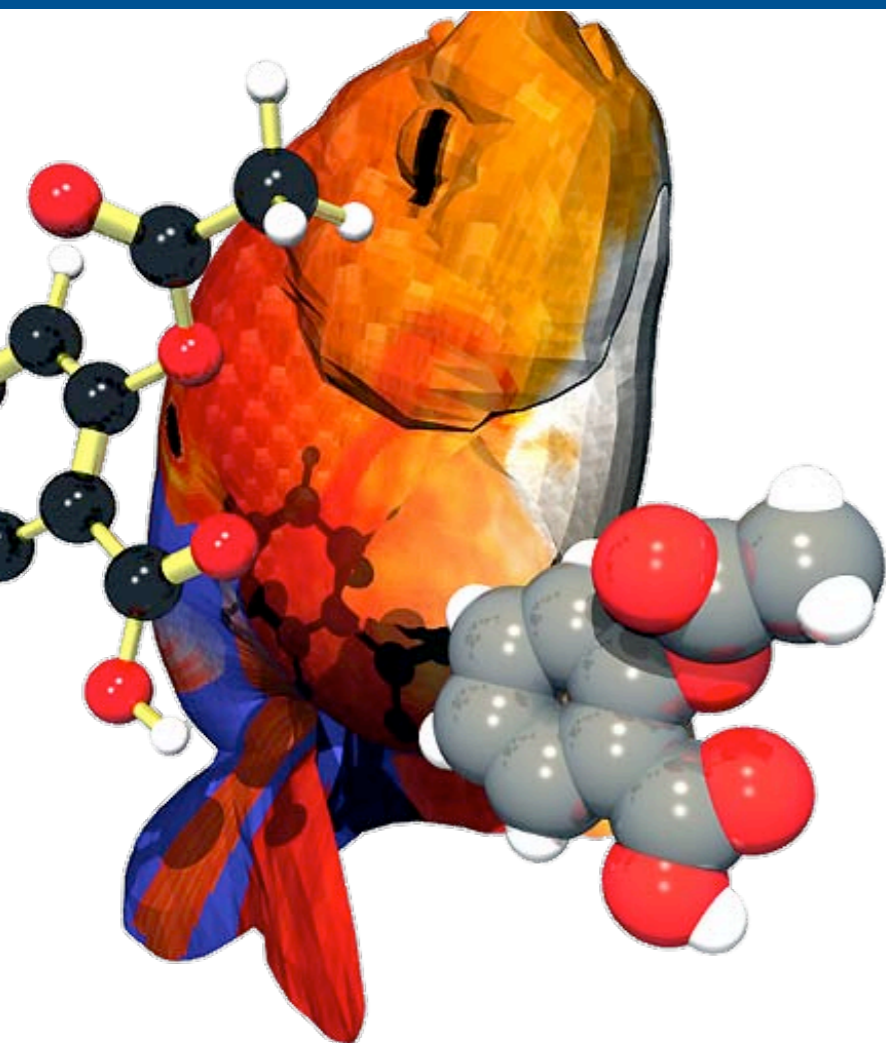


# Calming the Proliferation of Chemical Representations with Open Babel



Dr. Geoffrey Hutchison  
Cornell University

eChemInfo  
Spring 2004

# Acknowledgments

- Pat Walters
- Matt Stahl
- Roger Sayle
- Anthony Nicholls
- Joe Corkery
- Michael Banck
- Chris Morley
- Peter Murray-Rust
- Francesco Bresciani
- Jean Bréfort
- Alex Clark
- Vincent Favre-Nicolin
- Fabien Fontaine
- Malcolm Gillies
- Richard Gillilan



## Open Eye Scientific

- Brian Goldman
- Tommi Hassinen
- Bryan Herger
- Stefan Kebekus
- Erik Kruus
- Eugen Leitl
- David Mathog
- Sergei Pachovsky
- Steffen Reith
- Louis Richard
- Ajay Shah
- Bob Tolbert
- Pawel Wolinski
- Jörg Wegner

# Challenges: A Plethora of Chemical File Formats

## Currently supported input types

alc -- Alchemy file  
prep -- Amber PREP file  
bs -- Ball & Stick file  
caccrt -- Cacao Cartesian file  
ccc -- CCC file  
c3d1 -- Chem3D Cartesian 1 file  
c3d2 -- Chem3D Cartesian 2 file  
cml -- Chemical Markup Language file  
crk2d -- CRK2D: Chemical Resource Kit 2D file  
crk3d -- CRK3D: Chemical Resource Kit 3D file  
box -- Dock 3.5 Box file  
dmol -- DMol3 Coordinates file  
feat -- Feature file  
gam, gamout -- GAMESS Output file  
gpr -- Ghemical Project file  
mm1gp -- Ghemical MM file  
qm1gp -- Ghemical QM file  
hin -- HyperChem HIN file  
jout -- Jaguar Output file  
bin -- OpenEye Binary file  
mmd, mmod -- MacroModel file  
out, dat -- MacroModel file  
car -- MSI Biosym/Insight II CAR file  
sd, sdf -- MDL Isis SDF file  
mdl -- MDL Molfile file  
mol -- MDL Molfile  
mopcrt -- MOPAC Cartesian file  
mopout -- MOPAC Output file  
mmads -- MMADS file  
mpqc -- MPQC file  
bgf -- MSI BGF file  
nwo -- NWChem Output file  
ent, pdb -- PDB file  
pqs -- PQS file  
qcout -- Q-Chem Output file  
ins, res -- ShelX file  
smi -- SMILES file  
mol2 -- Sybyl Mol2 file  
unixyz -- UniChem XYZ file  
vmol -- ViewMol file  
xyz -- XYZ file

## Currently supported output types

alc -- Alchemy file  
bs -- Ball & Stick file  
caccrt -- Cacao Cartesian file  
cacint -- Cacao Internal file  
cache -- CACHe MolStruct file  
c3d1 -- Chem3D Cartesian 1 file  
c3d2 -- Chem3D Cartesian 2 file  
ct -- ChemDraw Connection Table file  
cht -- Chemtool file  
cml -- Chemical Markup Language file  
crk2d -- CRK2D: Chemical Resource Kit 2D file  
crk3d -- CRK3D: Chemical Resource Kit 3D file  
cssr -- CSD CSSR file  
box -- Dock 3.5 Box file  
dmol -- DMol3 Coordinates file  
feat -- Feature file  
fh -- Fenske-Hall Z-Matrix file  
gamin, inp -- GAMESS Input file  
gcart -- Gaussian Cartesian file  
gau -- Gaussian Input file  
gpr -- Ghemical Project file  
gr96a -- GROMOS96 (A) file  
gr96n -- GROMOS96 (nm) file  
hin -- HyperChem HIN file  
jin -- Jaguar Input file  
bin -- OpenEye Binary file  
mmod, dat, mmd -- MacroModel file  
sd, sdf -- MDL Isis SDF file  
mdl, mol -- MDL Molfile  
mopcrt -- MOPAC Cartesian file  
mmads -- MMADS file  
bgf -- MSI BGF file  
csr -- MSI Quanta CSR file  
nw -- NWChem Input file  
ent, pdb -- PDB file  
pov -- POV-Ray Output file  
pqs -- PQS file  
report -- Report file  
qcin -- Q-Chem Input file  
fix, smi -- SMILES file  
mol2 -- Sybyl Mol2 file  
txyz -- Tinker XYZ file

# Challenges: A Plethora of Chemical File Formats

## Currently supported input types

alc -- Alchemy file  
prep -- Amber PREP file  
bs -- Ball & Stick file  
caccrt -- Cacao Cartesian file  
ccc -- CCC file  
c3d1 -- Chem3D Cartesian 1 file  
c3d2 -- Chem3D Cartesian 2 file  
cml -- Chemical Markup Language file  
crk2d -- CRK2D: Chemical Resource Kit 2D file  
crk3d -- CRK3D: Chemical Resource Kit 3D file  
box -- Dock 3.5 Box file  
dmol -- DMol3 Coordinates file  
feat -- Feature file  
gam, gamout -- GAMESS Output file  
gpr -- Ghemical Project file

## Currently supported output types

alc -- Alchemy file  
bs -- Ball & Stick file  
caccrt -- Cacao Cartesian file  
cacint -- Cacao Internal file  
cache -- CACHe MolStruct file  
c3d1 -- Chem3D Cartesian 1 file  
c3d2 -- Chem3D Cartesian 2 file  
ct -- ChemDraw Connection Table file  
cht -- Chemtool file  
cml -- Chemical Markup Language file  
crk2d -- CRK2D: Chemical Resource Kit 2D file  
crk3d -- CRK3D: Chemical Resource Kit 3D file  
cssr -- CSD CSSR file  
box -- Dock 3.5 Box file  
dmol -- DMol3 Coordinates file

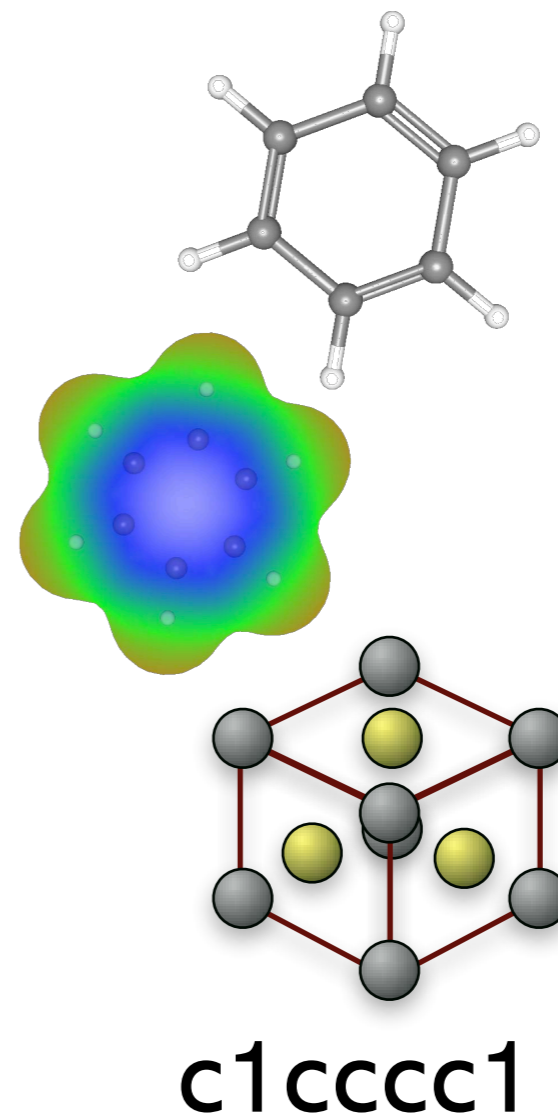
**PLUS: Multiple versions  
Non-standard implementations!**

sd,sdf -- MDL Isis SDF file  
mdl -- MDL Molfile file  
mol -- MDL Molfile  
mopcart -- MOPAC Cartesian file  
mopout -- MOPAC Output file  
mmads -- MMADS file  
mpqc -- MPQC file  
bgf -- MSI BGF file  
nwo -- NWChem Output file  
ent,pdb -- PDB file  
pqs -- PQS file  
qcout -- Q-Chem Output file  
ins,res -- ShelX file  
smi -- SMILES file  
mol2 -- Sybyl Mol2 file  
unixyz -- UniChem XYZ file  
vmol -- ViewMol file  
xyz -- XYZ file

hin -- HyperChem HIN file  
jin -- Jaguar Input file  
bin -- OpenEye Binary file  
mmod,dat,mmd -- MacroModel file  
sd,sdf -- MDL Isis SDF file  
mdl,mol -- MDL Molfile  
mopcart -- MOPAC Cartesian file  
mmads -- MMADS file  
bgf -- MSI BGF file  
csr -- MSI Quanta CSR file  
nw -- NWChem Input file  
ent,pdb -- PDB file  
pov -- POV-Ray Output file  
pqs -- PQS file  
report -- Report file  
qcin -- Q-Chem Input file  
fix,smi -- SMILES file  
mol2 -- Sybyl Mol2 file  
txyz -- Tinker XYZ file

# Challenges: Many Representations of Chemical Data

- Molecular Mechanics:  
Atom & bond types,  
No orbitals
- Quantum Mechanics:  
Atoms (no typing),  
No “bonds”
- Crystallography:  
Fractional coordinates
- Daylight SMILES  
Connectivity only  
No coordinates!



**PLUS: Explicit or implicit hydrogens?  
Different typing rules!**

# What is Babel? (A Brief History)

- **Babel: 1992-1996, Pat Walters & Matt Stahl (U.Arizona)**

*With this program we hope to implement a general framework for converting between file formats used for molecular modeling.*

*Additional options: center molecule, slice multi-molecule files, add/delete hydrogens*

- **OBabel: Pat Walters**
- **OELib: ~2000-2001 Matt Stahl, OpenEye**
- **Open Babel: 2001-Present**

*OpenBabel is a project designed to pick up where Babel left off, as a cross-platform program and library designed to interconvert between many file formats used in molecular modeling and computational chemistry.*

# What is Open Source?

*Open source promotes software **reliability** and quality by supporting **independent peer review** and rapid evolution of source code. To be OSI certified, the software must be distributed under a license that **guarantees** the right to **read, redistribute, modify, and use** the software **freely**.*

*— Open Source Definition (by Open Source Initiative)*

## Keys:

- **Access to source code**
- **Flexibility** — user can modify freely
- **Broad community of developers**
- **Standardizing** — promotes software reuse

# Additional Benefits of Open Source

- **Code reuse: stop reinventing the wheel!**  
*No need to write code for import/export*
- **Public verification and testing:**  
*Both algorithm and implementation in code*
- **User flexibility:**  
*Open file formats  $\Rightarrow$  no vendor “lock-in”*
- **Access to source code:**  
*Anyone can customize, fix bugs, add features...*

**No restrictions on use**  
**Only “restrictions” on distribution**

# Current Features in Summary

- Huge variety of chemical file formats  
*with thorough testing and bug-fixing!*
- Daylight SMARTS matching
- Flexible atom & bond typing
- Connectivity & bond order perception
- Aromatic & ring perception
- Chirality perception
- Gasteiger partial charge calculation
- Hydrogen addition/deletion
- Isotopes & common chemical data
- Cross-platform: Windows, UNIX, Mac...
- *More to come...*

# Current Features in Summary

- Huge variety of chemical file formats  
*with thorough testing and bug-fixing!*
- Daylight SMARTS matching
- Flexible atom & bond typing
- Connectivity & bond order perception

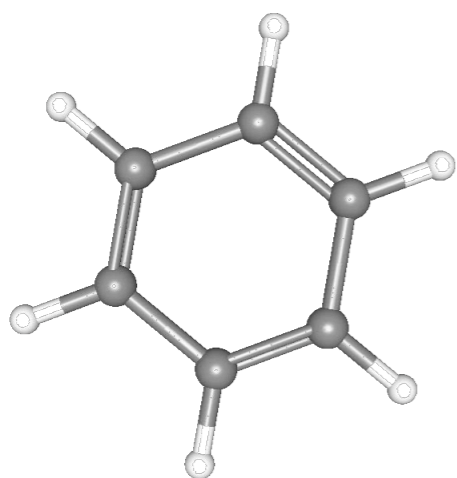
## **Key to implementation:**

**“Lazy perception” of missing data**

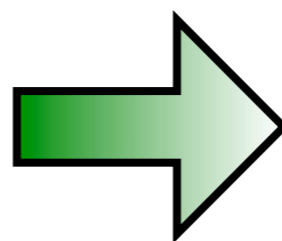
- Gasteiger partial charge calculation
- Hydrogen addition/deletion
- Isotopes & common chemical data
- Cross-platform: Windows, UNIX, Mac...
- *More to come...*

# Lazy Perception in Action...

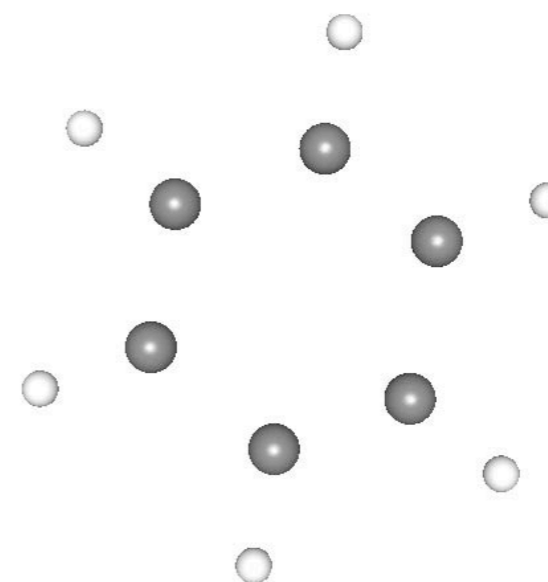
Sybyl Mol2



OBMol



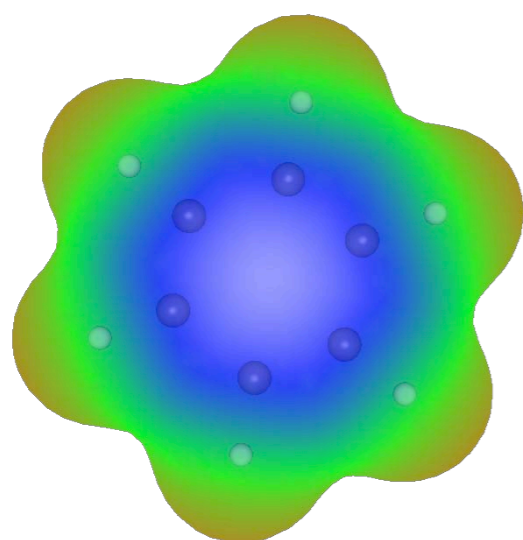
XYZ



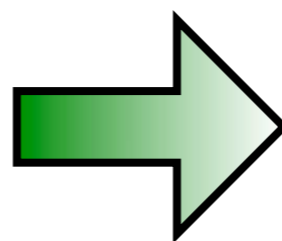
- XYZ format doesn't require partial charges  
Why compute them?
- No residue information, no chains...  
No atom type translation needed!
- Fast output

# Lazy Perception in Action...

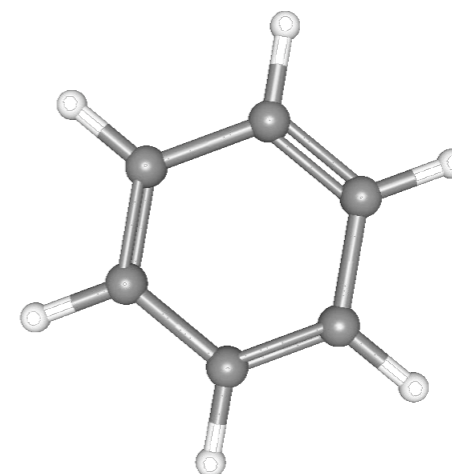
Gaussian 98 Output



OBMol

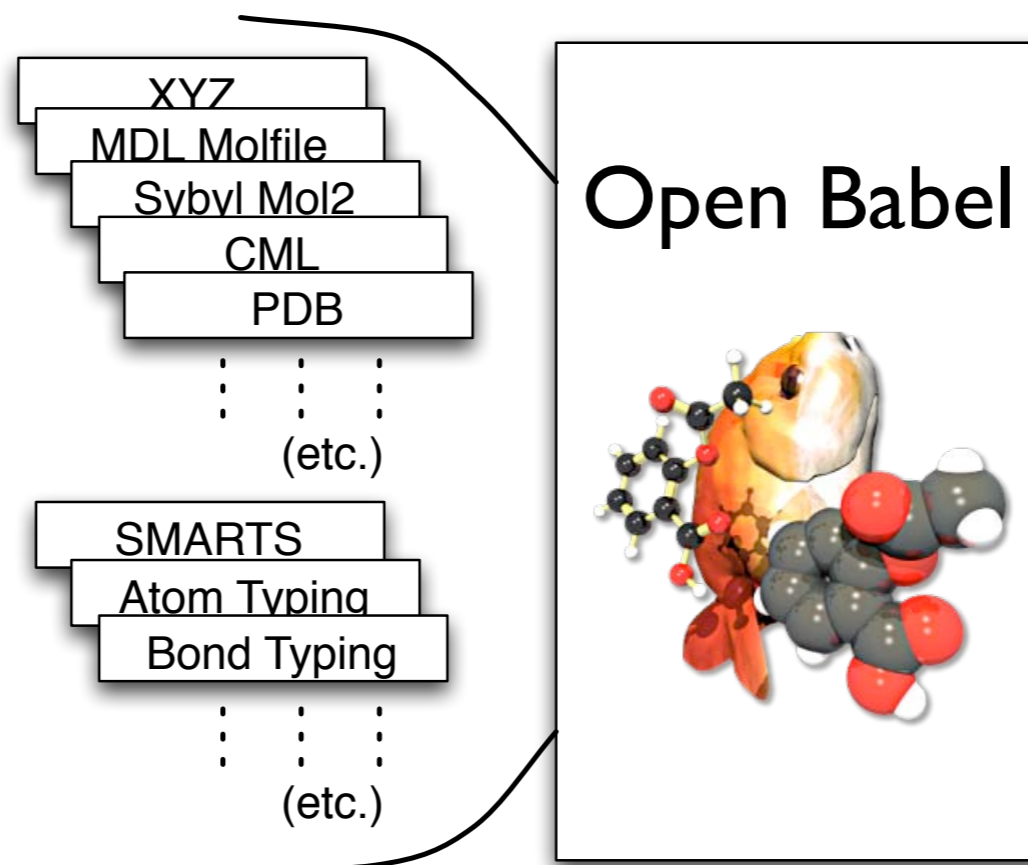


Sybyl Mol2



- Connectivity assignment
- Bond perception needed:  
double bonds, functional groups, aromaticity
- Atom typing & partial charges

# Solving the Chemical Representation “Problem”



- **Whole is greater than the sum of all parts:**  
*No one person handles all file formats*
- **Key goal reflected in “lazy evaluation”**  
*Leave no data behind, but “perceive” as little as possible — conversion should not “create” data!*

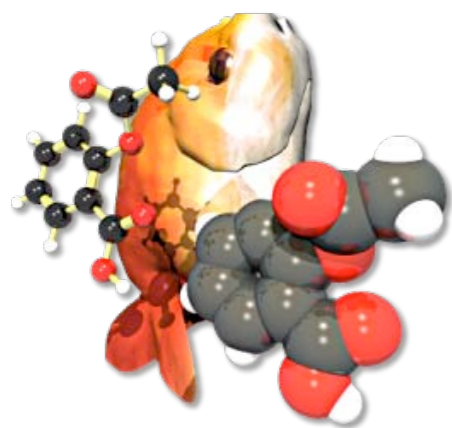
# Solving the Chemical Representation “Problem”

Molecular Editor

QSAR  
Program

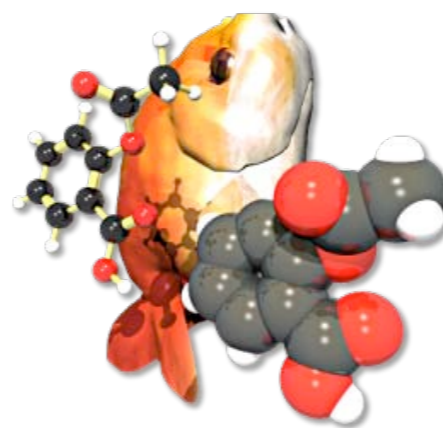
Molecular  
Database

Open Babel

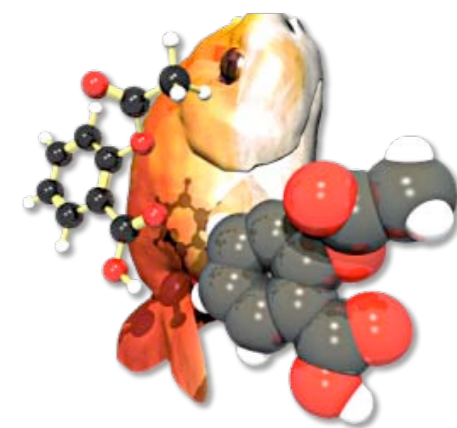


OpenGL  
Graphics

Open Babel



Open Babel



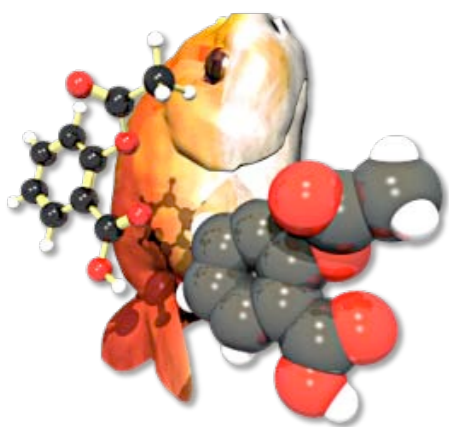
- Code reuse through open source code:  
*Focus on problems **beyond** the basics*  
*New science, not new software development*
- Rapid development
- Reduce non-standard file formats & bugs

# Code-Reuse Example: *obgrep*

## Match Molecular Patterns

Molecular  
Database

Open Babel



- **Total 216 lines of C++ code:**  
*Includes blank lines & comments!*
- **Contributed code, not originally part of Open Babel library**
- **Matches SMARTS molecular patterns in database file(s)**
- **Import/Export handled by Open Babel**  
*“Database” can be any file format, not just SMILES*

# New Directions and Future Plans

- **Improve “lazy evaluation”**  
*QM  $\Rightarrow$  QM requires no atom or bond typing!*
- **Coordinate refinement for SMILES**  
*User-request for 2D or 3D structure layout*
- **More flexible file format code**  
*Better support for multi-molecule files, trajectories, reactions, etc.*
- **Support for more chemical data**  
*Symmetry, molecular orbitals, charge density...*
- **Support for even more file formats**  
*Leave no orphaned data!*

# Other Related Projects and Links

- **Chemical Development Kit (CDK)**

*<http://cdk.sourceforge.net/>*

- **JOELib**

*<http://joelib.sourceforge.net/>*

•

- **Open Source Initiative**

*<http://opensource.org/>*

- **Open Science Project**

*<http://openscience.org/>*

<http://openbabel.sourceforge.net/>  
[openbabel-discuss@lists.sourceforge.net](mailto:openbabel-discuss@lists.sourceforge.net)