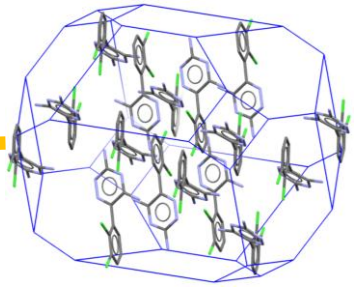
Creating a conformer validation test set using RDKit and the Cambridge Structural Database

Tjelvar Olsson and Jason Cole

The Cambridge Crystallographic Data Centre

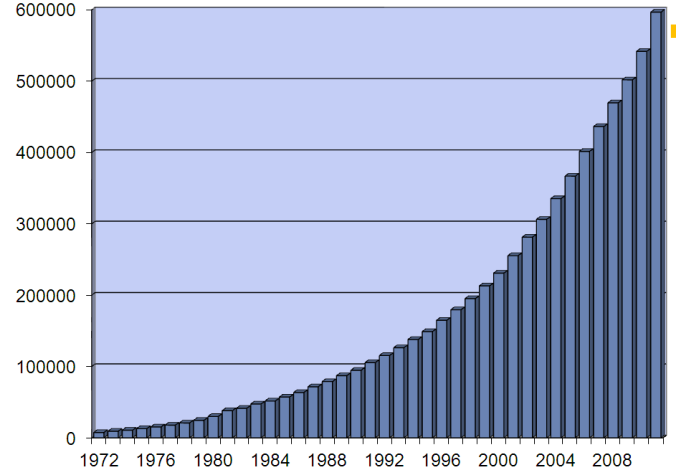
- A non-profit, charitable institution
- Self financing and self administering
- 43 employees
- Recognised institute for postgraduate degrees of the University of Cambridge
- Objectives
  - “advancement and promotion of the science of chemistry and crystallography for the public benefit”

Cambridge Structural Database System

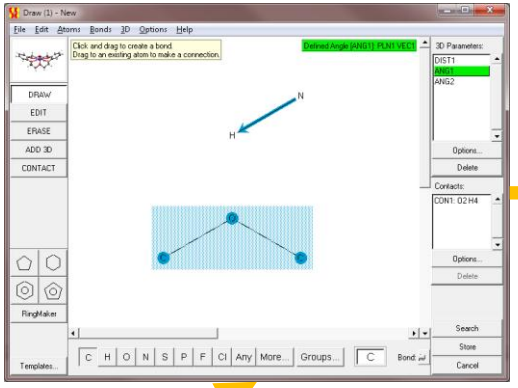


PreQuest: Create (in-house) database

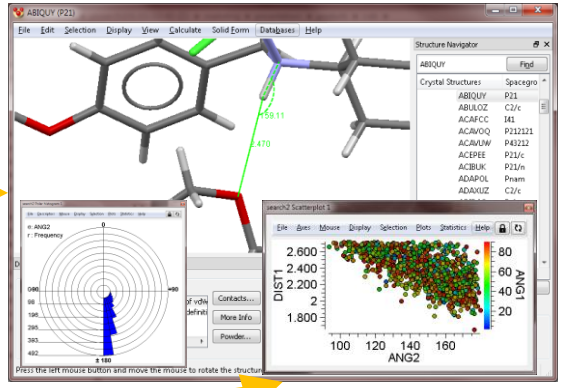
Cambridge Structural Database



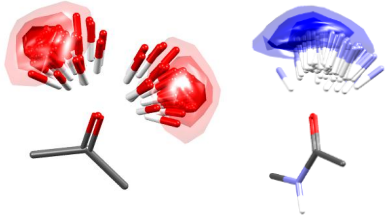
ConQuest: Advanced 3D searching



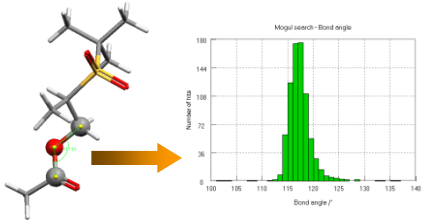
Mercury: Visualisation & data analysis



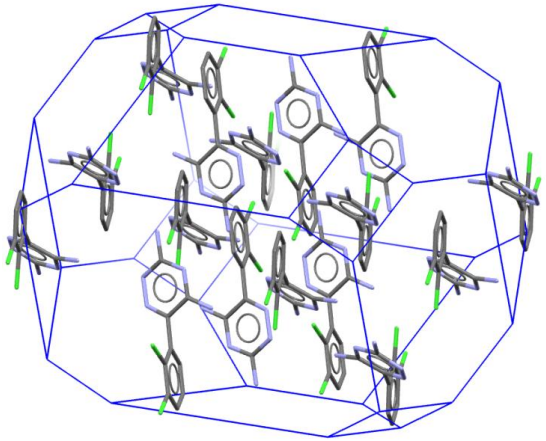
IsoStar: Molecular interaction analysis

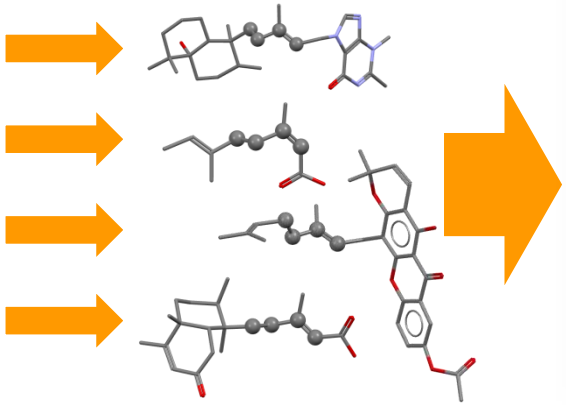


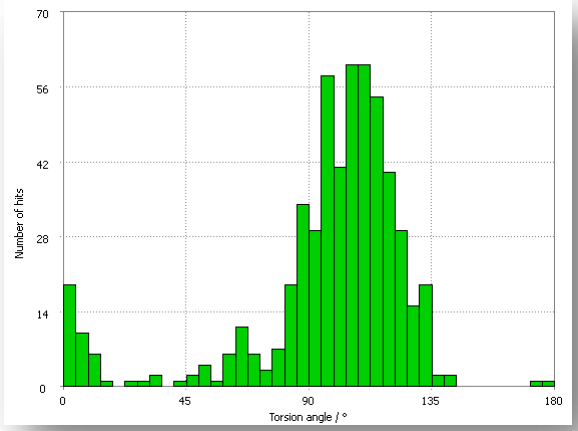
Mogul: Molecular geometry analysis

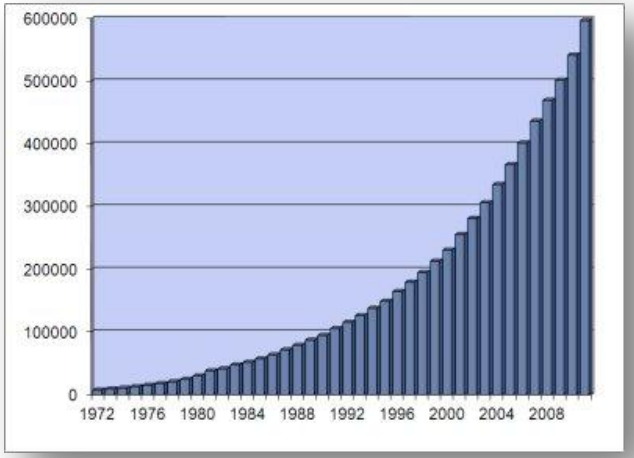


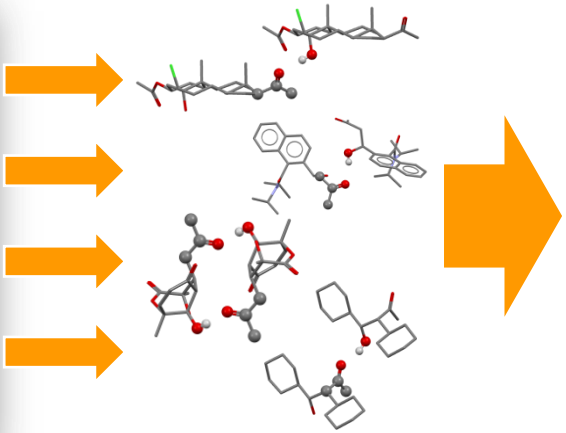
Cambridge Structural Database System



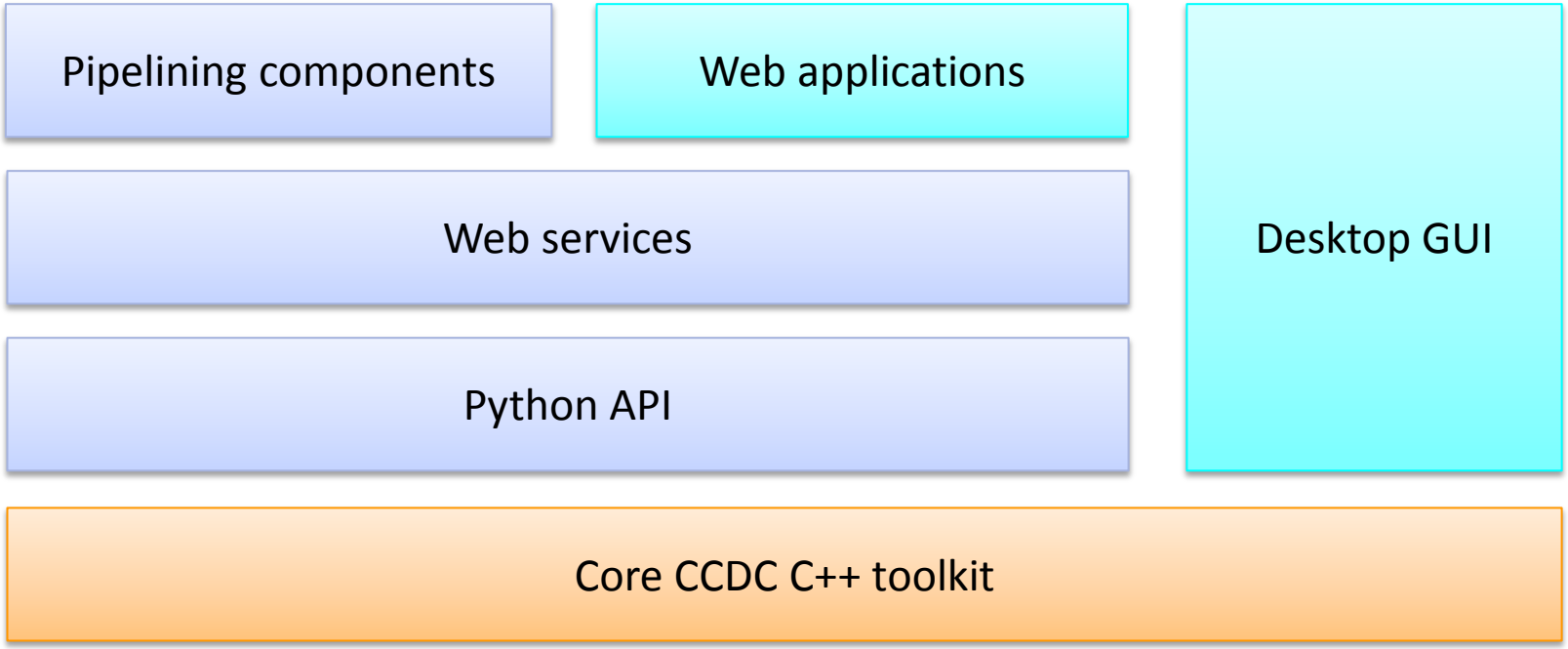






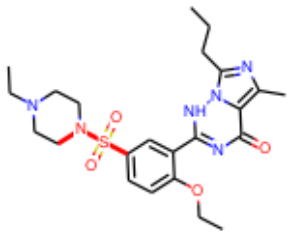
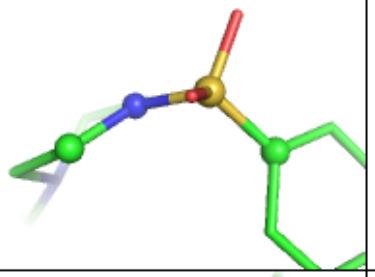
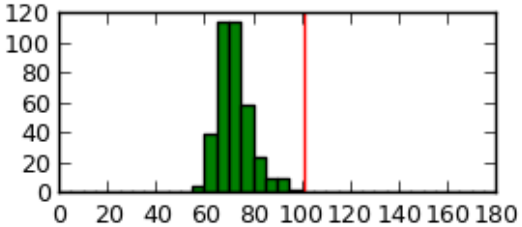
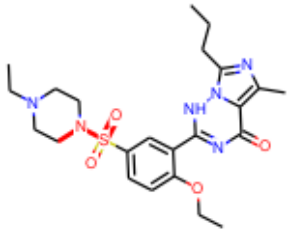
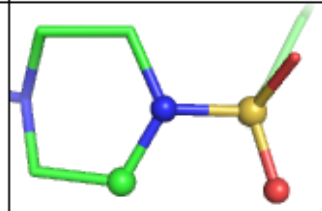
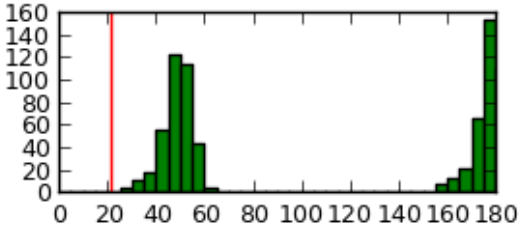


Recent developments: APIs



Example use of the API: Geometry report

Integration with 3rd party Python packages

			• Fragment: C7 S10 N14 C19 • Query: 101.29 • Num hits: 376 • **Local density: 0.53** • d(min): 6.74
			• Fragment: O11 S10 N14 C19 • Query: 21.65 • Num hits: 752 • **Local density: 1.73** • d(min): 0.12

RDKit
Open-Source Cheminformatics and Machine Learning

PyMOL

matplotlib

Torsion preferences from CSD data

Local similarity

Use environment of atoms in torsion and connected atoms:

- ring environment
- bond types
- atomic number
- number of connected atoms
- nature of connected atoms
  - hydrogen count
  - metal/non-metal

Automated classification

Bruno et al., J. Chem. Inf. Comput. Sci., 44, 2133-2144, 2004

Substructure rules

```
[*:1]~[CX3:2]!@[NX3:3]~[*:4]
[*:1]~[NX3:2]!@[NX2:3]~[*:4]
[*:1]~[NX3:2]!@[NX3:3]~[*:4]
[*:1]~[cX3:2]!@[NX2:3]~[*:4]
[*:1]~[CX4:2]!@[NX2:3]~[*:4]
[*:1]~[OX2:2]!@[P:3]~[*:4]
[*:1]~[CX4:2]!@[P:3]~[*:4]
…
[*:1]~[CX4:2]!@[CX3:3]~[*:4]
```

Hand curated classification

Schärfer et al., J. Med. Chem., 56, 2016-2028, 2013

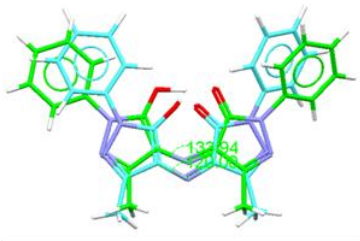
Automating the SMARTS searches using the Python API

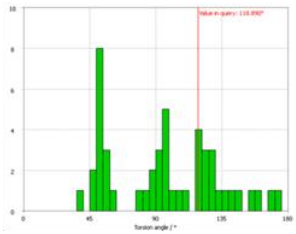
```
##########################################################

#   Top ten SMARTS patterns

    patts = [
        '[*:1]~[CX3:2]!@[NX3:3]~[*:4]',
        '[*:1]~[NX3:2]!@[NX2:3]~[*:4]',
        '[*:1]~[NX3:2]!@[NX3:3]~[*:4]',
        '[cH1:1]~[a:2]([cH1])!@[a:3]([s,o,n:4])
        '[*:1]~[cX3:2]!@[NX2:3]~[*:4]',
        '[*:1]~[CX4:2]!@[NX2:3]~[*:4]',
        '[cH1:1]~[a:2]([cH1])!@[a:3]([cHO])[cHO
        '[*:1]~[OX2:2]!@[P:3]~[*:4]',
        '[*:1]~[CX4:2]!@[P:3]~[*:4]',
        '[*:1]~[CX4:2]!@[CX3:3]~[*:4]',
    ]

##############################################
```

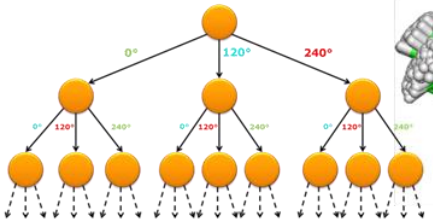
```
#################################################################

#   Do the search, and draw the histograms

    hist_filenames = []
    for i, p in enumerate(patts):
        t = time.time()
        q = SMARTSSubstructure(p)
        s = SubstructureSearch()
        s.add_substructure(q)
        s.add_torsion_angle_measurement('Torsion',
            0, q.label_to_atom_index(1),
            0, q.label_to_atom_index(2),
            0, q.label_to_atom_index(3),
            0, q.label_to_atom_index(4)
        )
        hits = s.search(max_hit_structures=args.max_hits,
                        max_hits_per_structure=1)
        hist_fn = os.path.join(args.out_dir, 'torsion_%02d.png' % i)
        hist_filenames.append(hist_fn)
        hist = Histogram(title=p, xlabel='Torsion', ylabel='Count',
                         file_name=hist_fn)
        xs = [abs(h.measurements['Torsion']) for h in hits]
        hist.add_plot(xs, color='blue', bins=18)
        hist.write()

#################################################################
```

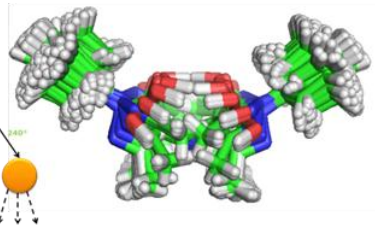
Recent developments: Conformer generator

- Use CSD bond lengths and angles to optimise 3D conformation
- Use CSD torsions to produce multiple conformations





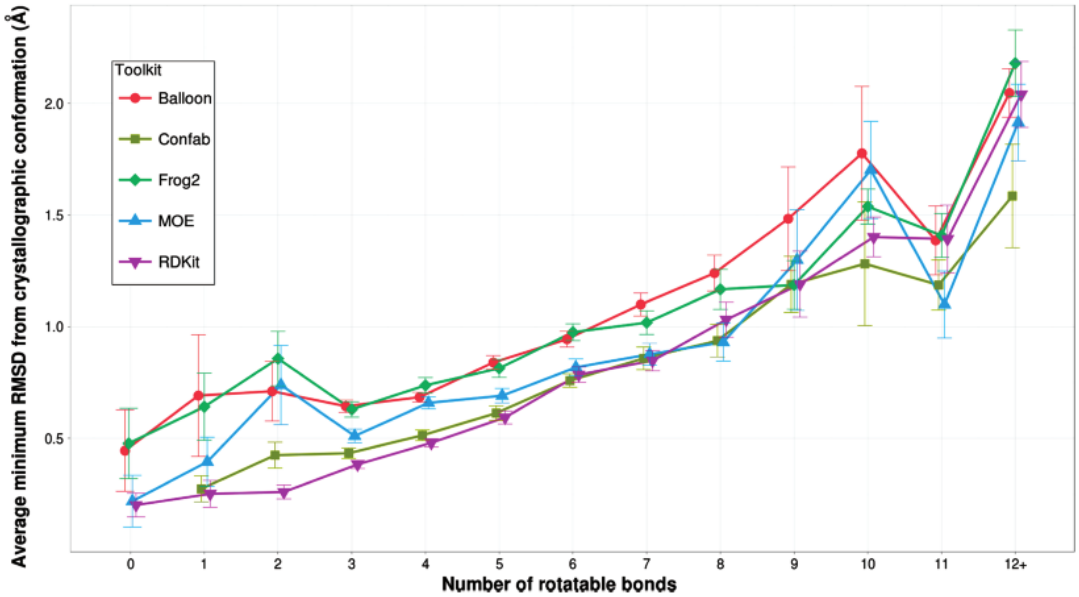




molecule → (optimise) → get rotamer and ring distributions → sample and score conformers → select diverse subset

InhibOx conformer generation test set

- 708 drug-like molecules from
  - OMEGA validation set
  - Astex diverse set



Ebejer J-P, Morris GM, Deane CM, J. Chem. Inf. Model., 52, 1146-1158, 2013

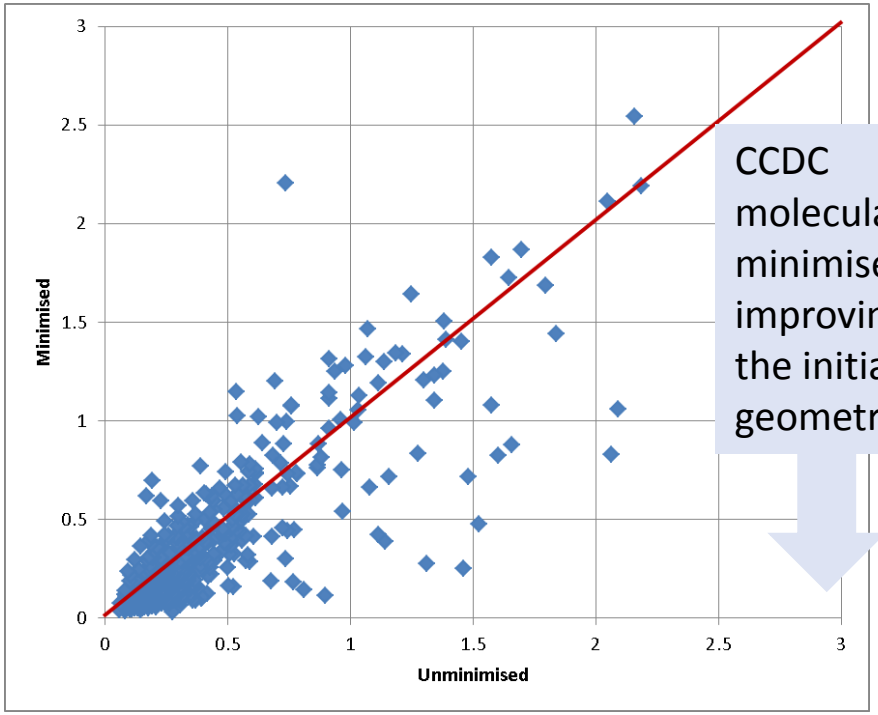
Effect of CCDC molecular minimiser on InhibOx test set

SMILES

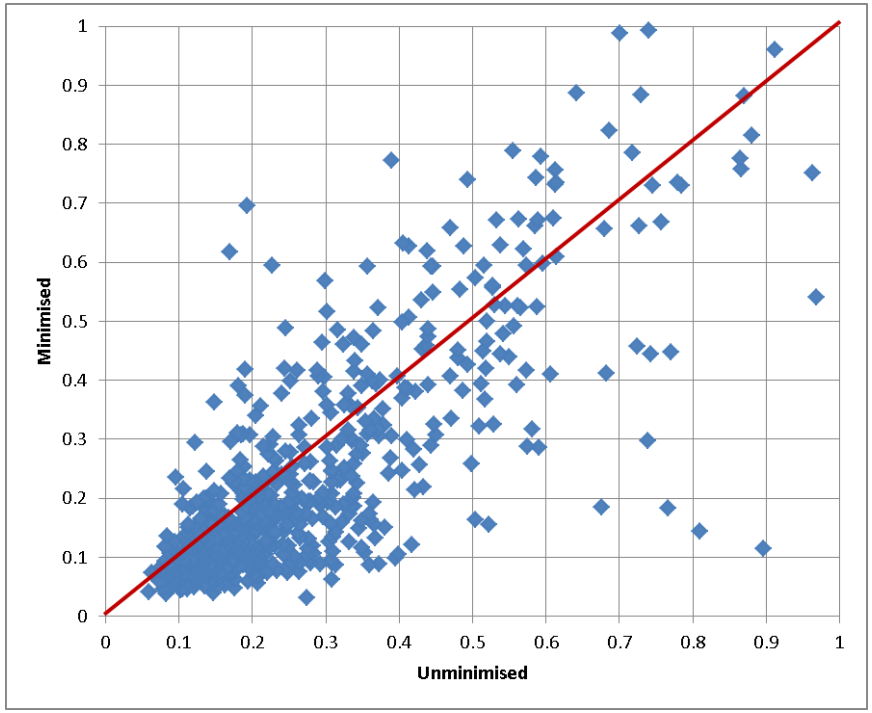
RDKit 3D from UFF

Flexible torsion overlay

CCDC molecular minimiser



CCDC molecular minimiser improving the initial geometry



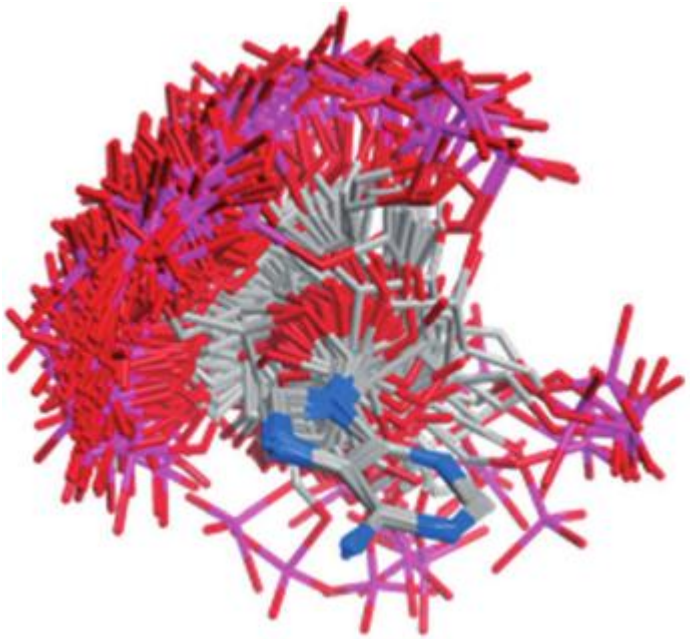
Results using InhibOx test set

	SMILES	SMILES
	RDKit 3D from UFF	RDKit 3D from UFF
		CCDC molecular minimiser
Using up to 200 conformers	CCDC conformer generator	CCDC conformer generator

	No minimization	With minimization
< 0.5 Å (%)	78.0	79.1
< 1.0 Å (%)	95.1	95.4

Validation using ensembles

"Finally, and more fundamentally, the fact that a single bioactive conformation can be reproduced does not prove that the conformation generator creates all relevant low-energy conformations."



343 conformers of AMP extracted from the PDB

Schärfer et al., Chem. Med. Chem., early access, 2013, doi: 10.1002/cmdc.201300242

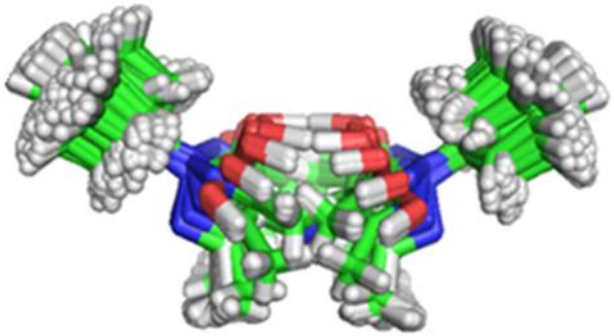
Conformational ensembles from CSD data

CSD

`[SMILES] → [(Refcode, Index)]`

Index map of canonical SMILES

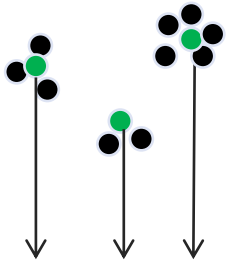
Sets of identical molecules with 3D coordinates



Superimpose

Sets of superimposed molecules

Cluster based on RMSD



Non-redundant set of superimposed molecules

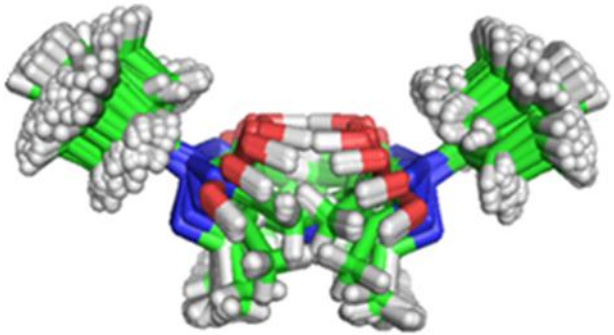
Conformational ensembles from CSD data

CSD

`[SMILES] → [(Refcode, Index)]`

Index map of canonical SMILES

Sets of identical molecules with 3D coordinates

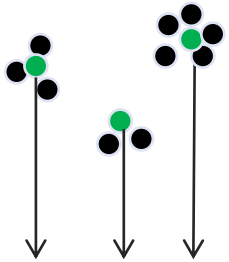


Superimpose

CCDC SMILES do not contain stereo chemistry

Sets of superimposed molecules

Cluster based on RMSD



Non-redundant set of superimposed molecules

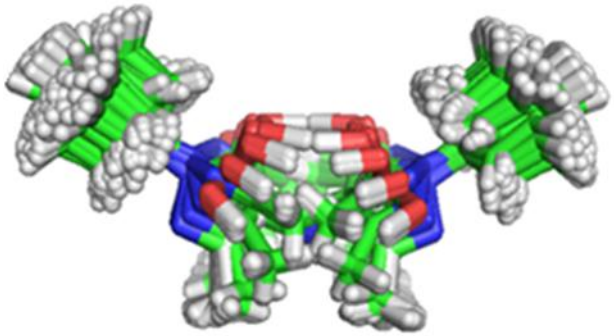
Conformational ensembles from CSD data

CSD

`[SMILES] → [(Refcode, Index)]`

Index map of canonical SMILES

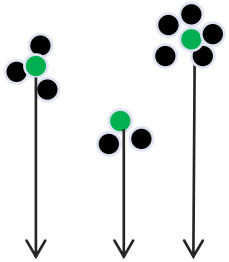
Sets of identical molecules with 3D coordinates



Superimpose

Sets of superimposed molecules

Cluster based on RMSD



Non-redundant set of superimposed molecules

Superimpose with flexible torsion overlay

Cluster based on RMSD

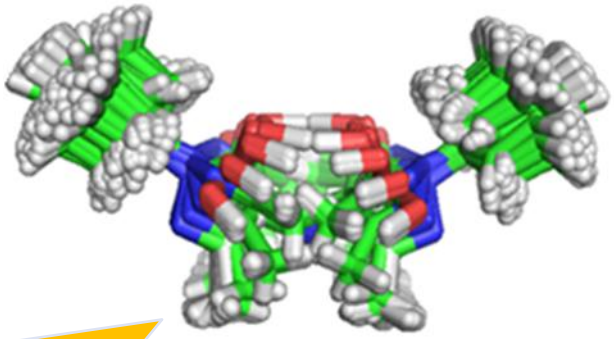
Conformational ensembles from CSD data

CSD

`[SMILES] → [(Refcode, Index)]`

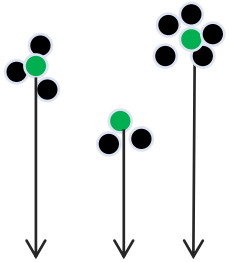
Index map of canonical SMILES

Sets of identical molecules with 3D coordinates



Superimpose with flexible torsion overlay

Conflates issues with conformational variability in rings, valence angles, bond distances



Cluster based on RMSD

Non-redundant set of superimposed molecules

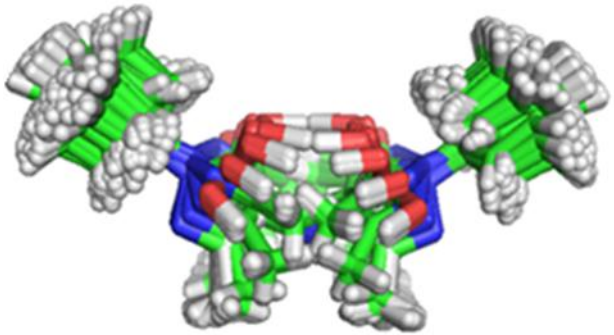
Conformational ensembles from CSD data

CSD

`[SMILES] → [(Refcode, Index)]`

Index map of canonical SMILES

Sets of identical molecules with 3D coordinates

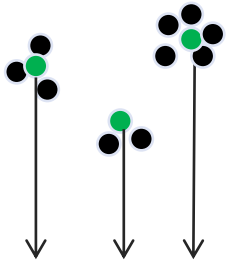


Superimpose

Sets of superimposed molecules

RDKit SMILES do contain stereo chemistry!

Cluster based on RMSD



Non-redundant set of superimposed molecules

“Drug” like filters

- Number of conformers before filtering: 22671

- Filters:
  - Reject if molecular weight > 650 or number of atoms > 150
  - Allowed elements: C, H, D, Cl, F, Br, I, N, O, S, P, B, Si
  - Reject if number of hydrogen bond acceptors > 15
  - Reject if number of hydrogen bond donors > 7
  - Reject if number of rotatable bonds > 20

- Number of conformers after filtering: 19878
- Number of conformers after RMSD clustering: **8581**

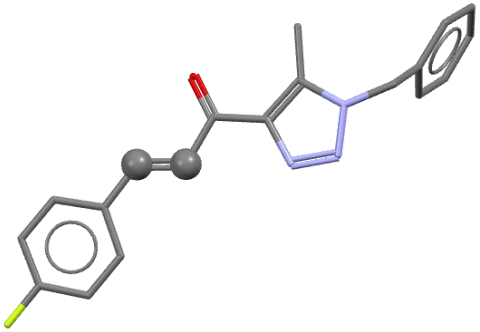
Getting chirality from structure in RDKit

conformer_test_set_utilities.py (Z:\scripts\ccdc_toolkit) - GVIM

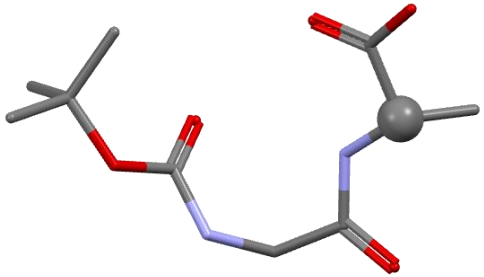
File Edit Tools Syntax Buffers Window Help

```python
import rdkit.Chem
import rdkit.Chem.AllChem
from rdkit.Chem import rdmolops

def get_rdkit_mol(mol):
    "Return RDKit molecule from a CCDC molecule."
    mol_block = mol.to_string('sdf')  # CCDC Python API
    rdkit_mol = rdkit.Chem.MolFromMolBlock(mol_block)
    if rdkit_mol is None:
        raise RuntimeError("Can't create RDKit molecule from %s" % mol_block)
    rdmolops.AssignAtomChiralTagsFromStructure(rdkit_mol)
    return rdkit_mol
```

4,0-1 5%



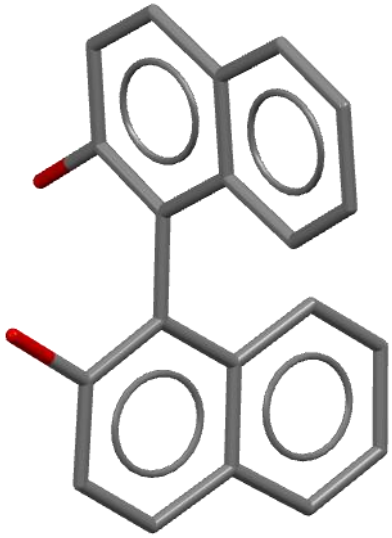
Refcode: ABABUC



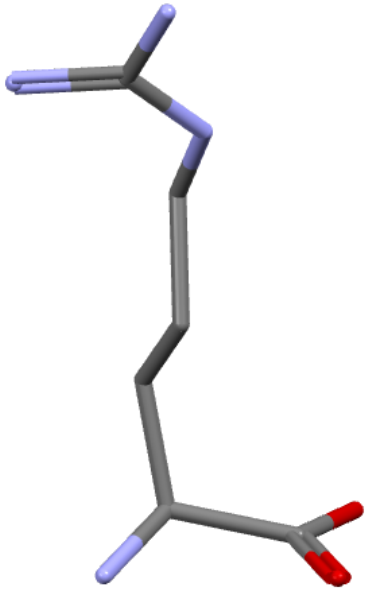
Refcode: BXGLAL

ABABUC: Cc1c(C(=O)/C=C/c2ccc(F)cc2)nnn1Cc1ccccc1
BXGLAL: C[C@H](NC(=O)CNC(=O)OC(C)(C)C)C(=O)O

Conformational ensemble examples



1,1'-Bi-2-Naphthol
Number of instances: 166
Number of conformations: 2
Refcode: ABOXAS



L-Arginine
Number of instances: 54
Number of conformations: 27
Refcode: ADAVAC

Generation of initial 3D conformations using RDKit

CSD 3D molecule

RDKit isomeric SMILES

RDKit 3D from UFF

CCDC molecular minimiser

CCDC conformer generator

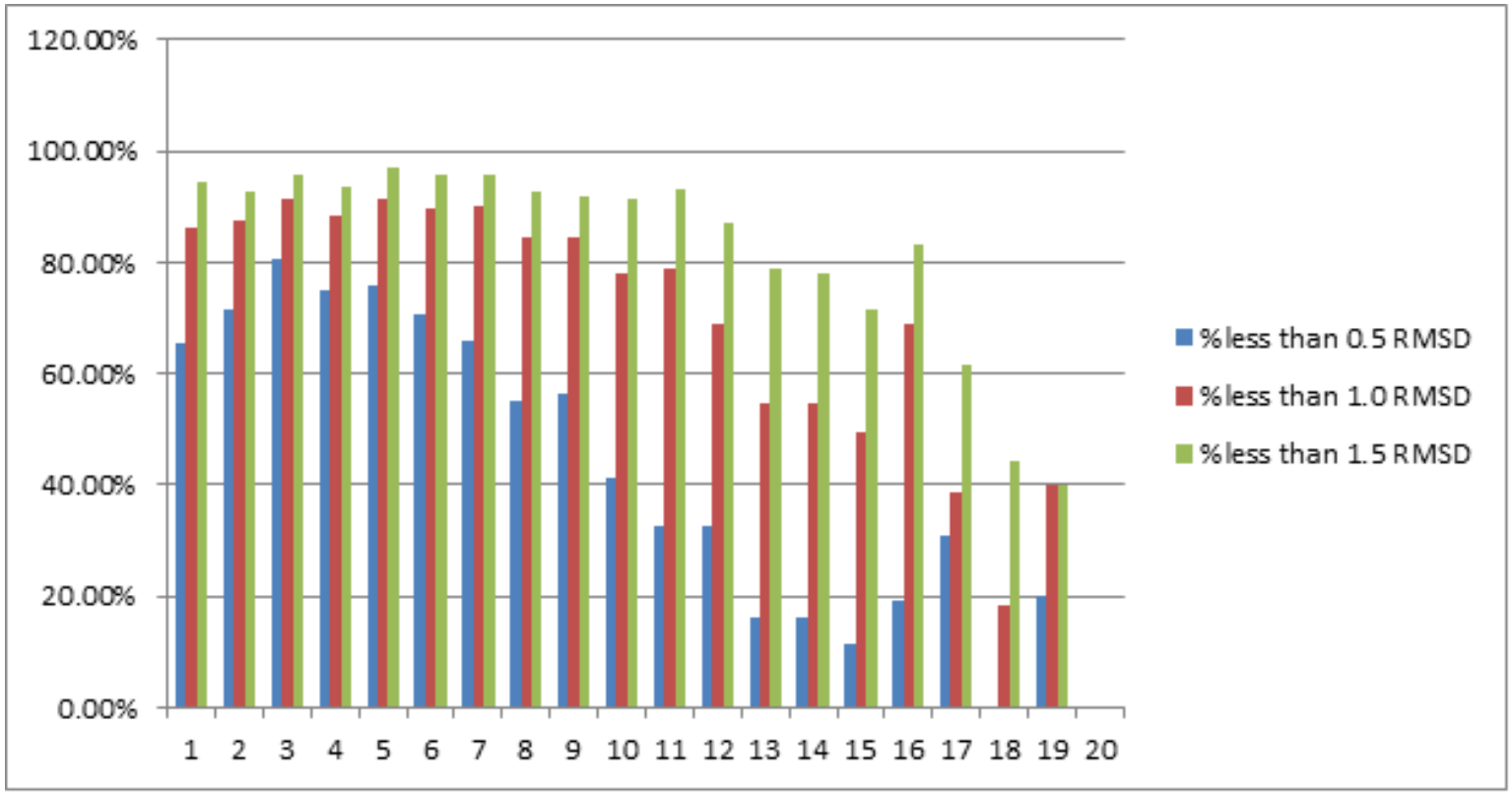
rdkit_mol2_to_smiles.py = (\\unix\co...Python_API\frequent_molecules) - GVIM1

File Edit Tools Syntax Buffers Window Help

```
def get_rdkit_3D_mol_from_smiles(isomeric_smiles, name):
    "Return RDKit molecule with hydrogen atoms and 3D coordinates."
    rdkit_mol = rdkit.Chem.MolFromSmiles(isomeric_smiles)
    rdkit_mol = rdmolops.AddHs(rdkit_mol)
    rdkit_mol.SetProp("_Name", name)
    if not rdkit.Chem.AllChem.EmbedMolecule(rdkit_mol) == 0:
        raise RuntimeError('rdkit.Chem.AllChem.EmbedMolecule failed for %s' % name)
    if not rdkit.Chem.AllChem.UFFOptimizeMolecule(rdkit_mol, maxIters=2000) == 0:
        raise RuntimeError('rdkit.Chem.AllChem.UFFOptimizeMolecule failed for %s' % name)
    return rdkit_mol
                                                                    36,0-1        53%
```

Validation!

Results using new test set



Number of rotatable torsions

That's a bit rubbish! What is going on…?

Issue 1: Where has my E-Z stereo-chemistry gone?

- Lot of the input 3D conformations had the wrong E-Z stereo-chemistry
- When processing structures in batch we found that we had more success parsing mol2 files than sdf files into RDKit. On first 1000 structures in CSD:
  - Mol2 better: 229
  - SDF better: 64
  - Both fail: 228
- So, switched to converting molecules using mol2 file format…
- However
  - From sdf:
    `Cc1c(C(=O)/C=C/c2ccc(F)cc2)nnn1Cc1ccccc1`
  - From mol2:
    `Cc1c(C(=O)C=Cc2ccc(F)cc2)nnn1Cc1ccccc1`

Hindsight: sdf vs mol2

	First 1000 entries	First 1000 metal-organics	First 1000 organics
Mol2 better	229	506	11
SDF better	64	31	141
Both fail	228	326	43

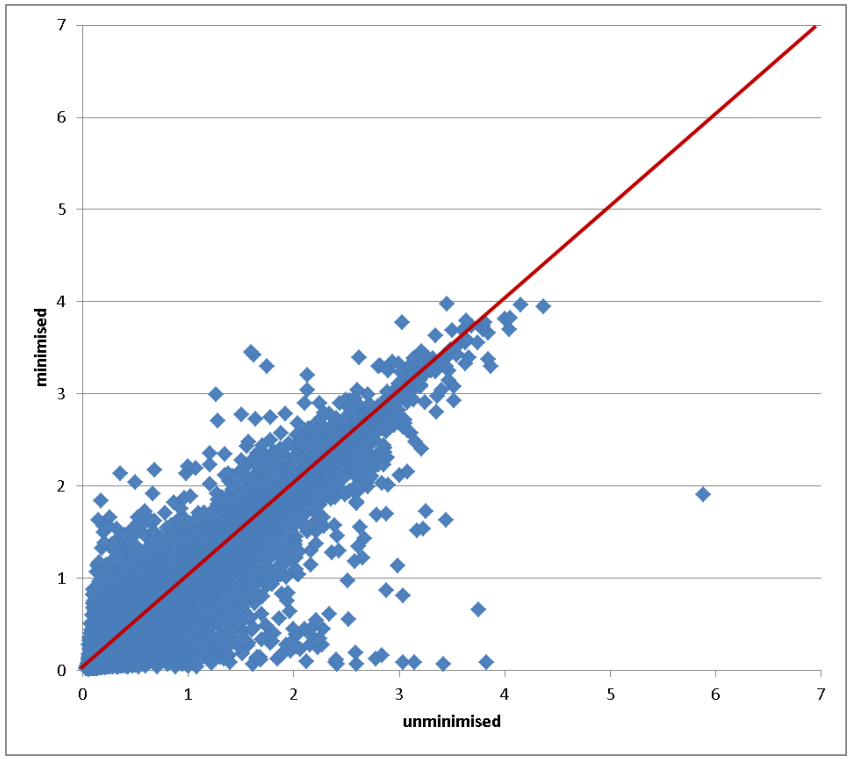
Issue 2: Some molecules “too complicated” for UFF…

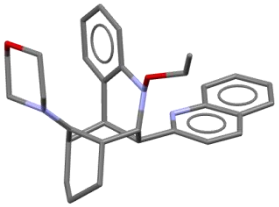
SMILES

RDKit 3D from UFF

Flexible torsion overlay

CCDC molecular minimiser





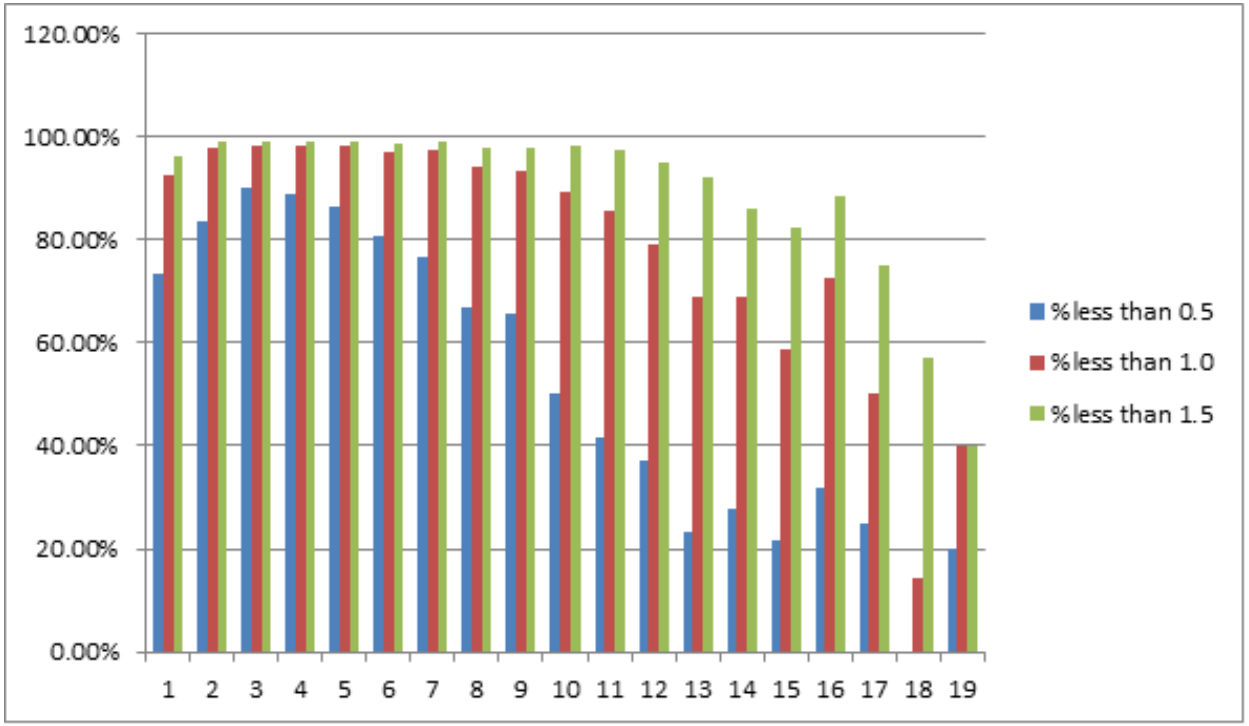
Refcode: EQZBMI

“Too complicated” = Not possible to reproduce experimental structure to < 1 Å RMSD using a flexible torsion overlay

… an alternative view is that our conformer minimiser/generator is not good enough (i.e. it should be able to fix more issues with macro-cycles, bridged-ring systems, etc…

Results using new test set

- Fixed SMILES to include E-Z stereo-chemistry
- Filtered out any conformations where the initial UFF conformer could not reproduce the experimental structure to < 1.0 Å using a flexible torsion overlay



Number of rotatable torsions

Summary

- Work in progress…
  - Creating test sets is always a painful business
  - Always discover new nuances
  - **Are there better ways of converting molecules between toolkits?**
- Use of RDKit at CCDC
  - 2D diagram generation for internal structure reports
  - Calculating stereo chemistry from 3D coordinates
  - Generating initial 3D conformations from SMILES
- CSD is a great test set for validating chemistry toolkits
  - **Chemistry in CSD can be very challenging**
  - Will provide feedback to RDKit once we understand why and where the issues are arising

Acknowledgments

- Jason Cole
- Oliver Korb
- Patrick McCabe
- Robin Taylor
- Richard Sykes

Thank you for your attention

Are there any questions?