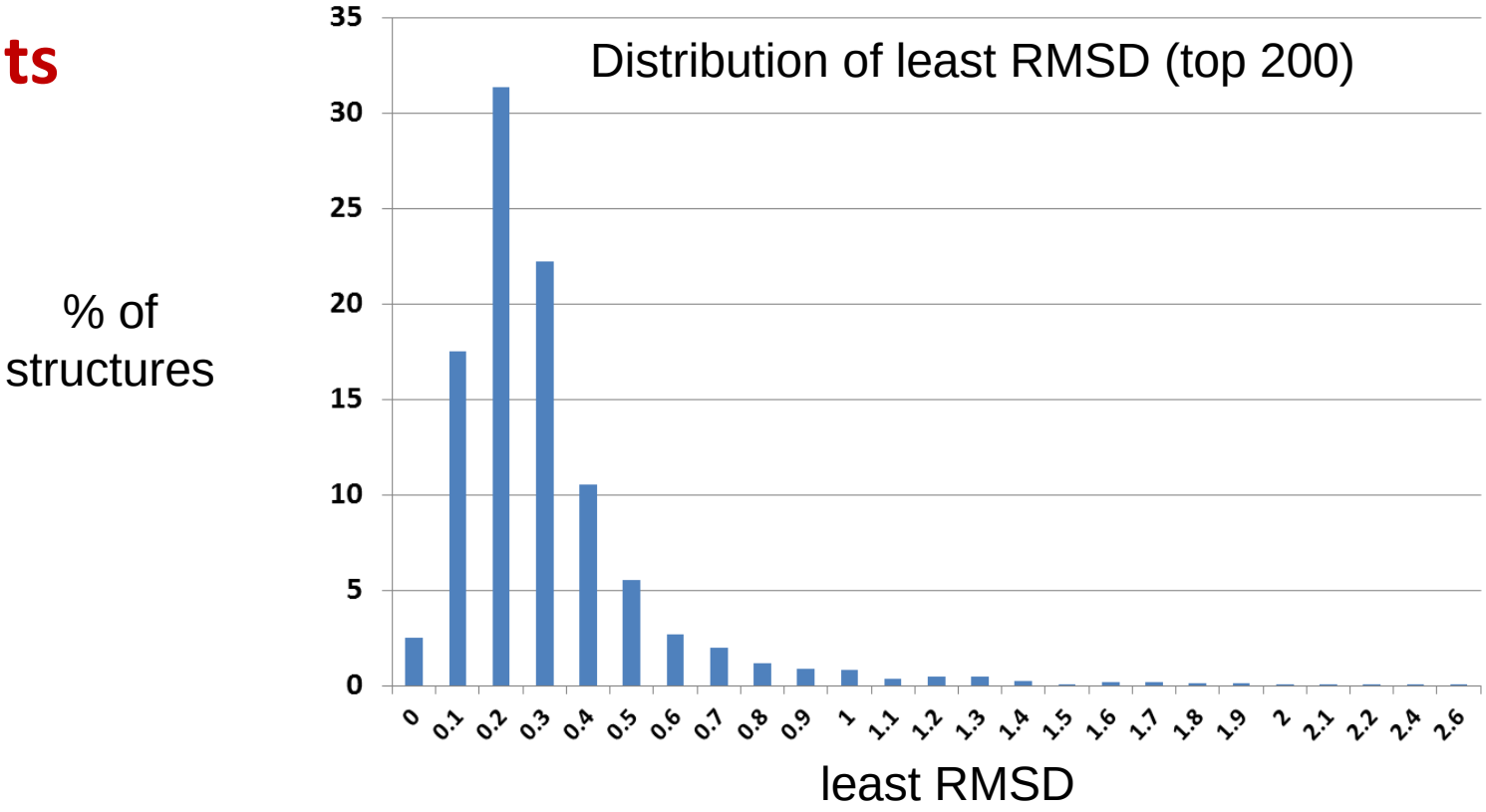
Result Generation



- Average of least RMSD?
- proportion least RMSD< 0.5Å?

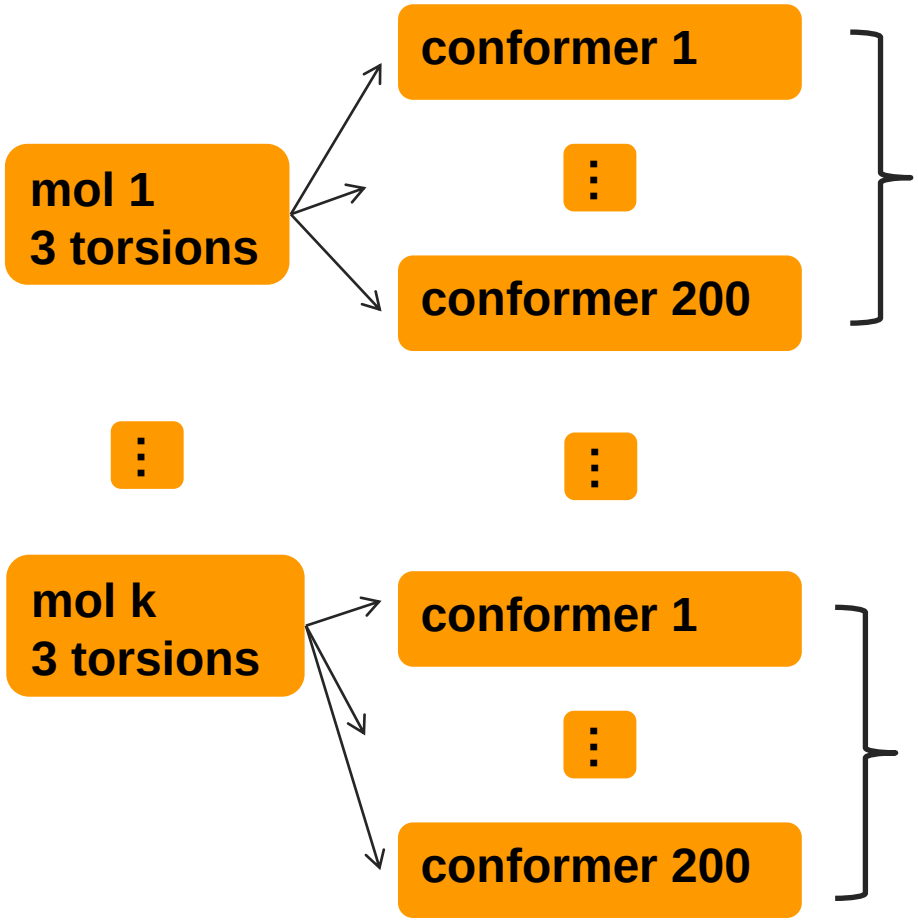
In top (1,50,100,200)

Results



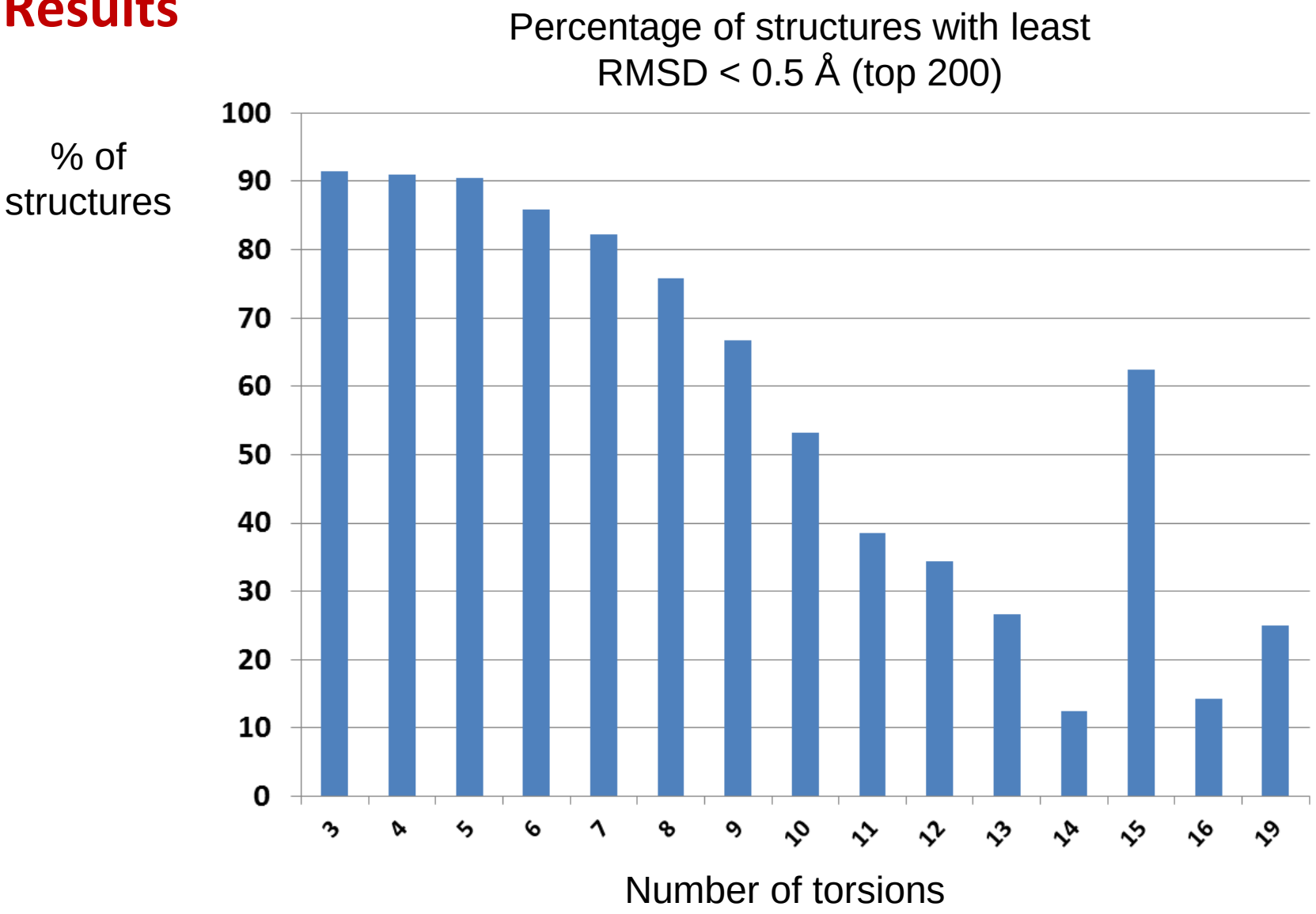
	Top n conformers				
n	1	25	50	100	200
least RMSD < 0.5 Å (%)	24	73	78	82	84
< least RMSD >	1.12	0.45	0.41	0.38	0.36

Results by Number of Torsions



- Structures with 3 torsions
- Repeat for structures with 4,5… torsions

Results


Percentage of structures with least RMSD < 0.5 Å (top 200)
% of structures
100
90
80
70
60
50
40
30
20
10
0
3
4
5
6
7
8
9
10
11
12
13
14
15
16
19
Number of torsions

Conclusions

- Using CSD data directly for conformer generation is successful
- Pre-minimisation aids performance
- This tool will improve as the CSD increases in size (particularly for rarer chemistries)
- Where Next: Easily facilitate use of user's in-house structural data to augment Mogul libraries
  - Extend coverage of company-relevant conformational data

Acknowledgements

- Robin Taylor
- Oliver Korb
- Patrick McCabe
- John Liebescheutz

And you, for your attention!