

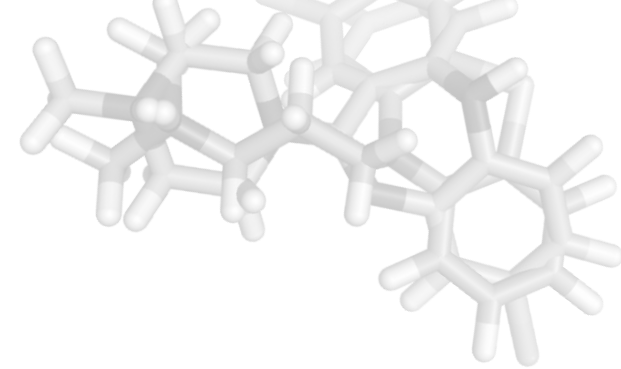
Open-Source Cheminformatics  
and Machine Learning

6th RDKit UGM 2017

# Conformation generation with RDKit: Comparative study and application reproducing bioactive overlays with PharmScreen

*Enric Herrero, PhD CTO*

[info@pharmacelera.com](mailto:info@pharmacelera.com)



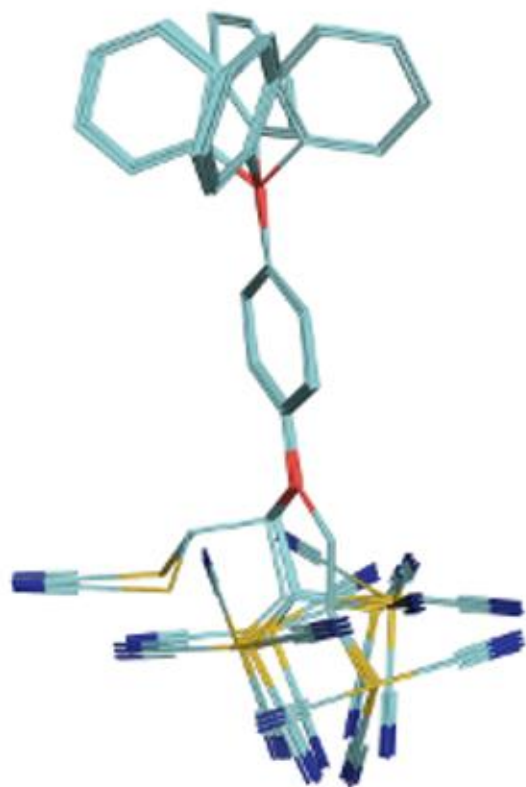
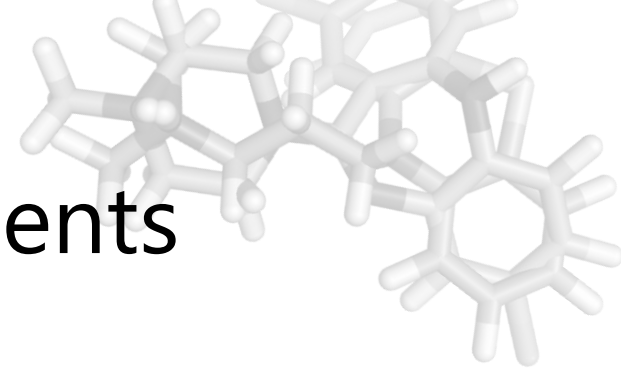
# Brief introduction

CADD Software company based in Barcelona

- PharmScreen: high precision ligand-based **virtual screening** software
- PharmQSAR: Quantitative Structure-Activity Relationship (**QSAR**) software package



# Conformation generation requirements

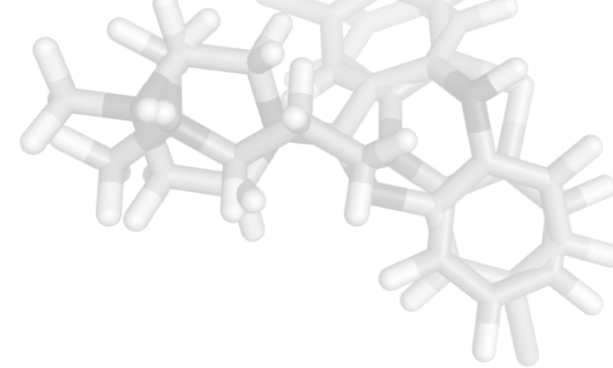


What are we looking in a conformer generator tool?

- **Accuracy** on reproducing the experimentally determined structures
- Low **number of conformers**
- Reasonable computation **time**

The balance of these three properties leads to an optimal conformer generator tool.

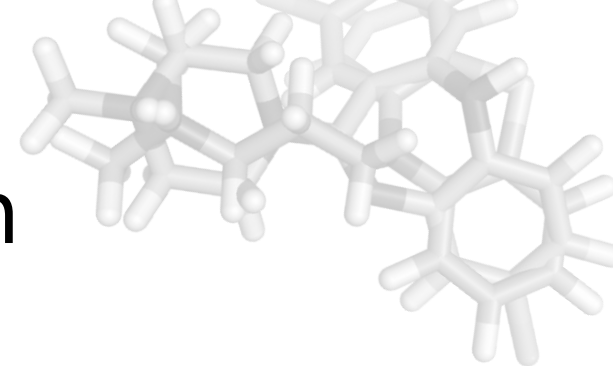
# Outline



1. RDKit Conformation generation
2. Optimization of the conformer generation strategy with RDKit
3. Flexible superposition study using PharmScreen



# Conformation generation tool comparison



JOURNAL OF  
**CHEMICAL INFORMATION  
AND MODELING**

Article


pubs.acs.org/jcim

## Freely Available Conformer Generation Methods: How Good Are They?

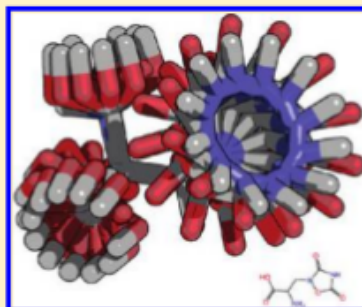
Jean-Paul Ebejer,<sup>†,‡</sup> Garrett M. Morris,<sup>‡</sup> and Charlotte M. Deane<sup>\*,†</sup>

<sup>†</sup>Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K., and

<sup>‡</sup>InhibiOx Limited, Oxford Centre for Innovation, New Road, Oxford, OX1 1BY, U.K.

 Supporting Information

**ABSTRACT:** Conformer generation has important implications in cheminformatics, particularly in computational drug discovery where the quality of conformer generation software may affect the outcome of a virtual screening exercise. We examine the performance of four freely available small molecule conformer generation tools (BALLOON, CONFAB, FROG2, and RDKit) alongside a commercial tool (MOE). The aim of this study is 3-fold: (i) to identify which tools most accurately reproduce experimentally determined structures; (ii) to examine the diversity of the generated conformational set; and (iii) to benchmark the computational time expended. These aspects were tested using a set of 708 drug-like molecules assembled from the OMEGA validation set and the Astex Diverse Set. These molecules have varying physicochemical properties and at least one known X-ray crystal structure. We found that RDKit and CONFAB are statistically better than other methods at generating low rmsd conformers to the known structure. RDKit is particularly suited for less flexible molecules while CONFAB, with its systematic approach, is able to generate conformers which are geometrically closer to the experimentally determined structure for molecules with a large number of rotatable bonds ( $\geq 10$ ). In our tests RDKit also resulted as the second fastest method after FROG2. In order to enhance the performance of RDKit, we developed a postprocessing algorithm to build a diverse and representative set of conformers which also contains a close conformer to the known structure. Our analysis indicates that, with postprocessing, RDKit is a valid free alternative to commercial, proprietary software.



Proposed RDKit configuration:

$$n = \begin{cases} 50 & \text{if } n_{rot} \leq 7 \\ 200 & \text{if } n_{rot} \geq 8 \text{ and } n_{rot} \leq 12 \\ 300 & \text{otherwise} \end{cases}$$

RMSD threshold 0.5 and 0.35 Å

2012

ation

### Used options:

- numConfs: Number of conformers to be generated
- pruneRmsThresh: RMS cutoff to keep the conformers
- useExpTorsionAnglePrefs: use experimental information about torsion angles
- useBasicKnowledge: use basic knowledge such as flat aromatic rings and linear triple bonds.

Article  
pubs.acs.org/cim

Sereina Riniker<sup>\*,†</sup> and Gregory A. Landrum<sup>‡</sup>

<sup>†</sup>Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland

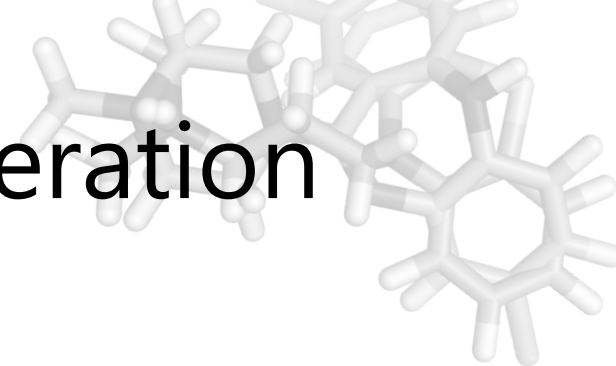
<sup>†</sup>Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland

**S** Supporting Information

Distance Geometry + Experimental Torsional-Angle Preferences = ETKDG

2015

## 2. Optimization of the conformer generation strategy with RDKit

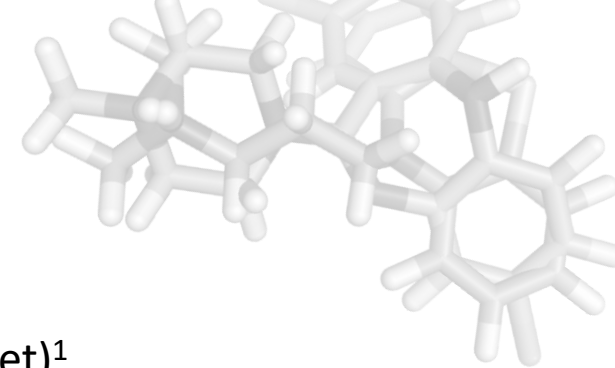


Dataset: AstraZeneca Overlays Validation Test Set (1456 molecules divided in 121 set)<sup>1</sup>  
Available in the Cambridge Crystallographic Data Centre



1. Optimize conformer generation strategy
  1. Generate different **number of conformers** for each molecule and find a relation with the rotatable bond number.
  2. Remove conformers based on their **energy**.
  3. Prune the generated conformers based on the **RMSD**.

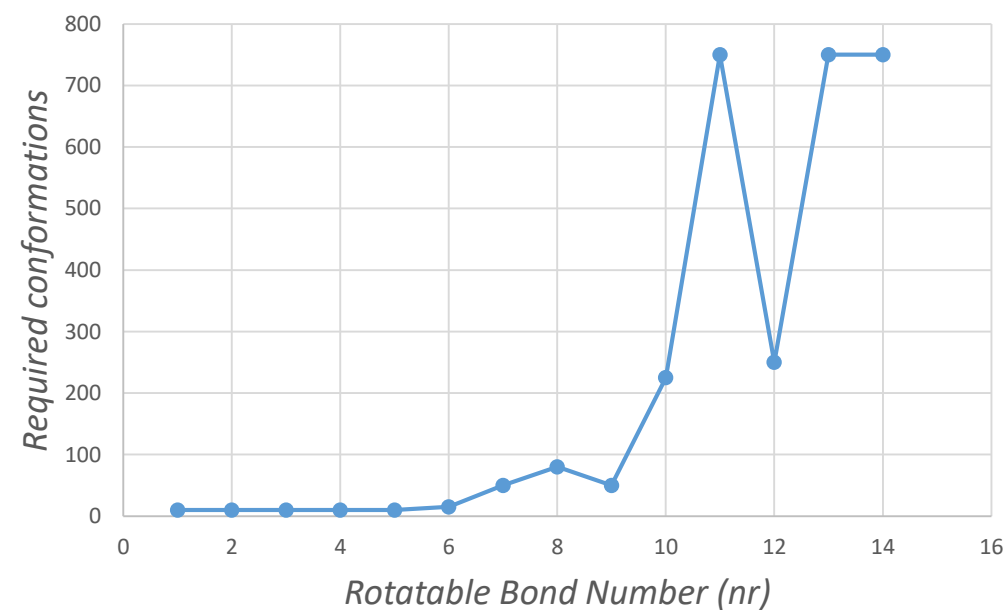
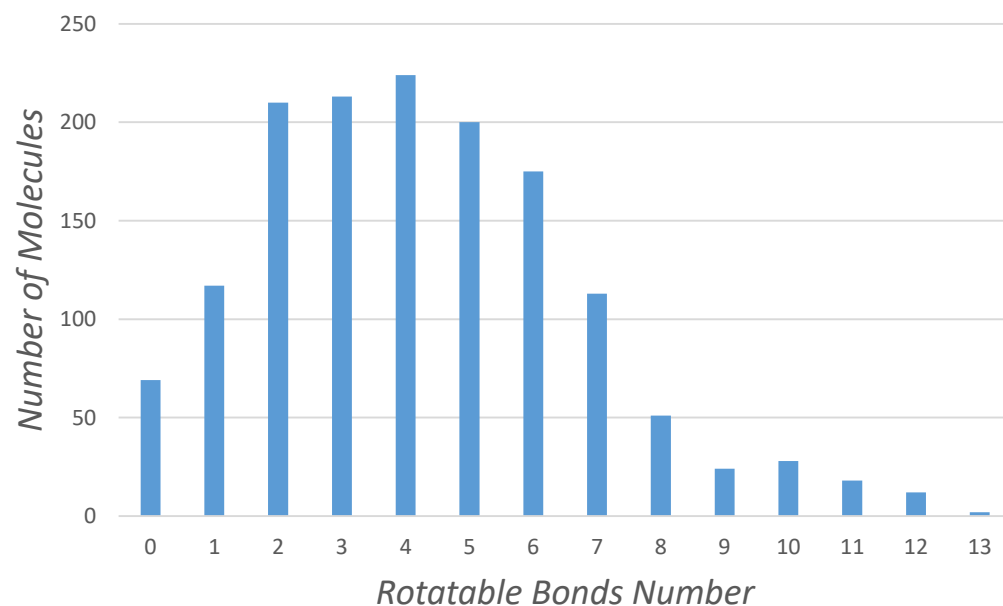
1. Giangreco I, Cosgrove DA, Packer MJ (2013) An extensive and diverse set of molecular overlays for the validation of pharmacophore programs. J Chem Inf Model 53:852–866.



# Dataset analysis

Dataset: AstraZeneca Overlays Validation Test Set (1456 molecules divided in 121 set)<sup>1</sup>

## Rotatable Bond Number Distribution



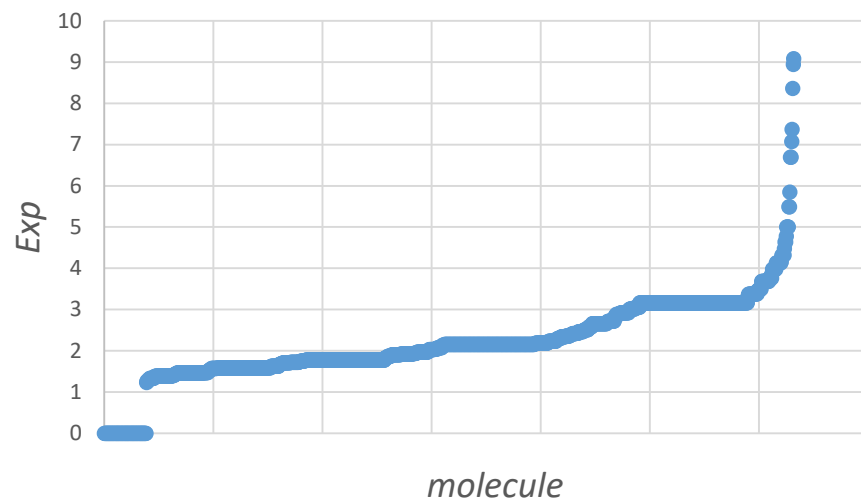
Would it have sense to use an exponential function to select the required number of conformations?

1. Giangreco I, Cosgrove DA, Packer MJ (2013) An extensive and diverse set of molecular overlays for the validation of pharmacophore programs. J Chem Inf Model 53:852–866.





# Conformers-Rotatable Bond Number Relationship



## Thresholds

50 for  $nr < 3$

300 for  $nr > 6$

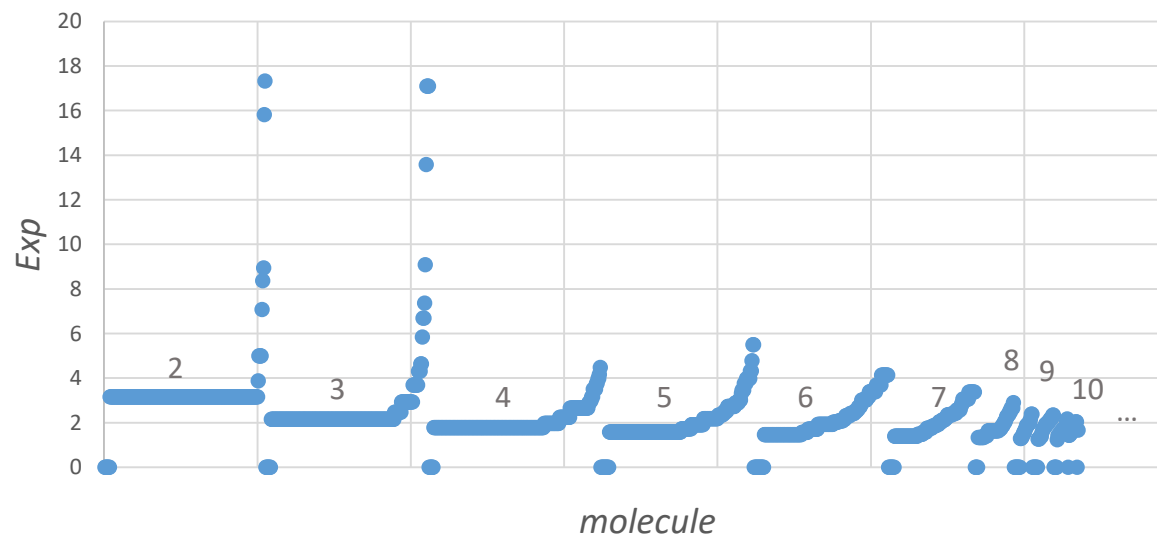
$$Exp = \sqrt[nr]{m}$$

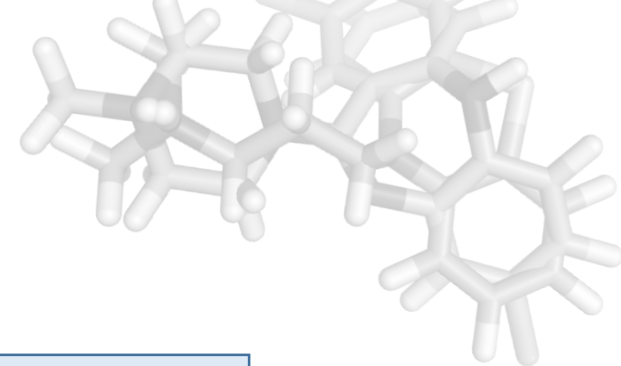
$nr$  = rotatable bonds number

$m$  = minimum conformer number needed

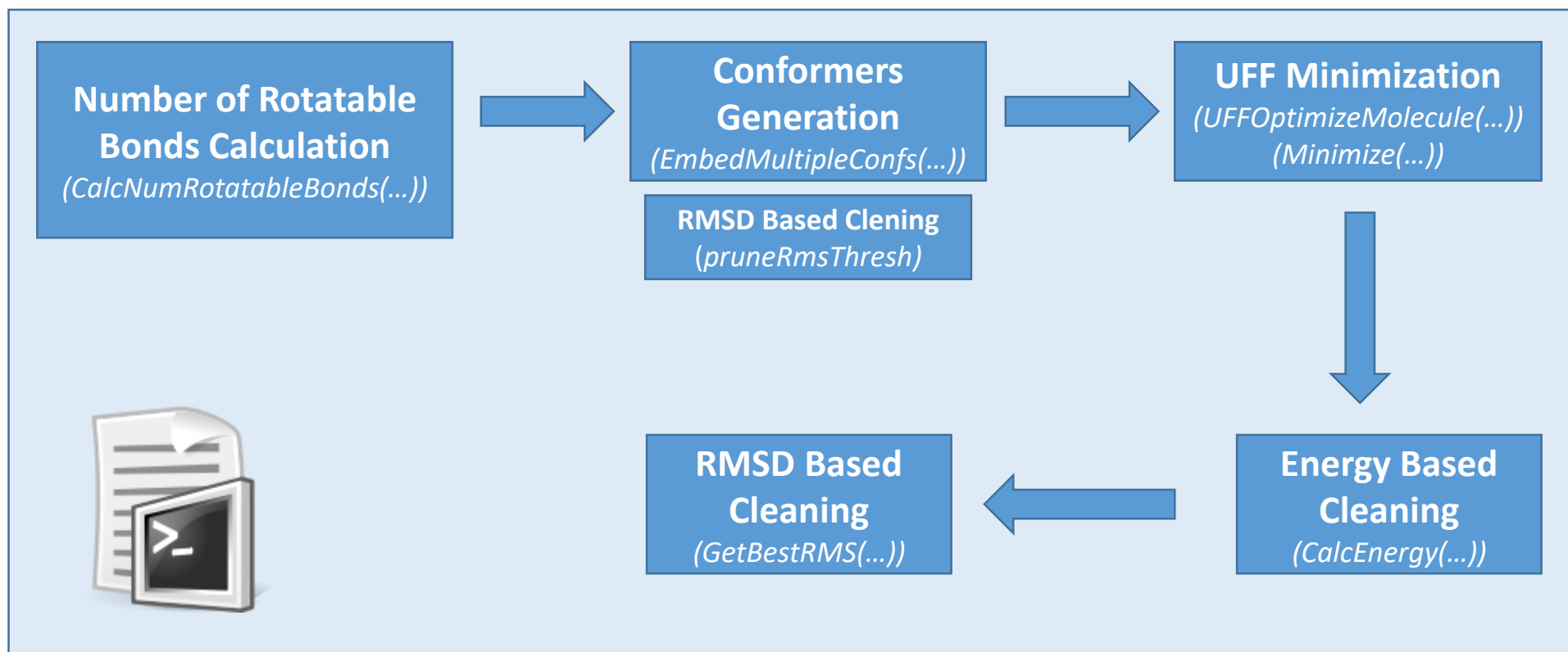
Average value without outliers is 2

3 covers most of the molecules



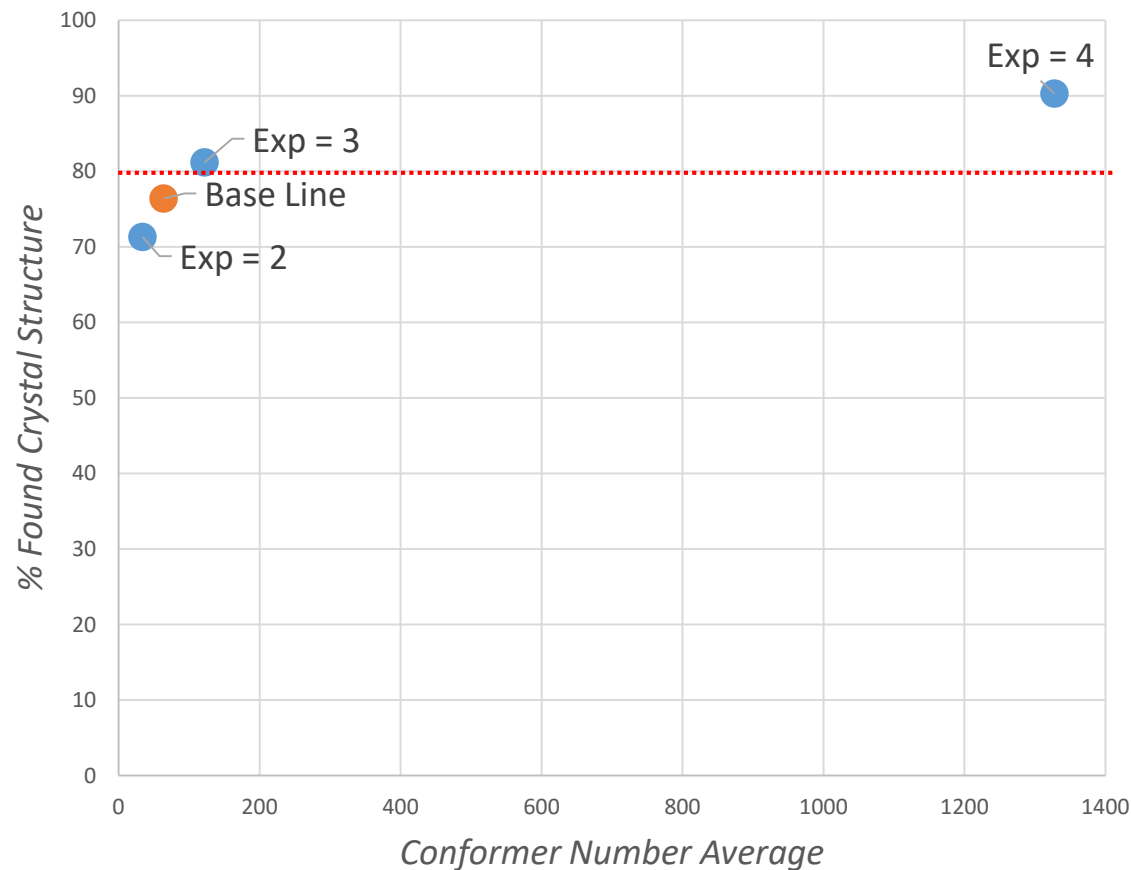


# Conformers Generation Script

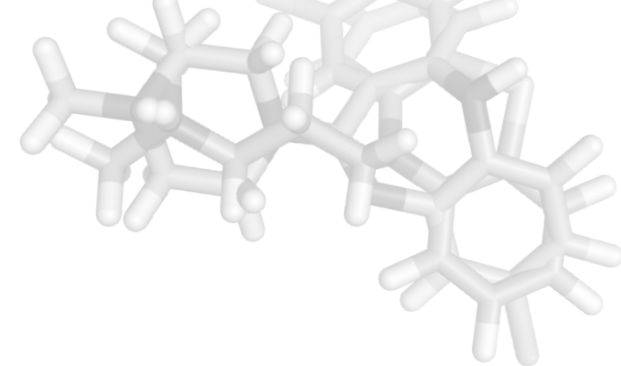


Script available at [www.pharmacelera.com/scripts/rdkit-conformation-generation-script](http://www.pharmacelera.com/scripts/rdkit-conformation-generation-script)

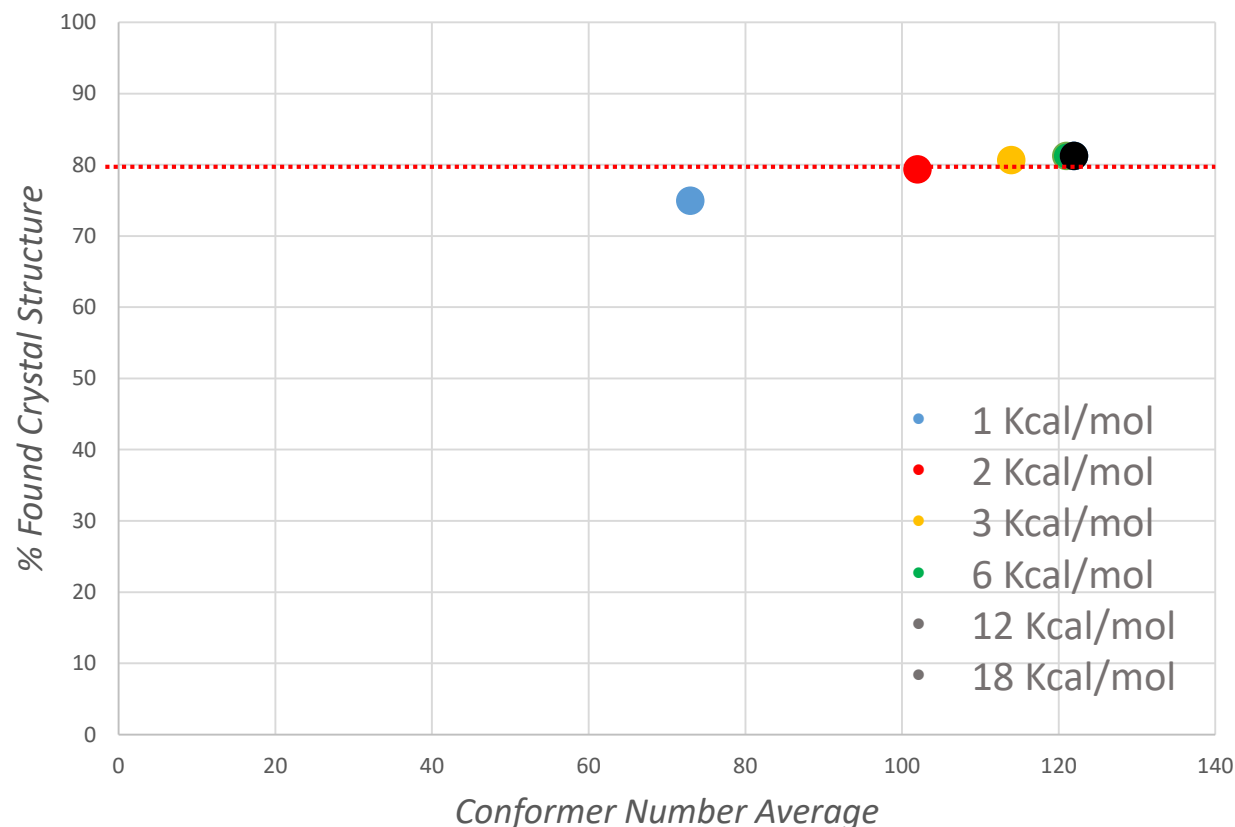
# Conformers-Rotatable Bond Number Relationship



Using exp= 3 gives a good compromise between accuracy and number of conformers.



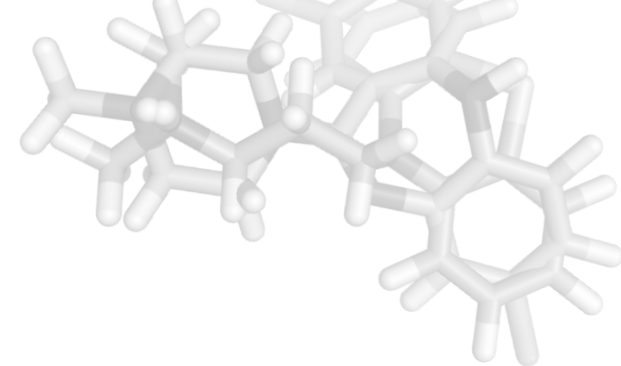
# Energy Cleaning Results



```
ff =  
AllChem.UFFGetMoleculeForceField()  
ff.Minimize()  
ff.CalcEnergy()
```

Six different energy windows were tested (1, 2, 3, 6, 12, 18 Kcal/mol).

The energy cleaning did not show important effects neither on reducing the number of conformers neither on the percentage of correct conformers.

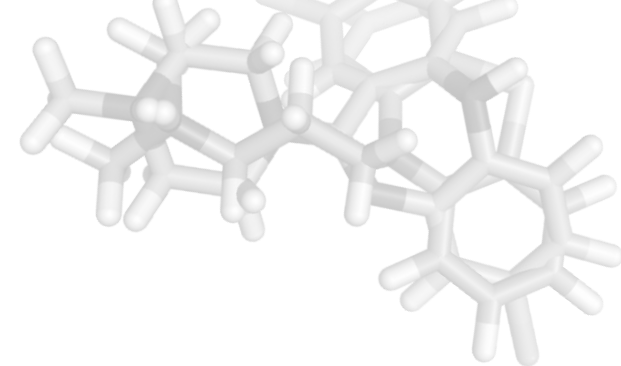


# RMSD Cleaning Strategy

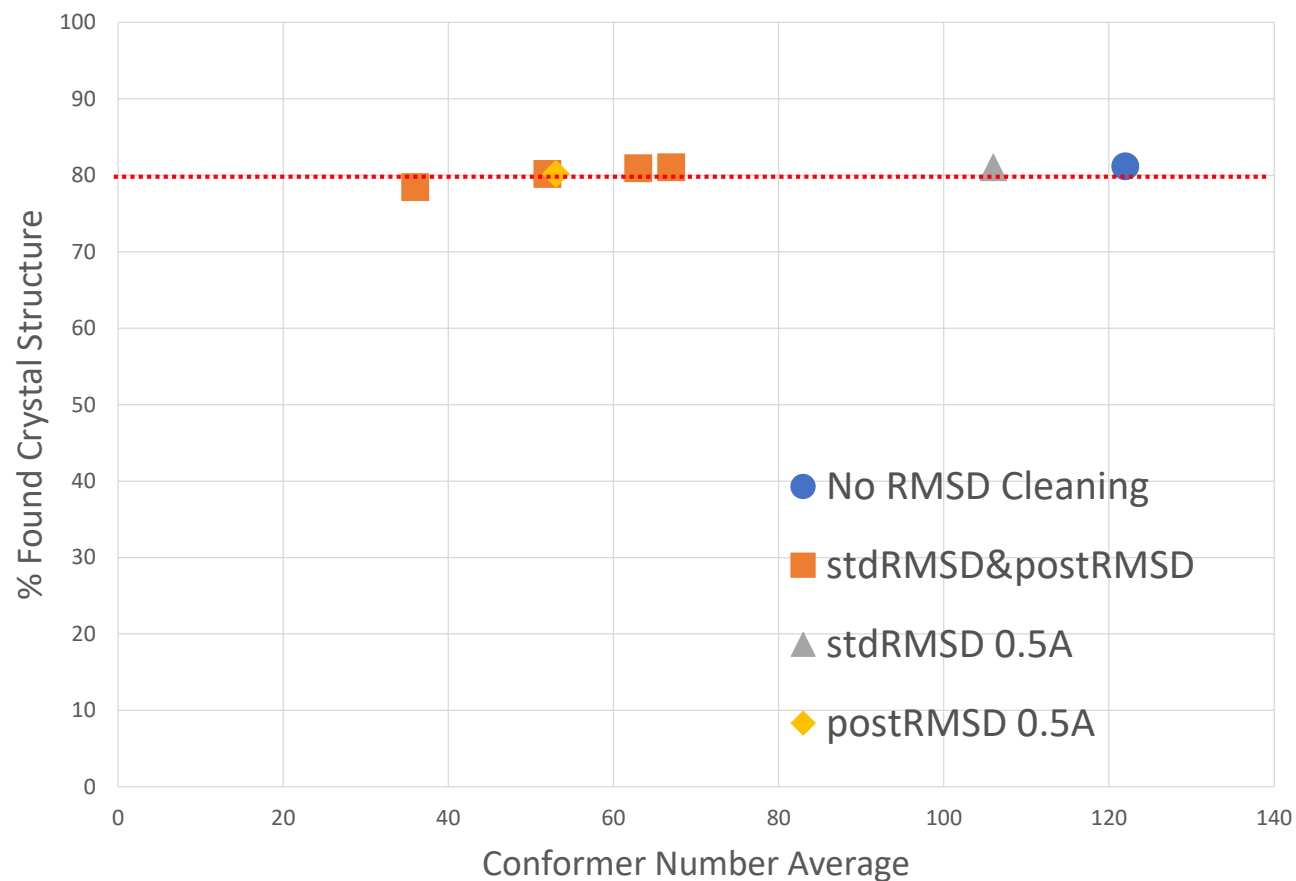
Two possible RMSD cleaning:

1. During Conformers Generation (stdRMSD): with *pruneRmsThresh* an option included on *EmbedMultipleConfs()* function to keep only conformations that are at least a particular RMSD threshold apart one another.
2. Post Minimization (postRMSD): the conformers were sorted by increasing energy value after minimization and kept only those with RMSD value higher than a choose.  
*sort()*  
*GetBestRMS()*





# RMSD Cleaning Results

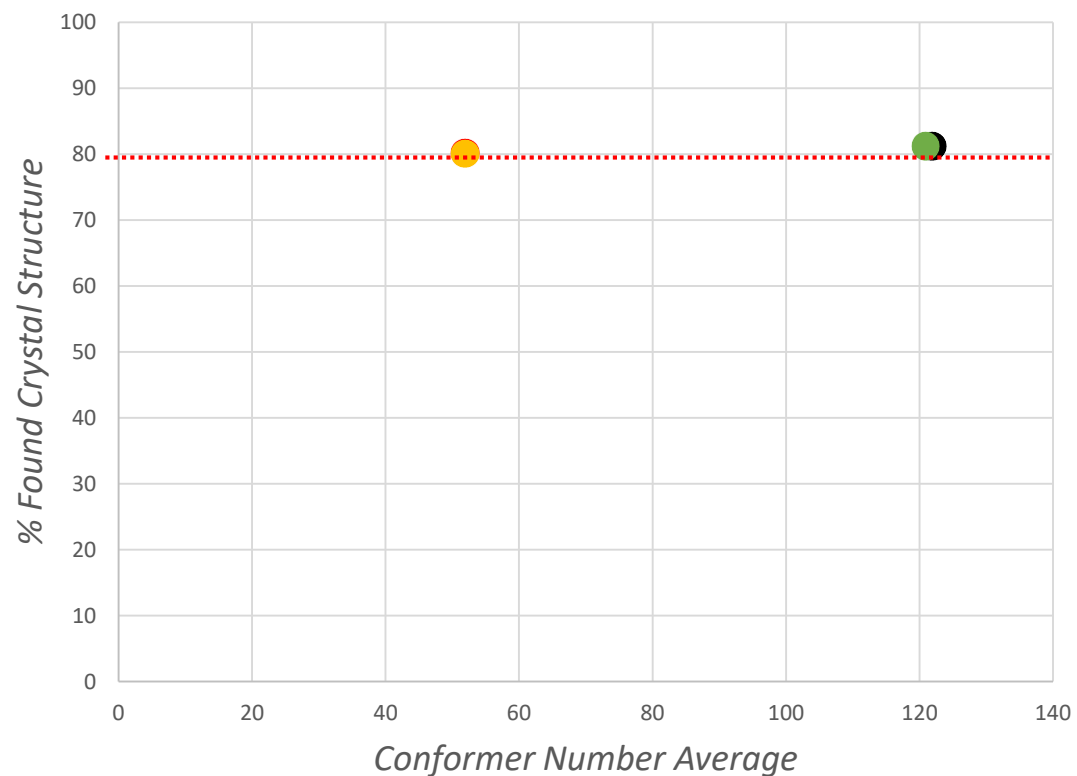
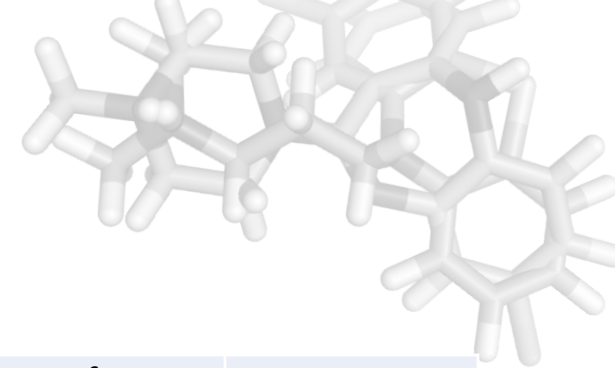


The two RMSD cleaning steps were performed with the same filtering value, 0.25, 0.35, 0.50 and 0.75 Å.

stdRMSD does not reduce conformers but reduces simulation time speeding up minimization and postRMSD

postRMSD reduces significantly the number of conformers without degrading performance

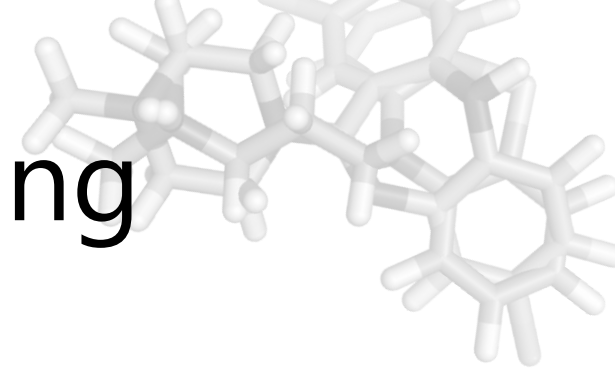
# Final Configuration



Strategy	Conformer Number Average	Found Crystal Structure
No Cleaning	122	81.18 %
Energy Cleaning 6.0 Kcal/mol	121	81.18 %
RMSD Cleaning 0.50 Å	52	80.15 %
Final Configuration: Energy Cleaning 6.0 Kcal/mol & RMSD Cleaning 0.50 Å	52	80.01 %

- No Cleaning
- Energy Cleaning 6.0 Kcal/mol
- RMSD Cleaning 0.50 Å
- Final Configuration: Energy Cleaning 6.0 Kcal/mol & RMSD Cleaning 0.50 Å

# 3. Flexible superposition study using PharmScreen



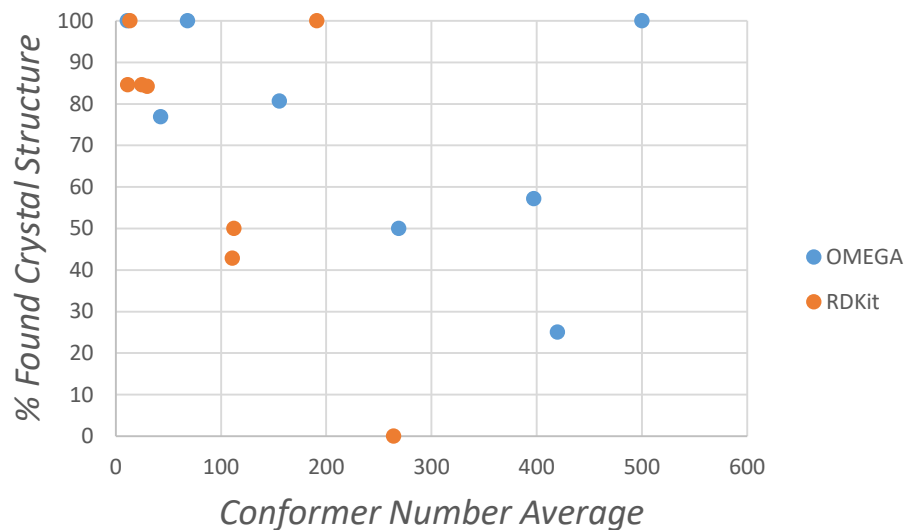
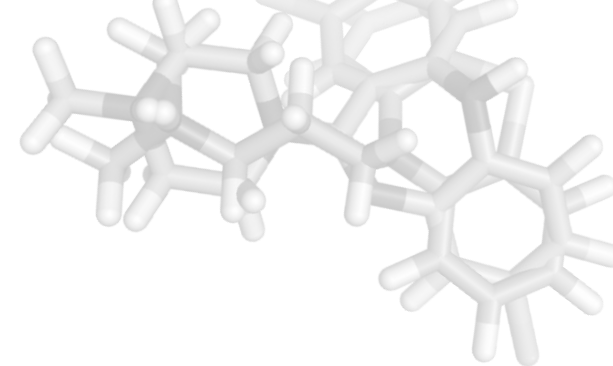
**Goal:** Reproduce bioactive overlays of flexible compounds

**Sets:** From Eli Lilly study<sup>1</sup>

<u>Name</u>	<u>Number of Molecules</u>	<u>Rotatable Bond Number Average</u>
• CDK2	57	4
• Elastase	7	8
• ESR1	13	4
• HIV	28	15
• p38	13	3
• Rhinovirus	8	9
• Thermolysin	12	8
• Trypsin	7	3

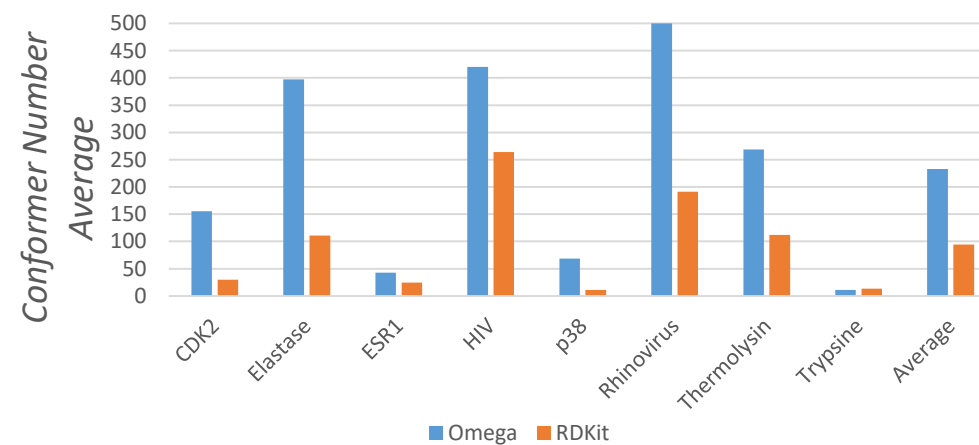
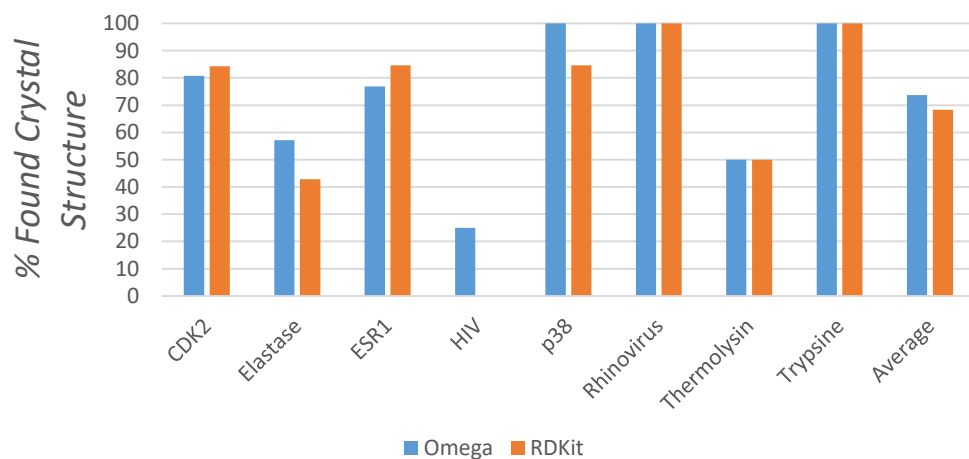
1. Chen Q, Higgs R.E, Vieth M (2006) Geometric Accuracy of Three-Dimensional Molecular Overlays. J Chem Inf Model 46(5): 1996-2002.

# RDKit vs OMEGA



Both tools provide a similar performance finding crystals but **RDKit requires significantly less conformers**

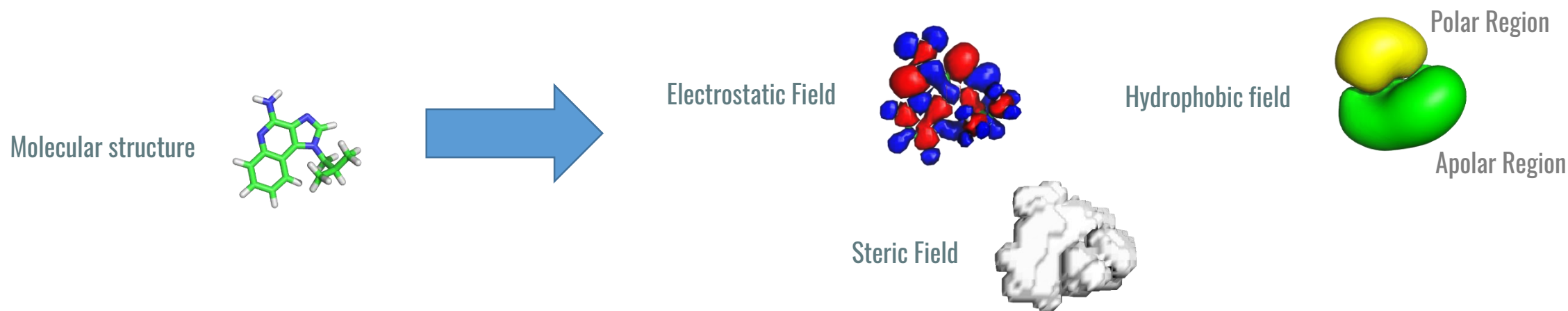
Omega is better in HIV, the most complex structure



# PharmScreen field-based alignment

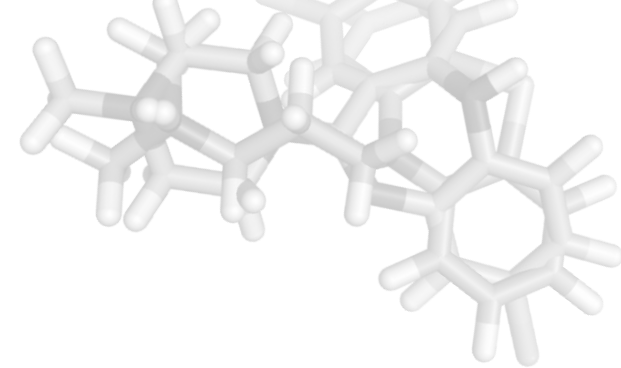
PharmScreen, a Virtual Screening tool which allows choosing different setup options:

- ✓ Determine atomic charges with various methods: Gasteiger; AM1-BCC; RM1-ESP
- ✓ Choose the best descriptor fields among Shape, Electrostatic and/or Hydrophobic to perform the superpositions
- ✓ Perform single or multi fields experiment
- ✓ Assign the best combination weight to the chosen fields





# Alignment options



PharmScreen New Experiments View Files Settings Logout

New Experiment [User Manual](#)

Experiment Name ^

Name  Destination folder: **CDK2**

Molecule Library ^

[Select File](#) [Upload File](#) Selected molecule library: **CDK2\_alessandro/CDK2\_conf\_par.mol2.zip**

Reference Molecules ^

[Select File](#) [Upload File](#) Selected reference molecules: **CDK2\_alessandro/CDK2\_set.mol2.zip**

Advanced Options (Optional) ^

Molecular charges ^

LogP ^

Similarity Component Weight ^

Setup Components & Weights ☒

Field 1:  Weight (%):

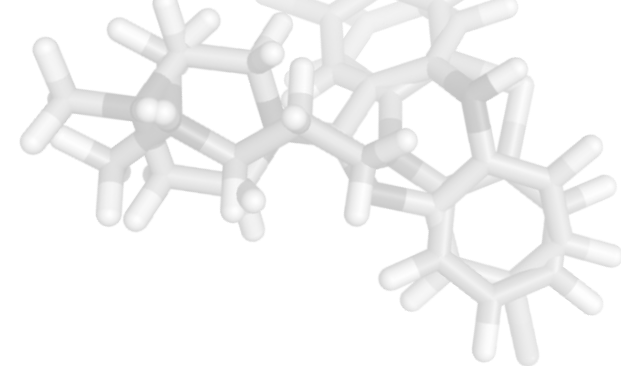
Field 2:  Weight (%):

Field 3:  Weight (%):

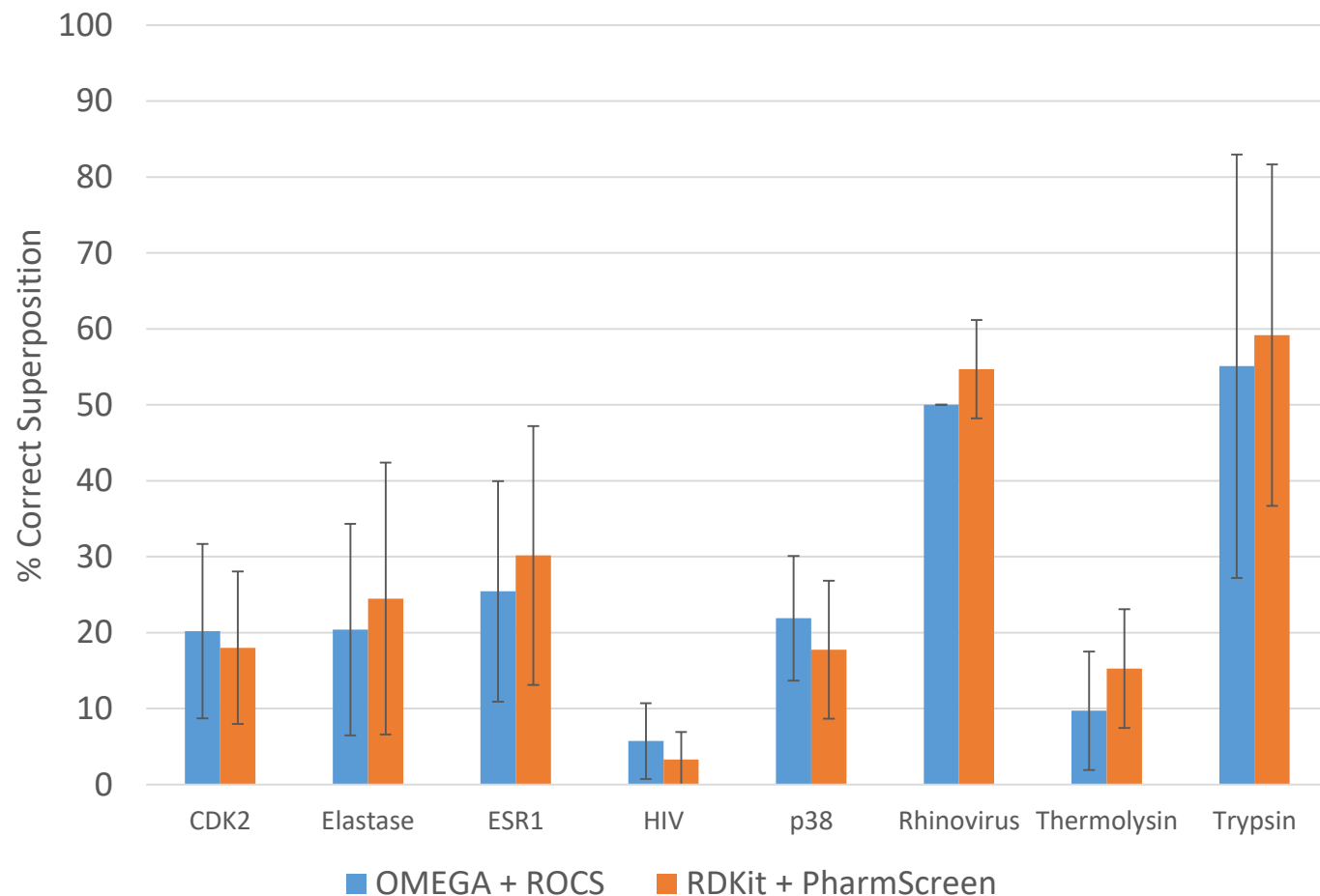
All Crystal structures are tested as the reference

Fields and scoring function:

1.  $\text{LogP}_{\text{ele}}$  15%
2.  $\text{LogP}_{\text{cav}}$  55%
3. H-Bond 30%



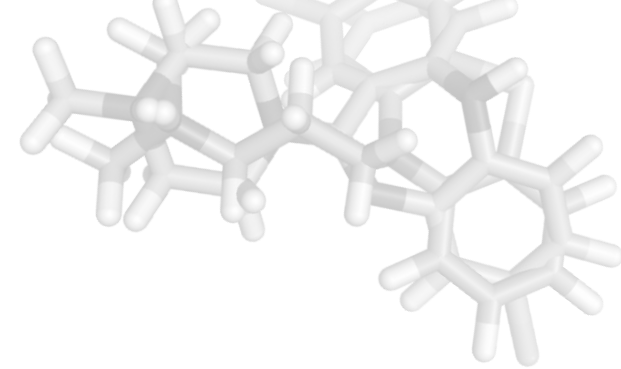
# Flexible alignment results



Alignments are considered correct if RMSD vs the Crystal structure is less than 2Å

In 5 out of 8 sets RDKit + PharmScreen outperforms OMEGA + ROCS reproducing bioactive overlays.<sup>1</sup>

1. Chen Q, Higgs R.E, Vieth M (2006) Geometric Accuracy of Three-Dimensional Molecular Overlays. J Chem Inf Model 46(5): 1996-2002.



# Summary

- RDKit provides high quality conformers
- Further cleaning reduces the number of conformers without affecting quality
  - Available script at [www.pharmacelera.com/scripts/rdkit-conformation-generation-script](http://www.pharmacelera.com/scripts/rdkit-conformation-generation-script)
- RDKit and PharmScreen can reproduce bioactive overlays accurately which can be used for:
  - Virtual screening
  - QSAR analysis

# Thank you!



## Pharmacelera<sup>TM</sup>

NEXT GENERATION DRUG DISCOVERY SOLUTIONS

- Contact information:
- [info@pharmacelera.com](mailto:info@pharmacelera.com)