CHEMINFORMATICS TOOLKITS: A PERSONAL PERSPECTIVE

ROGER SAYLE

NEXTMOVE SOFTWARE LTD
CAMBRIDGE UK

1st RDKit User Group Meeting, London, UK, 4th October 2012

OVERVIEW

- Models of Chemistry
  - Implicit and Explicit Hydrogens
  - Atom types
  - Aromaticity Models
  - Valence Models
- The “mdlbench.sdf” benchmark
- Some words on performance

ABOUT THE AUTHOR

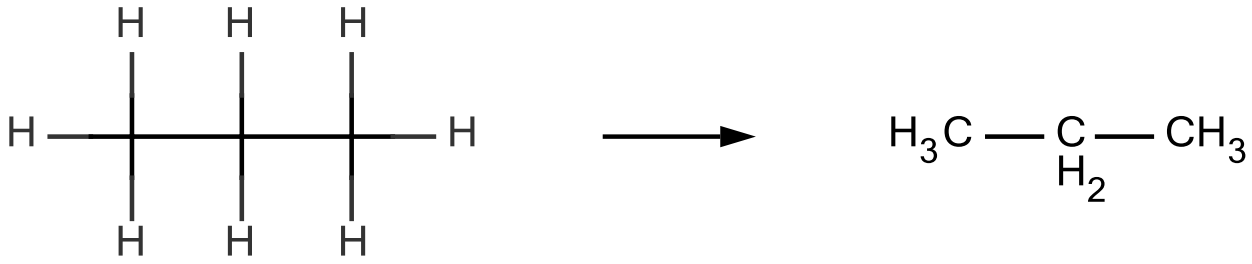
- RasMol, 1989-1997.
- GSK (Glaxo Wellcome), 1994-1997.
- Daylight/Metaphorics, 1998-2001.
- OpenEye Scientific Software, 2001-2010.
- AstraZeneca, Abbott & NCBI, 2009.
- NextMove Software, 2010-

MODELS OF CHEMISTRY

- Algorithms + Data Structures = Programs
- Abstraction of storage and representation from assumptions about the “real world”.

CONNECTION TABLE COMPRESSION

- A distinguishing feature between computational chemistry and cheminformatics is the representation of hydrogens.
- Typically in organic chemistry, approximately half of the bonds in a “full” connection table are the bonds to terminal hydrogen atoms.



HYDROGEN REPRESENTATIONS

- The two representations of the structure are equivalent, and either form can be deduced from the other [or even a hybrid form].
- In OEChem terminology:
  - OESuppressHydrogens(mol)
  - OEAddExplicitHydrogens(mol)
- Confusingly, Daylight's SMILES toolkit auto-supresses hydrogens during dt_modoff, and a bug in OEChem sprouts chiral hydrogens.

HANDLE WITH CARE

- Isotopes of Hydrogen; deuterium, tritium and protium.
- Charged hydrogens.
- Bridging hydrogens.
- Hydrogens with atom maps.
- Hydrogens with none single bonds.

- Preserving chirality on tetrahedral parent.

HYDROGENS IN FILE FORMATS

- Implicit/Explicit is also preserved with file I/O.
- SMILES: [H]C([H])([H])[H] vs C (or [CH4]).
- MDL connection tables:

```
[CH4]
   RDKit

  1  0  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
M  END
$$$$
```

```
OpenBabel10021215582D

  5  4  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
  1  5  1  0  0  0  0
M  END
$$$$
```

HYDROGEN RELATED PROPERTIES

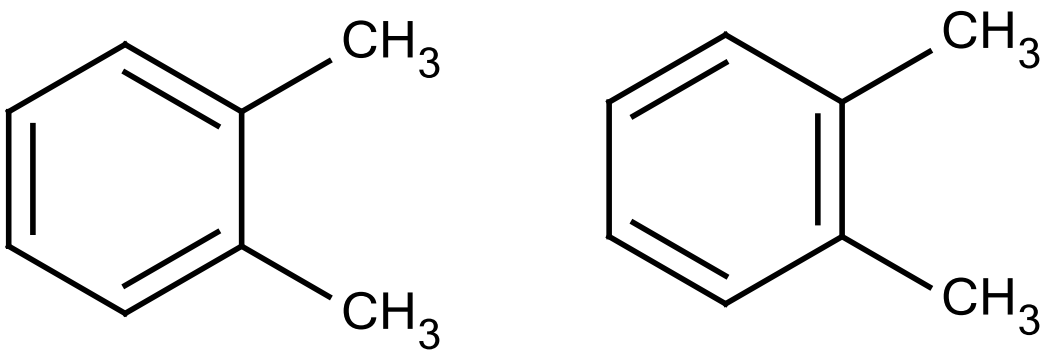
- A number of properties are influenced by the choice of hydrogen representation.
  - NumAtoms()
  - ImplicitHCount() / ExplicitHCount()/TotalHCount()
  - Degree() / ExplicitDegree() / HvyDegree()
  - Valence() / ExplicitValence() / HvyValence()
- Several of these are exposed by SMARTS.

ATOM TYPES

- Another distinction between computational chemistry toolkits and cheminformatics toolkits concerns the roles of “atom types”.
- In cheminformatics, atoms can be described by the atomic number and charge, but in molecular modeling simulations more is required → Domain of applicability.
- Ideally, integer atom types should be an annotation not a fundamental pre-requisite.

AROMATICITY

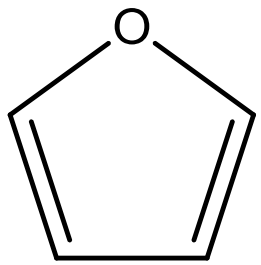
- The primary function of aromaticity in cheminformatics is to treat the multiple Kekulé forms of benzenoid rings as equivalent.



SLIPPERY SLOPE

- Unfortunately the problem with aromaticity is that its a useful physical property of a molecule with uses beyond cheminformatics!
- Conjugated structures behave differently to saturated systems, and hence the degree to which a pi-system is shared and resonance stablized is useful in computational chemistry.

AROMATICITY AND QSAR



Name: Furan
Test: #96
Exptl: 1.34

Original XLogP

R-O-R:	0.327	0.327
R=CHX	-0.166	-0.332
R=CHR	0.236	0.472
H	0.046	0.184
Total:		0.651

Aromatic Furan XLogP

R-O-R	0.327	0.327
R-CH-X	0.142	0.284
R-CH-R	0.281	0.562
H	0.046	0.184
Total:		1.357

AROMATICITY

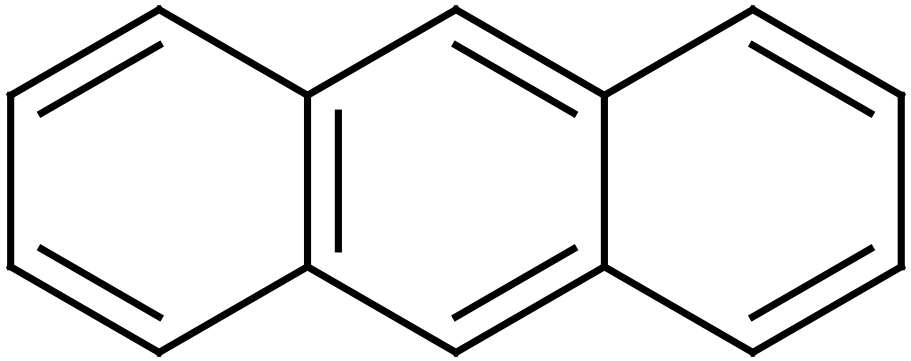
- The upshot is that different notions of aromaticity make sense in different contexts, with no one universally accepted [acceptable] answer.
  - *Aromaticity is one of those unpleasant topics that is simultaneously simple and impossibly complicated. Since neither experimental nor theoretical chemists can agree with each other about a definition, it's necessary to pick something arbitrary and stick to it. This is the approach taken in the RDKit.*

AROMATICITY MODELS

- A better approach is to acknowledge the different (conflicting) definitions of aromaticity and support them much like atom types; using the Tripos aromaticity model for XLogP and handling Sybyl mol2 files, the Daylight aromaticity model for SMILES strings, the MDL aromaticity model for MACCS keys, the CACTVS aromaticity model for PubChem fingerprints and so on.

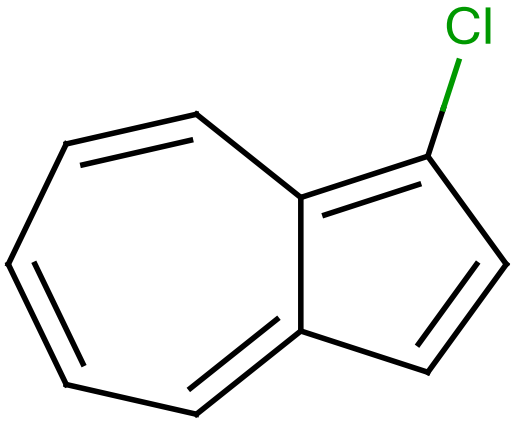
QUINONOID FORMS

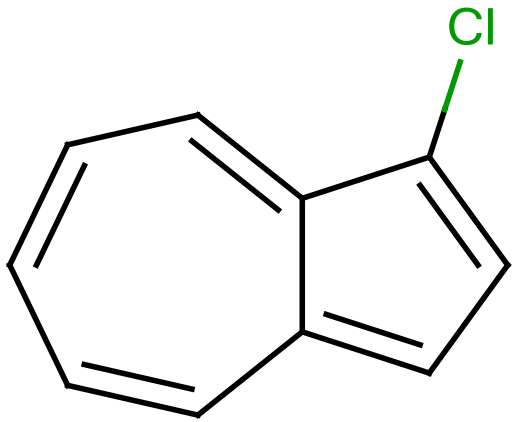
- Unfortunately, quinonoid forms make even the simplest aromaticity models (MDL's) more complex than just alternating single and double bonds around a six membered ring.



LONGER PATHS

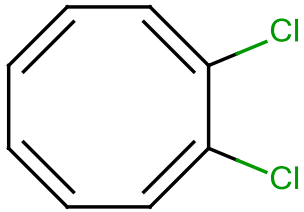
- MDL's aromaticity model doesn't consider azulene to be aromatic.

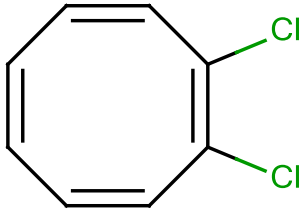


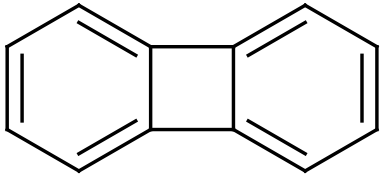


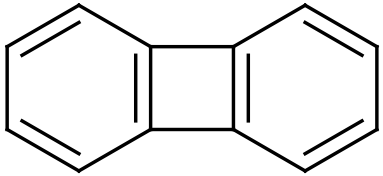
ANTI-AROMATICITY

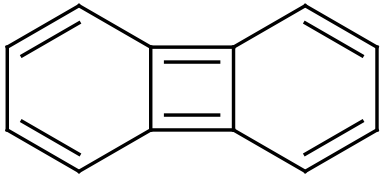
- Some conjugated rings systems perhaps shouldn't be considered equivalent...



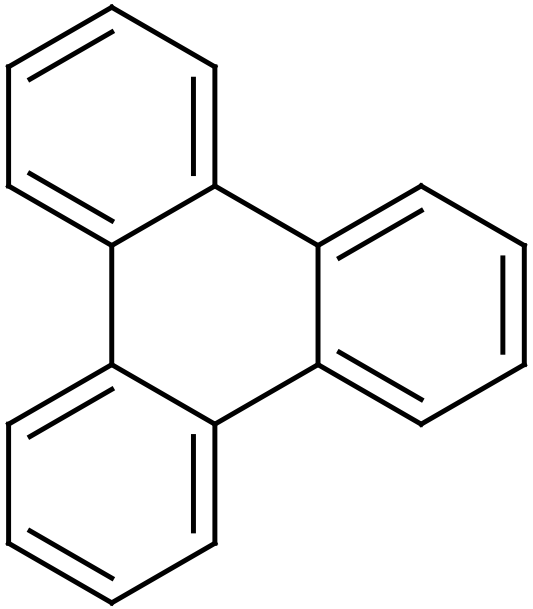


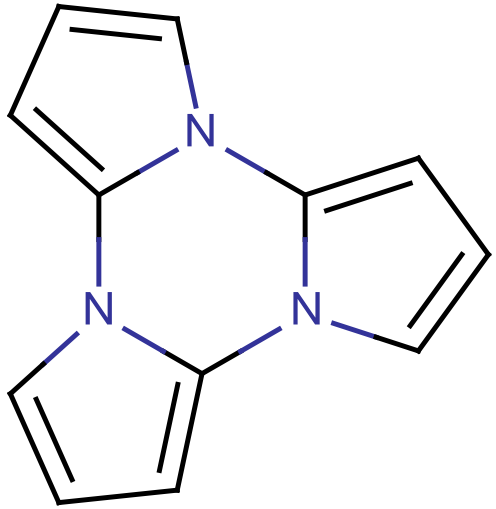




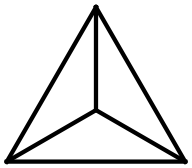


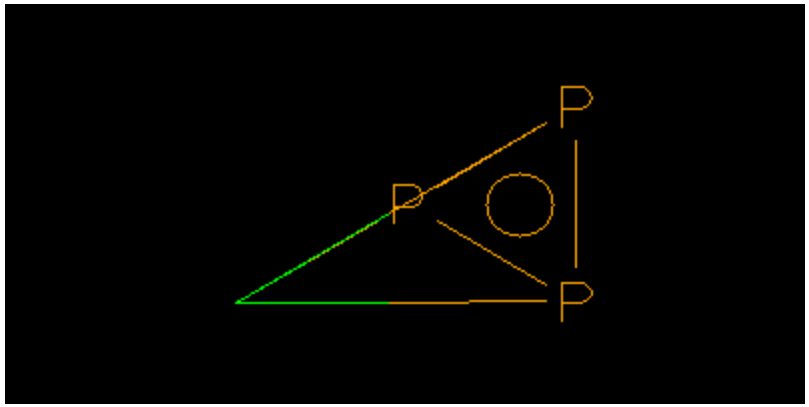
BURIED RING AROMATICITY

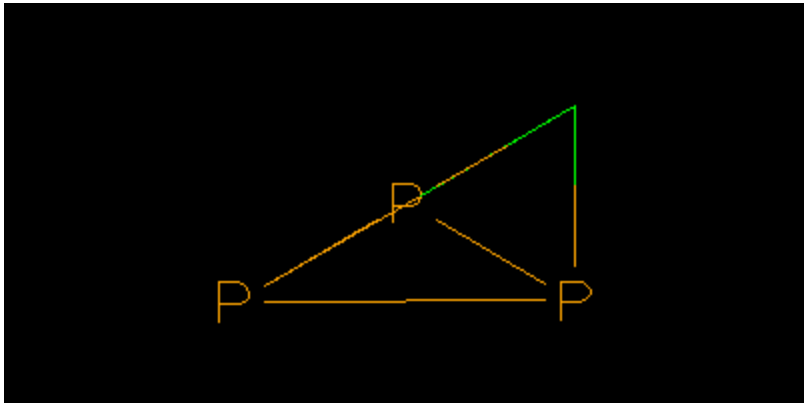




AVOID SSSR FOR AROMATICITY



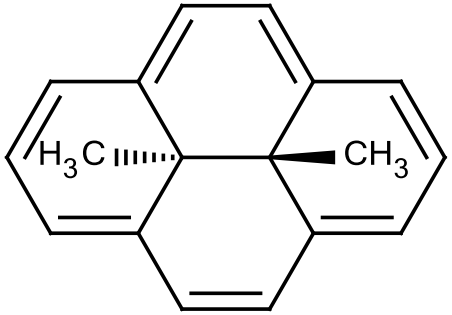


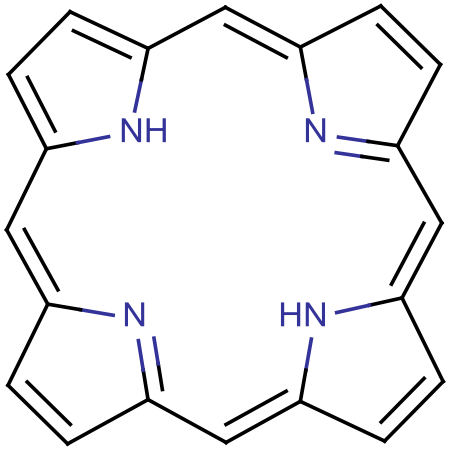


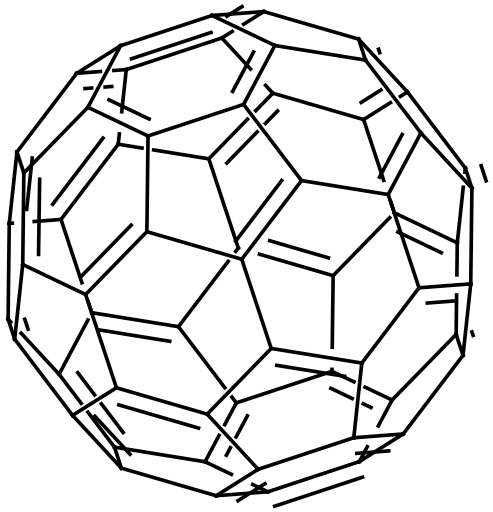
SMILES: C12P3P1P23

SMILES: P12C3P1P23

YOU DON'T WANT SSSR ANYWAY!







AROMATICITY MODELS BY ATOM TYPE

ToolKit	*CH=*	*O*	*N*	*C(=O)*	*As=*	*Te*	*B*	*B=*	*S(=O)*
MDL	1	N	N	N	N	N	N	N	N
Tripos	1	N	N	N	N	N	N	N	N
Merck	1	2	2	N	N	N	N	N	N
Daylight	1	2	2	0	1	N	N	N	2
OEChem	1	2	2	0	1	2	N	N	2
RDKit	1	2	2	0	N	2	N	1	2
ChemAxon	1	2	2	0	1	N	N	N	2
OpenBabel	1	2	2	0	N	N	N	N	2
Indigo	1	2	2	N	1	2!	0	1	N
CDK	1	2	2	N	1	N	N	N	N
Derwent	1			1?					

π Electrons contributed to Huckel 4n+2 calculation

VALENCE MODELS

- To provide life an interesting challenge, developers of cheminformatics file formats that support implicit hydrogens (Daylight and MDL) can optionally omit the number of implicit hydrogens on an atom, to save space, instead relying on the default number of implicit hydrogens for a given atoms environment.

EXAMPLE VALENCE MODEL

- For example, Daylight's valence model for SMILES is that all aromatic nitrogens have no implicit hydrogens by default.
- If a hydrogen is present, it has to be specified as "[nH]".
- This means that "[n]" should never occur in a generated/canonical SMILES string.

THE "MDLBENCH.SDF" BENCHMARK

- To test the fidelity of MDL valence implementations in the SD file readers, we evaluate the SMILES generated from a standard reference test file.
- This SD file contains 10,208 connection tables:
  - 114 different elements plus “D” and “T”
  - 11 charge states (from -4 to +6 inclusive)
  - 8 environments (minimum valences from 0 to 7)
  - 116*11*8 = 10208

MDLBENCH.SDF EXAMPLES

- The correct valence is specified by MDL/ISIS.

- Neutral carbon should be four valent.
  - Everyone agrees on this (hopefully).
- +1 Nitrogen cation should be four valent.
- +1 Carbon cation should be three valent.
- -4 Boron should a have valence 1.
  - OEChem says 0, OpenBabel says 3, RDKit says 7…

MDLBENCH.SDF RESULTS

ToolKit	Failures	Incorrect	Correct	Recall	Precision
OEChem v1.9	264	22	9922	97.20%	99.78%
CDK v1.4.13	264	486	9458	95.11%	95.11%
OpenBabel v2.3.90	176	668	9364	91.73%	93.34%
CACTVS v3.407	352	511	9339	91.49%	94.81%
MDL Direct v8.0	968	22	9218	90.30%	99.76%
ChemAxon v5.10	440	685	9083	88.98%	92.99%
Pipeline Pilot v9.0	968	243	8997	88.14%	97.37%
ChemDraw v12.0	704	548	8956	87.74%	94.23%
GGA Indigo v1.1.4	2797	184	7227	70.80%	97.52%
MOE v2011.10	4221	936	5051	49.48%	84.37%
RDKit v2012_09	4095	4723	1390	13.62%	22.74%

MDLBENCH.SDF SUMMARY

- Although many toolkits can read MDL SD files, they don't all agree on the semantics.
- Different readers, sometimes from the same company, can generate different SMILES from the exact same MOL file.
- Fortunately, most compounds encountered in the pharmaceutical industry fall into the widely understood "well-behaved" subset.

A LITTLE ABOUT PERFORMANCE

- Time to find molecules containing chlorine in the 250,251 SMILES of the NCI August 2000 data set.

ToolKit	Times (secs)	Unfair	Count
UNIX grep	0.06	0.06	43509
OpenEye OEChem v1.8	4.7	2.5	43509
Daylight Toolkit v4.95	8.0		43509
OELib CVS (2002-04-01)	32.8	3.25	43509
ChemAxon JChem v5.10	39.4		43509
GGA Indigo Toolkit v1.1.4	41.6		43508*
RDKit v2011_03_2	109.6		42692*
OpenBabel v2.3.1	148.0		43509
CDK v1.4.13	1278		43501*
PerlMol v0.35	1775		43509

CONCLUSIONS

- There is a one-to-one mapping between MDL connection tables and SMILES, and perfect portable round-tripping should be possible.
- A major step towards this is to separate the chemistry from the computer science in cheminformatics toolkits to permit flexible models of valence and aromaticity.

ACKNOWLEDGEMENTS

- Sorel Muresan and Plamen Petrov, AstraZeneca, Molndal, Sweden.
- Richard Hall, Astex Pharmaceuticals, Cambridge, UK.
- Markus Sitzmann, NCI, Washington DC, USA.
- Daniel Lowe and Noel O'Boyle, NextMove Software, Cambridge, UK.

- Thank you for you time. Questions?