JOHN MAYFIELD

EGON WILLIGHAGEN

CHEMISTRY DEVELOPMENT KIT V2.0

CINF 64: ACS Fall 2017, Washington, D.C.

Willighagen *et al. J Cheminform (2017) 9:33*
DOI 10.1186/s13321-017-0220-4

Journal of Cheminformatics

SOFTWARE Open Access

The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching

Egon L. Willighagen[1*], John W. Mayfield[2], Jonathan Alvarsson[3], Arvid Berg[3], Lars Carlsson[4], Nina Jeliazkova[5], Stefan Kuhn[6], Tomáš Pluskal[7], Miquel Rojas-Chertó[8], Ola Spjuth[3], Gilleain Torrance[9], Chris T. Evelo[1], Rajarshi Guha[10] and Christoph Steinbeck[11]

http://cdk.github.io

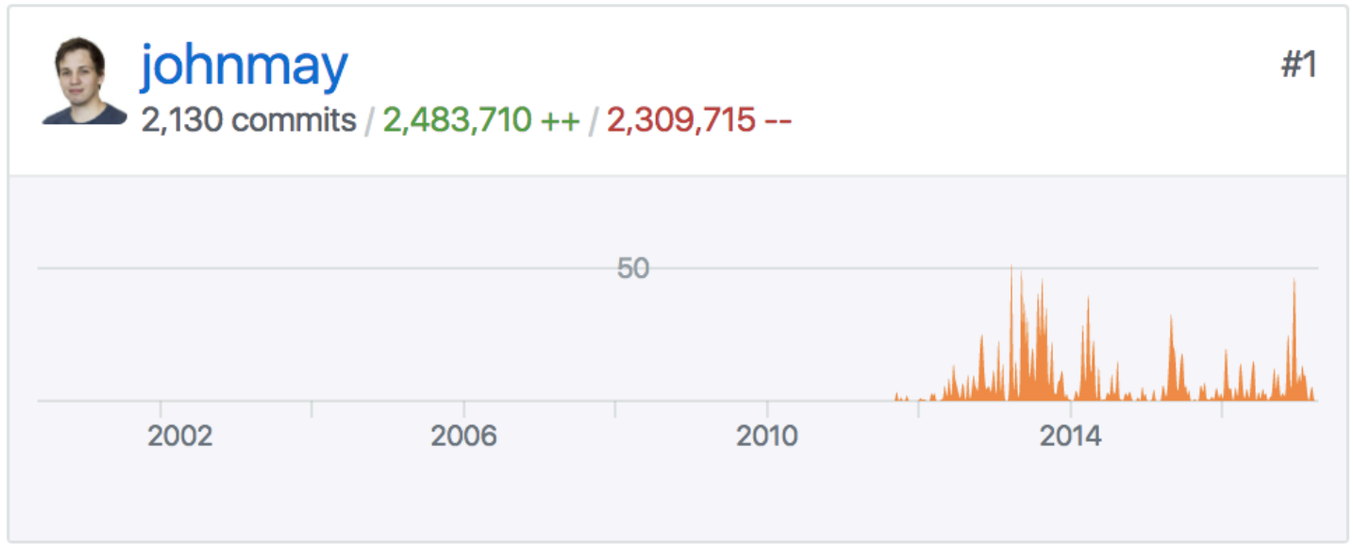
http://efficientbits.blogspot.com

HISTORY

- Project started in **2000**
- **97** contributors, key people:
  - Christoph Steinbeck, Egon Willighagen, Dan Gezelter, Rajarshi Guha
- Originally based on Christoph's "compchem" (1997)
- Lineage in Compute Assisted Structure Elucidation

ABOUT ME

- Joined CDK Project 2012
  - During “Cheminformatics for Genome-Scale Metabolic Reconstructions” BBSRC-CASE (Unilever) funded PhD
- At NextMove Software since Oct 2014



Thesis: https://www.repository.cam.ac.uk/handle/1810/246652

CDK IN ACTION

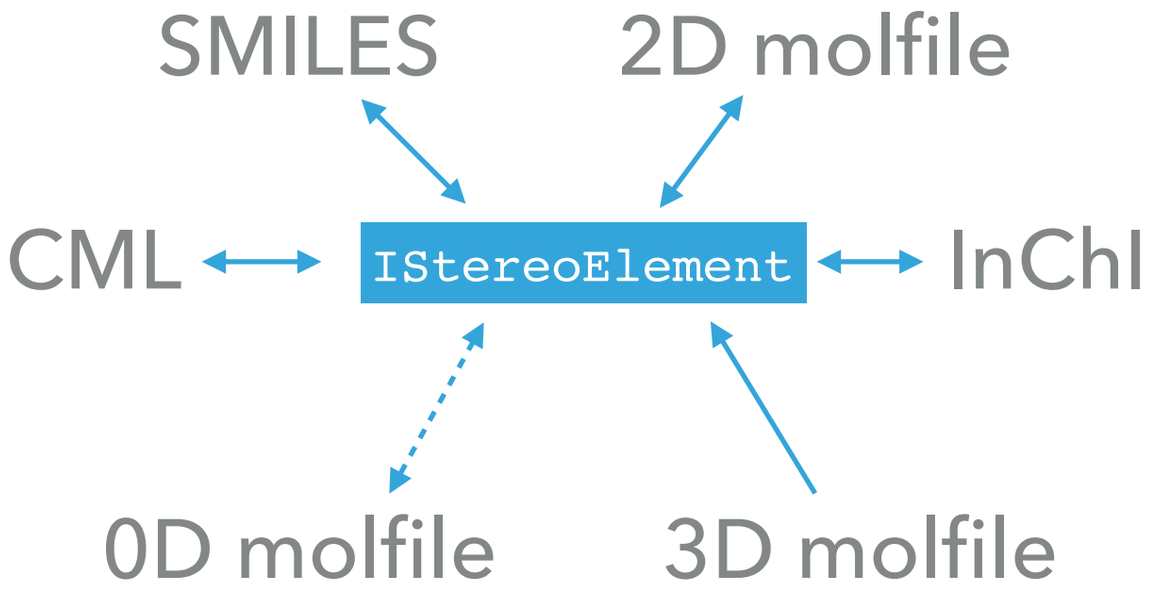
- Stephan Beisken (EMBL-EBI). **KNIME-CDK**: Workflow-driven cheminformatics.
- Rajarshi Guha (Indiana/NIH). **RCDK**: Chemical Informatics Functionality in R
- Chun Wei Yap (National University of Singapore), **PaDEL-descriptor**: molecular descriptors and fingerprints
- Ola Spjuth, **Bioclipse**: Workbench for life science
- Kazuya Ujihara, **NCDK:** .NET port
- TU Dortmund, **Scaffold Hunter**: Visual analysis of large and complex data sets
- IDEA consult Ltd. **AMBIT**: open-source cheminformatics data management
- Scooter Moris (UCSF), **chemViz2**: Cheminformatics Plugin for Cytoscape
- Pistoia Alliance, **HELM 2.0**: Hierarchical Editing Language for Macromolecules
- NextMove Software Ltd. **Casandra**: Real time alerting server for Electronic Lab Notebooks
- Kevin Lawson (Syngenta), **LICSS System** (excel-cdk): MS Excel Integration
- Edmund Duesbury (University of Sheffield), et al. **MCS-based Data Fusion in Similarity Searching**
- **etc…**

THIS TALK

- Key **improvements** and **innovations** in **2.0**:
  - Speed many core algorithms replaced or optimised
  - Quality correctness of results
  - Robustness tolerance and recovery to unexpected input
  - Testing continuous integration of passing 20,000+ test
- **Future plans -** what's missing

STEREOCHEMISTRY I

- Disparate support pre **2.0**
- **Central** representation



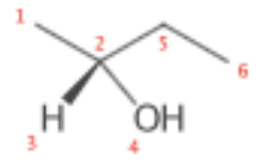
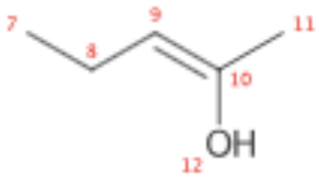
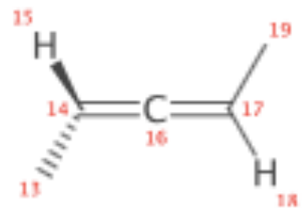
	Type	Focus	Carriers	Configuration
	Tetrahedral	2	1,5,3,4 1,5,4,3 5,1,3,4	CCW CW CW
	Cis-trans	9-10	8-9,10-11 8-9,10-12	Opposite Together
	Extended Tetrahedral	16	15,19,13,18 19,15,13,18	CCW CW

Fig. 2 Relative storage of stereochemistry, the type and focus of stereochemistry are fixed for a given stereocenter description but the carriers and configuration are relative. The multiple rows for each stereochemistry type are different internal representation that would be considered equivalent. In the tetrahedral types, hydrogens may be suppressed in a molecular graph so the focus is reused in the carriers list as a placeholder

Image: E Willighagen *et* al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. ***J. Cheminf*** (2017) 9:33

STEREOCHEMISTRY II

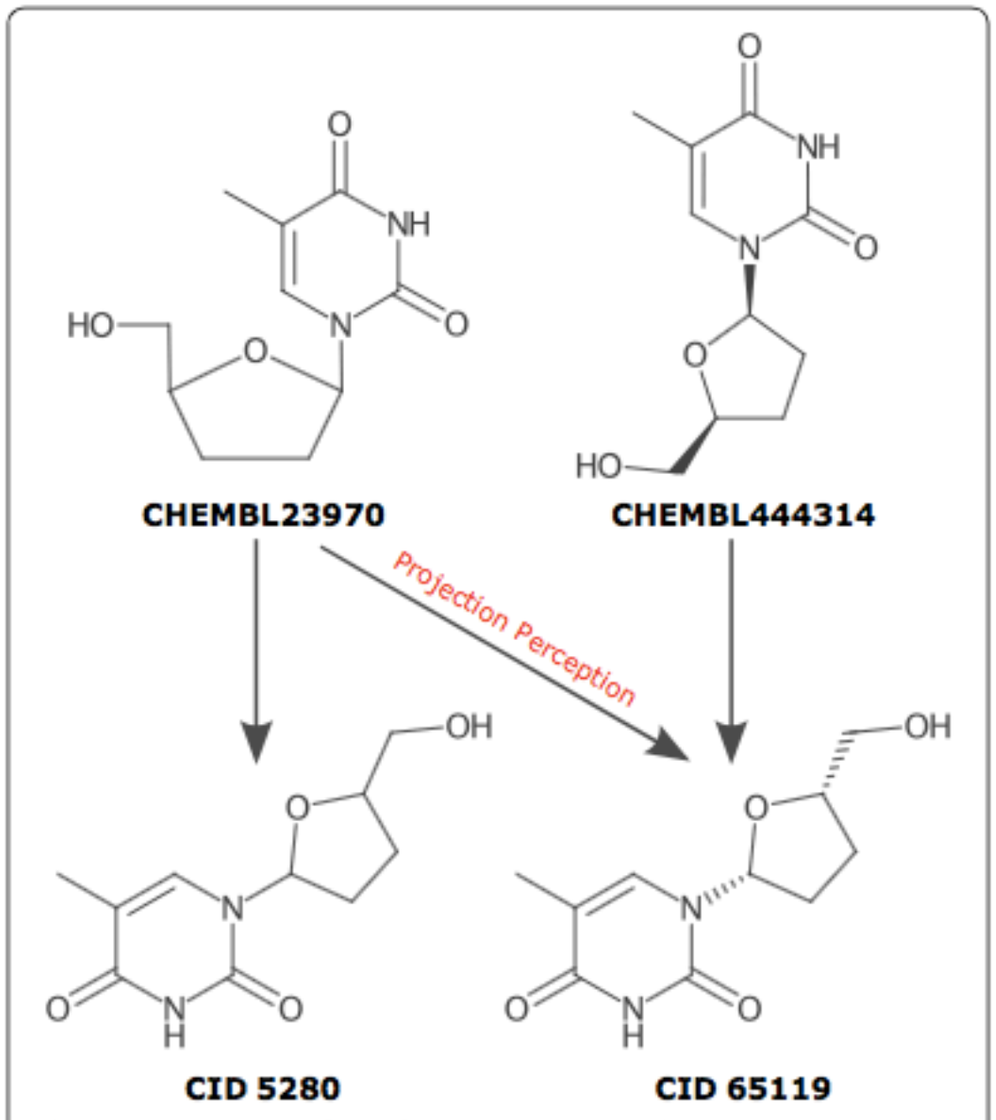


Fig. 3 The raw input files of CHEMBL23970 and CHEMBL444314 are displayed (ChEMBL 21). Without perceiving the stereochemistry indicated by Haworth projection in CHEMBL23970, the database entries are incorrectly considered distinct. Down stream aggregation databases mirror this separation (PubChem CID 5280, CID 65119)

- Substructure matching
- SMARTS
  - `[C@,Si@@]`
  - `[!@!@@]`
- Haworth, Chair, and Fischer projections

Image: E Willighagen *et* al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. ***J. Cheminf*** (2017) 9:33

SMILES

- Memory footprint matters
- Daylight CIS databases could keep SMILES in **memory**
- Most toolkits ~ same per molecule

	SDfile	SMILES	CDK	RDKit	OB
NCI Aug 00	3.9KB	52	8.5KB	14.9KB	12.1KB
ChEMBL 19	2.1KB	71	11.7KB	20.6KB	16KB
eMolecules 2014	-	63	10.1KB	8.7KB	13.8KB

SMILES

- Beam sub-library for SMILES (github.com/johnmay/beam)
  - Very fast parser and generator
  - Kekulization algorithms

Benchmark	Input	FMT	1.4.19			2.0			
			Skipped	Elapsed	Per Min	Skipped	Elapsed	Per Min	Improve
Count Heavy	ChEBI 149	smi	2112	22.51s	108.2K				
		sdf	0	7.21s	355.4K				
	ChEMBL 22.1	smi	0	8m39.3s	193.9K				
		sdf	0	3m17.29s	510.4K				

SMILES

- Beam sub-library for SMILES (github.com/johnmay/beam)
  - Very fast parser and generator
  - Kekulization algorithms

Benchmark	Input	FMT	1.4.19			2.0			Improve
			Skipped	Elapsed	Per Min	Skipped	Elapsed	Per Min	
Count Heavy	ChEBI 149	smi	2112	22.51s	108.2K	9	0.85s	2.9M	**26.48**
		sdf	0	7.21s	355.4K	25	3s	854.1K	**2.4**
	ChEMBL 22.1	smi	0	8m39.3s	193.9K	9	10.74s	9.4M	**48.35**
		sdf	0	3m17.29s	510.4K	0	53.27s	1.9M	**3.7**

CTFILE SGROUPS

- **Annotation** layer over **part** of a **structure**
- Round-trip via **ChemAxon Extended SMILES**

Display Shortcut

Polymer

Mixture

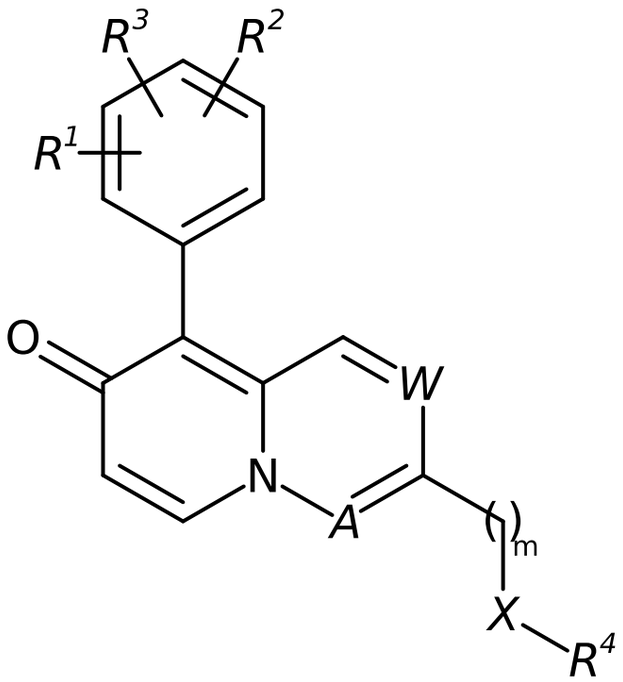
Data

Gushurst *et al*. The substance module: the representation, storage, and searching of complex structures. ***J. Chem. Inf. Comput. Sci.*** (1991)

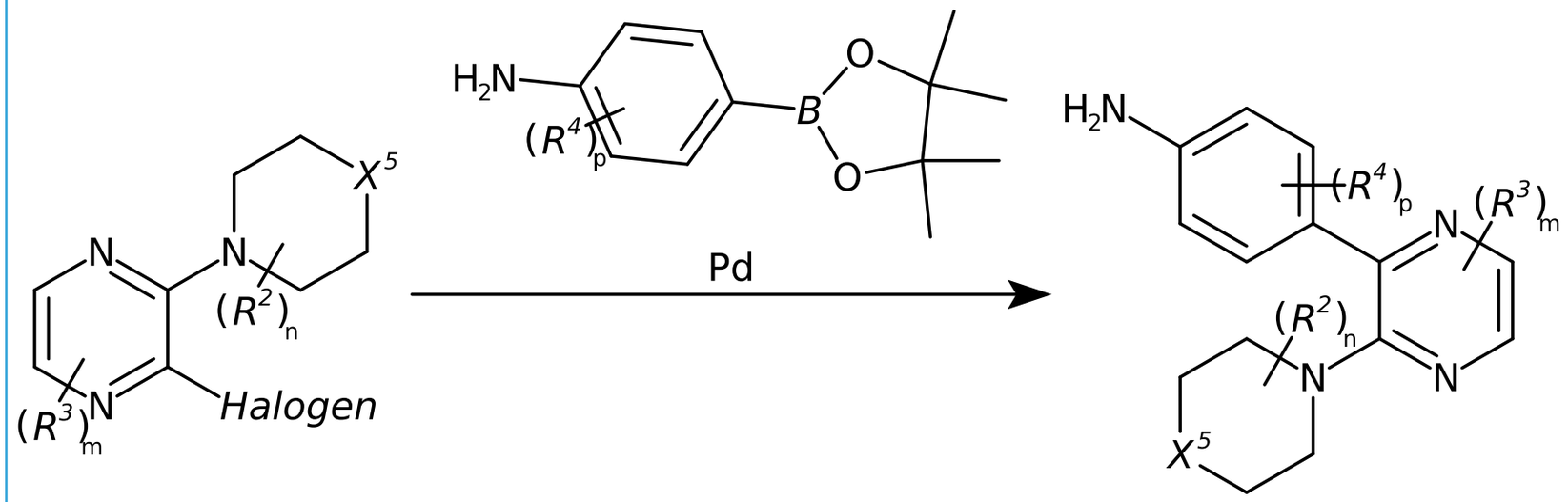
Blanke G. Sgroups – Abbreviations, Mixtures, Formulations, Polymers, Structures with Statistical Distribution and Other Special Cases. ***Online - StructurePendium Technologies GmbH***

SGROUP LAYOUT AND RENDERING

Generated from ChemAxon Extended SMILES (CXSMILES):



US 2007/0129372 (I)

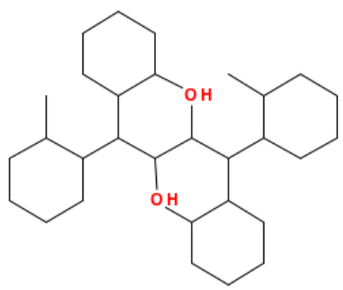
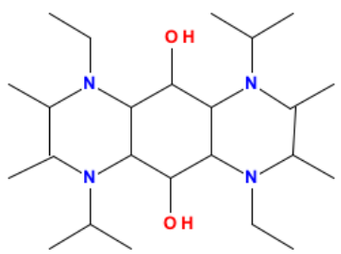
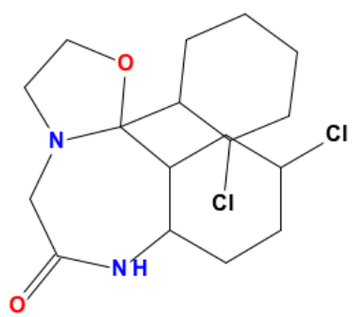
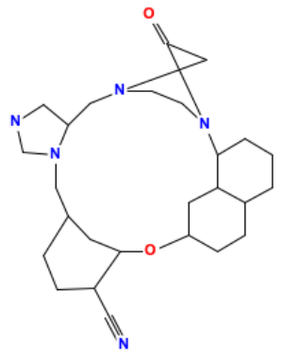
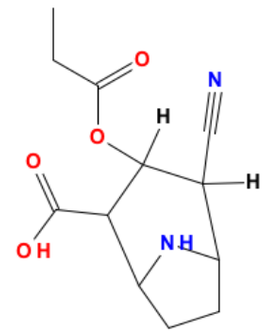
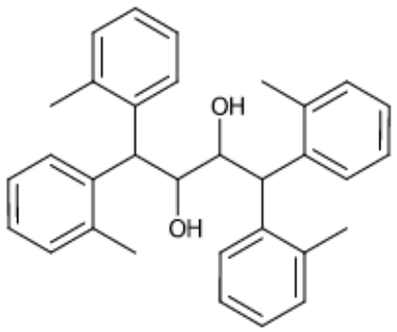
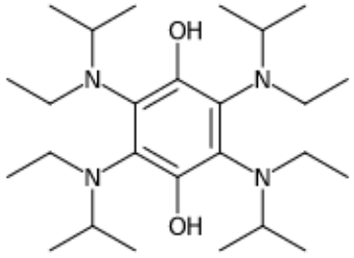
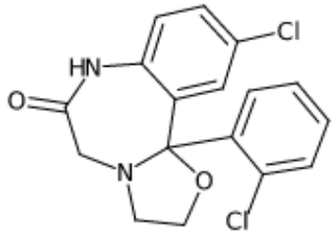
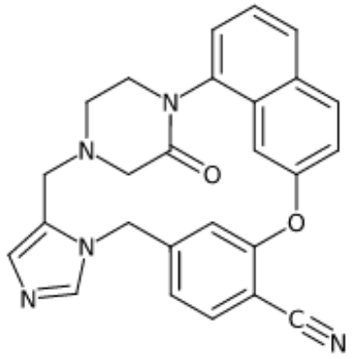
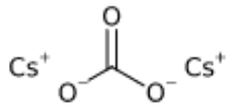
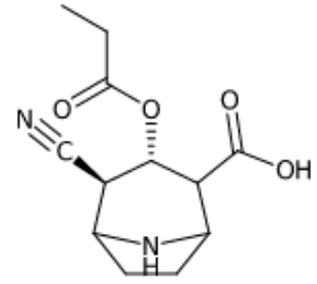


US 2017/233380 A1 (Synthetic scheme 5)

Praline, NextMove Software Ltd

John May, *et al.* Sketchy Sketches: Hiding Chemistry in Plain Sight. Seventh Joint Sheffield Conference on Cheminformatics. 2016

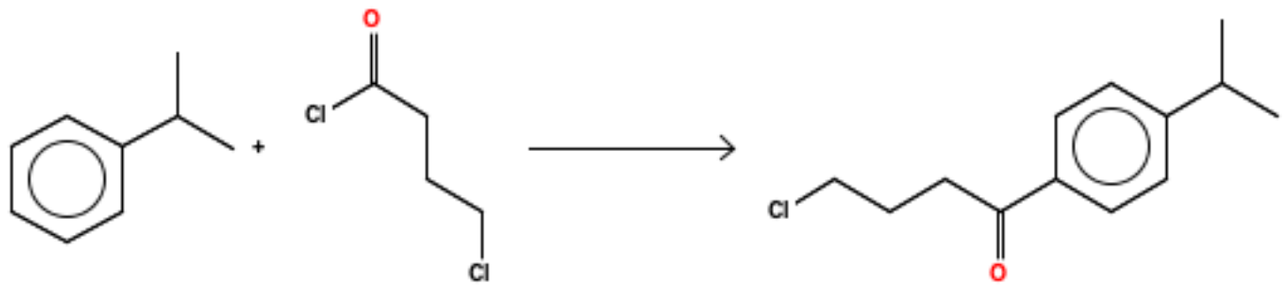
LAYOUT AND RENDERING

1.4.19					ERROR	
2.0						

Try it: http://cdkdepict-openchem.rhcloud.com/

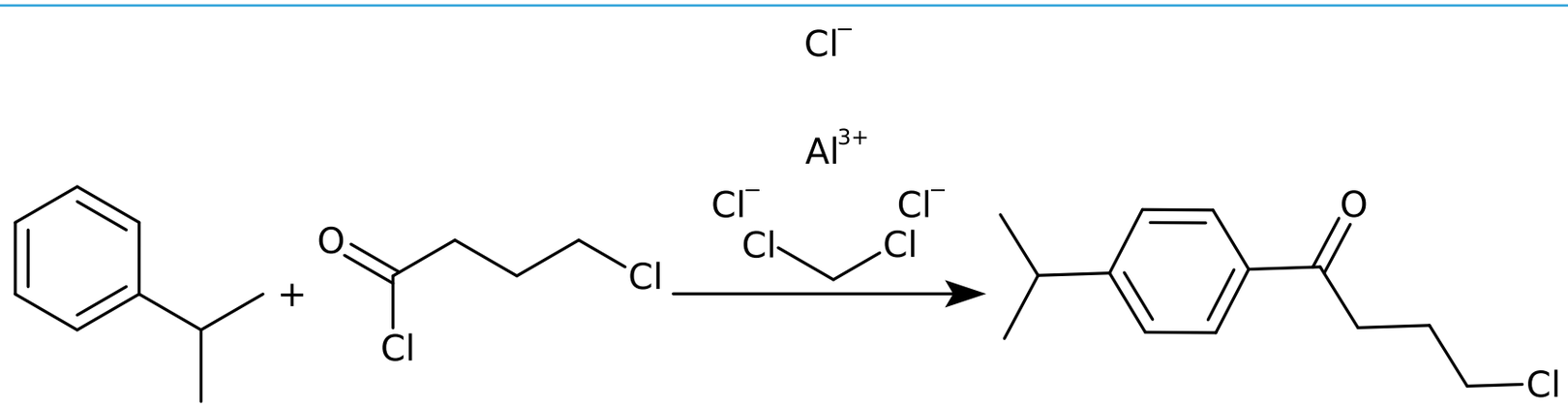
Examples 1-4: Clark A, et al, 2D structure depiction. **JCIM**, 46, 1107-1123 (2006)

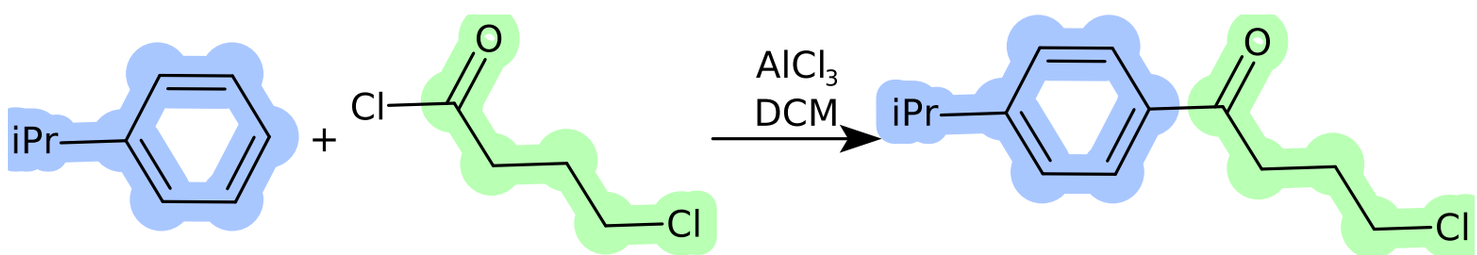
REACTION DIAGRAMS



1.4.19

2.0





2.0++

Try it: http://cdkdepict-openchem.rhcloud.com/

Alex Clark. Rendering Molecular Sketches for Publication Quality Output. ***Molecular Informatics***. 2013

DEPICTION API

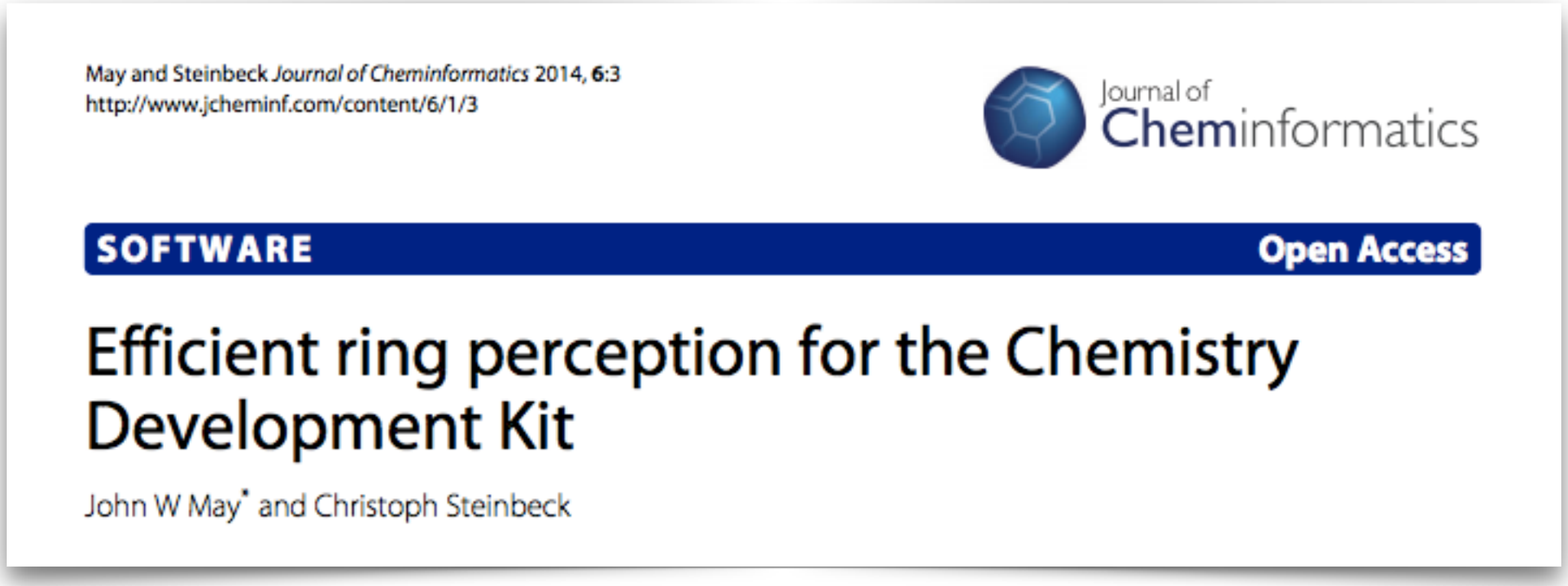
1.4.19

```
StructureDiagramGenerator sdg = new StructureDiagramGenerator();
sdg.setMolecule(mol, false);
sdg.generateCoordinates();

AtomContainerRenderer renderer =
        new AtomContainerRenderer(Arrays.asList(new BasicSceneGenerator(),
                                                new BasicBondGenerator(),
                                                new ExtendedAtomGenerator()),
                                  new AWTFontManager());
BufferedImage img = new BufferedImage(256, 256, BufferedImage.TYPE_4BYTE_ABGR);
Graphics2D g2 = img.createGraphics();

g2.setBackground(Color.WHITE);
g2.clearRect(0, 0, 256, 256);

IDrawVisitor visitor = new AWTDrawVisitor(g2);
Rectangle2D bounds = new Rectangle2D.Double(0, 0, 256, 256);

renderer.paint(mol, visitor, bounds, true);
g2.dispose();

ImageIO.write(img, "PNG", new File("myoutput.png"));
```

2.0

```
new DepictionGenerator().withSize(256, 256)
                        .depict(mol)
                        .writeTo("molecule.png");
```

RING PERCEPTION



May and Steinbeck *Journal of Cheminformatics* 2014, **6**:3
http://www.jcheminf.com/content/6/1/3

Journal of Cheminformatics

SOFTWARE Open Access

Efficient ring perception for the Chemistry Development Kit

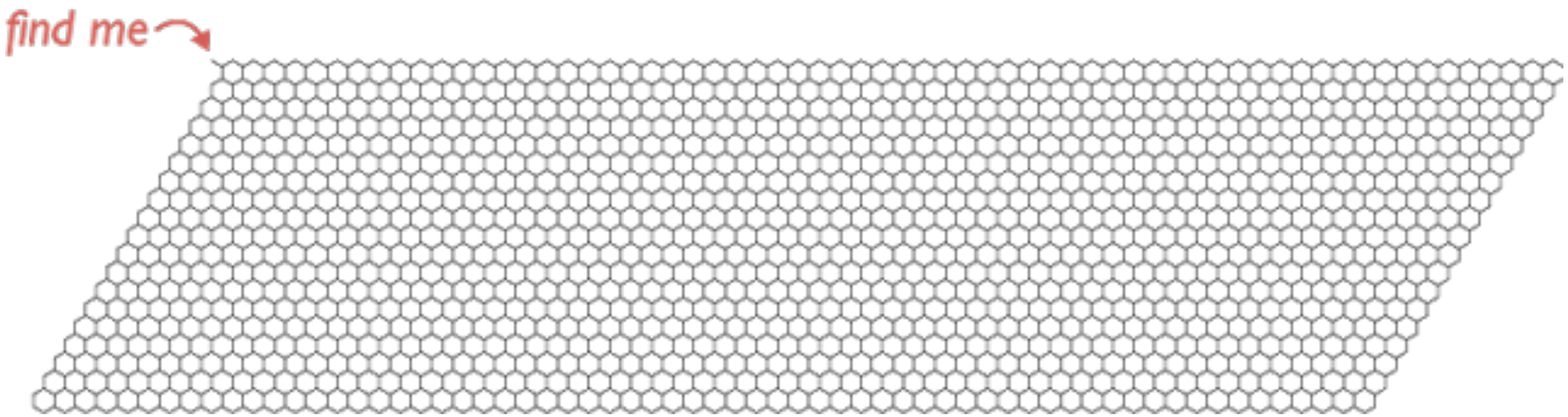
John W May* and Christoph Steinbeck

Rewrote and optimised:

- **Membership** - *fundamental*
- **Cycle basis** (SSSR, Elementary Cycles, Relevant Cycles)
- **All Rings** - *useful for delocalising rings*

RING PERCEPTION II

- Stress test



- Was **73** seconds
- Now **0.011** seconds

Efficient Bits **2012** - http://efficientbits.blogspot.com/2012/12/scaling-up-faster-ring-detection-in-cdk.html

RING PERCEPTION II

			1.4.19			2.0			
Benchmark	**Input**	**FMT**	**Skipped**	**Elapsed**	**Per Min**	**Skipped**	**Elapsed**	**Per Min**	**Improve**
Membership	ChEBI 149	smi	2112	22.91s	106.3K				
		sdf	0	8.71s	294.2K				
	ChEMBL 22.1	smi	0	8m45.78s	191.5K				
		sdf	0	4m12.01s	399.6K				
SSSR	ChEBI 149	smi	2112	27.4s	88.9K				
		sdf	0	11.84s	216.4K				
	ChEMBL 22.1	smi	0	12m4.62s	139K				
		sdf	0	7m9.58s	234.4K				
All Rings	ChEBI 149	smi	2126	45.28s	53.8K				
		sdf	16	36.56s	70.1K				
	ChEMBL 22.1	smi	88	12m40.2s	132.5K				
		sdf	90	8m5.64s	207.4K				

RING PERCEPTION II

Benchmark	Input	FMT	1.4.19 Skipped	1.4.19 Elapsed	1.4.19 Per Min	2.0 Skipped	2.0 Elapsed	2.0 Per Min	Improve
Membership	ChEBI 149	smi	2112	22.91s	106.3K	9	1.06s	2.3M	**21.61**
		sdf	0	8.71s	294.2K	25	3.11s	823.9K	**2.8**
	ChEMBL 22.1	smi	0	8m45.78s	191.5K	9	17.09s	5.9M	**30.77**
		sdf	0	4m12.01s	399.6K	0	1m6.54s	1.5M	**3.79**
SSSR	ChEBI 149	smi	2112	27.4s	88.9K	9	1.43s	1.7M	**19.16**
		sdf	0	11.84s	216.4K	25	3.78s	677.8K	**3.13**
	ChEMBL 22.1	smi	0	12m4.62s	139K	9	27.16s	3.7M	**26.68**
		sdf	0	7m9.58s	234.4K	0	1m8.17s	1.5M	**6.3**
All Rings	ChEBI 149	smi	2126	45.28s	53.8K	26	1.26s	1.9M	**35.94**
		sdf	16	36.56s	70.1K	40	3.51s	730K	**10.42**
	ChEMBL 22.1	smi	88	12m40.2s	132.5K	9	24.97s	4M	**30.44**
		sdf	90	8m5.64s	207.4K	0	1m5.68s	1.5M	**7.39**

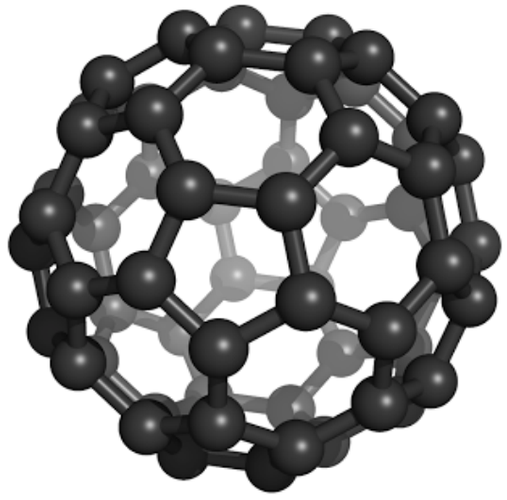
STRUCTURE MATCHING I

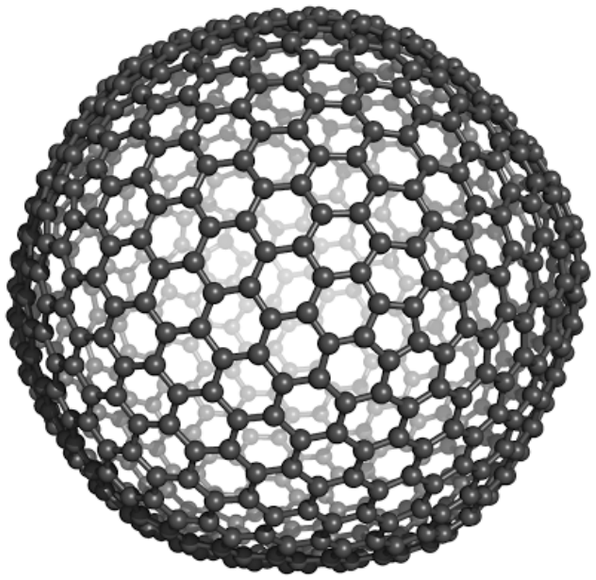
- Previously MCS based `UniversalIsompormismTester`
- Reaction SMARTS (substructure queries)
- New 'lazy' **`Pattern`** API

```
1  Pattern ptrn = Smarts.create("O=[C,N]aa[N,O;!H0]");
2  for (IAtomContainer mol : mols) {
3    boolean b = ptrn.matches(mol);
4    boolean b = ptrn.matchAll(mol).atLeast(5)
5    int     n = ptrn.matchAll(mol).count();
6    int     n = ptrn.matchAll(mol).countUnique();
7    int     n = pattern.matchAll(mol).uniqueBonds().count();
8  }
```

STRUCTURE MATCHING II

- Counting automorphisms





Structure	Matched	UIT	SMSD	Ullmann	VentoFoggia
	n	*t (s)*	*t (s)*	*t (s)*	*t (s)*
ferrocene	200	0.281	0.166 (*n*=180?)	0.007	0.001
fullerene C60	120	279	7.5 (*n*=1)	1.2	0.017
fullerene C70	60	n/a	9.8 (*n*=1)	2	0.058
fullerene C320	120	n/a	n/a	n/a	0.527

Roger Sayle, Efficient matching of multiple chemical subgraphs. 9th ICCS, Noordwijkerhout, **2011**
Efficient Bits, **2013** - http://efficientbits.blogspot.com/2013/11/improved-substructure-matching.html

STRUCTURE MATCHING II

- Find `O=[C,N]aa[N,O;!H0]` in **NCI Aug 00** (250K)
- "Death by a thousand paper cuts" - majority of gain from overheads (String finding, aromaticity, SMILES)

Version	Matched	Errors	Time (Machine A)	Time (Machine B)
1.4.15	11,145	64	819 s 860 ms	
1.5.3	11,145	37	279 s 725 ms	
1.5.4a	11,146	37	225 s 639 ms	
1.5.4b	10,136	37	160 s 713 ms	
1.5.4c	10,136	0	16 s 18 ms	8 s 126 ms
2.0	10,136	0		7 s 446 ms

Improvement from last slide

D Weininger. Chemistry Cartridge CGI Examples. **EMug '98**
Efficient Bits, **2013** - http://efficientbits.blogspot.com/2013/10/all-small-things.html

FINGERPRINTS

Benchmark	Input	FMT	1.4.19 Skipped	1.4.19 Elapsed	1.4.19 Per Min	2.0 Skipped	2.0 Elapsed	2.0 Per Min	Improve
Path Fingerprints	ChEBI 149	smi	2112	1m38s	24.9K	9	10.28s	236.9K	**9.53**
		sdf	0	2m11.03s	19.6K	25	13.03s	196.6K	**10.06**
	ChEMBL 22.1	smi	0	42m56.15s	39.1K	9	6m34.67s	255.2K	**6.53**
		sdf	0	47m5.58s	35.6K	0	7m52.32s	213.2K	**5.98**
MACCS 166	ChEBI 149	smi	2150	1h37m35s	416	9	19.51s	124.8K	**300.1**
		sdf	48	1h44m17s	409	25	21.25s	120.6K	**294.45**
	ChEMBL 22.1	smi	214	20h24m57s	1.4K	9	13m31.21s	124.1K	**90.6**
		sdf	225	24h41m46s	1.1K	0	13m26.41s	124.9K	**110.25**
Circular Fingerprint	ChEBI 149	smi	0	NA	-	9	4.37s	557.4K	
		sdf	0	NA	-	25	6.81s	376.2K	
	ChEMBL 22.1	smi	0	NA	-	9	2m43.45s	616.1K	
		sdf	0	NA	-	0	3m42.01s	453.6K	

Circular Fingerprint - Clark A, *et al.* New target prediction and visualization tools incorporating open source molecular fingerprints for TB Mobile 2.0. ***J. Cheminf*** (2014) 6(1):38

FUTURES

- New connection table implementation (in progress)
  - Minor write performance trade-off, major read performance gain (see blog)
- More types of stereochemistry (in progress)
  - Complete: Atropisomerism
  - Planned: Square-Planar, Octahedral, etc
  - Maybe: Enhanced stereochemistry?
- Structure Transforms (SMIRKS)
  - Currently provided by AMBIT
- 3D Coordinate Generation - Can you help?

CONCLUSIONS

- Many improvements in **CDK 2.0**
- Performance now competitive with commercial toolkits
  - Relative comparisons are deceptive
- Progress required
  - Robustness
  - APIs - some still very complex/convoluted
  - Training materials
  - Push performance even further…

ACKNOWLEDGEMENTS

CDK Project

- Egon Willighagen
- Christoph Steinbeck
- Stephan Beisken
- Venkata Chandrasekhar Nainala
- Rajarshi Guha
- All CDK Contributors and Users

Discussions/Papers

- Roger Sayle (NextMove Software)
- Noel O'Boyle (NextMove Software)
- Daniel Lowe (MineSoft)
- Alex Clark (CDD)