

---

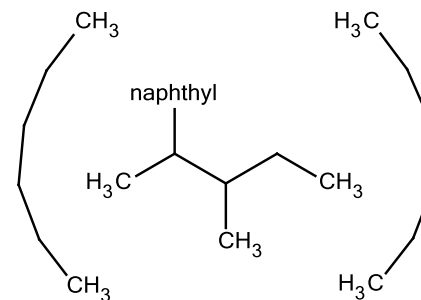
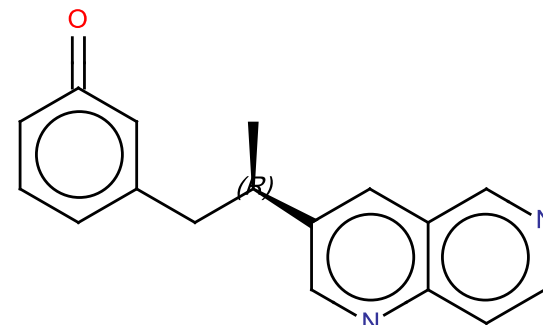
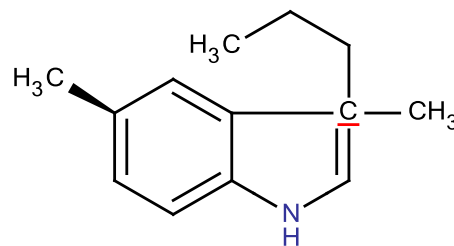
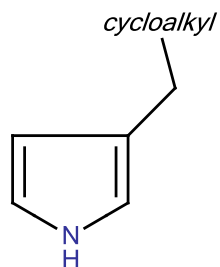
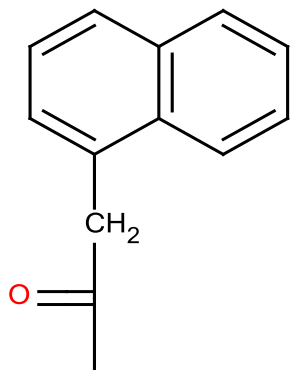
# Structure Checker – *in silico* plastic surgery for molecules

György Pirok, Zsolt Mohácsi, **Wei Deng (David)**



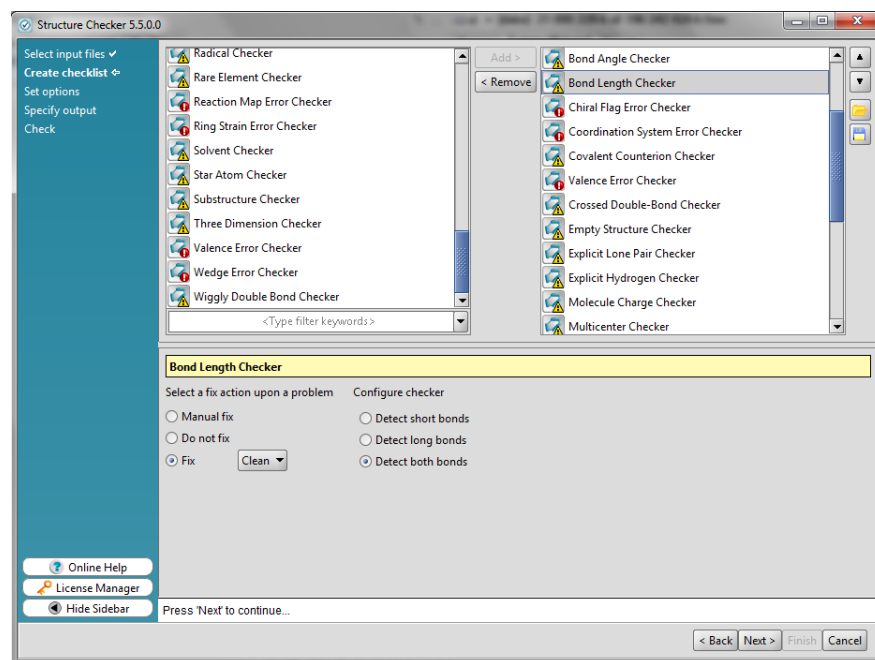
# Structure Issues

- Drawing errors
- OSR Scanning errors
- Inconsistent representations
- Aliases
- etc.



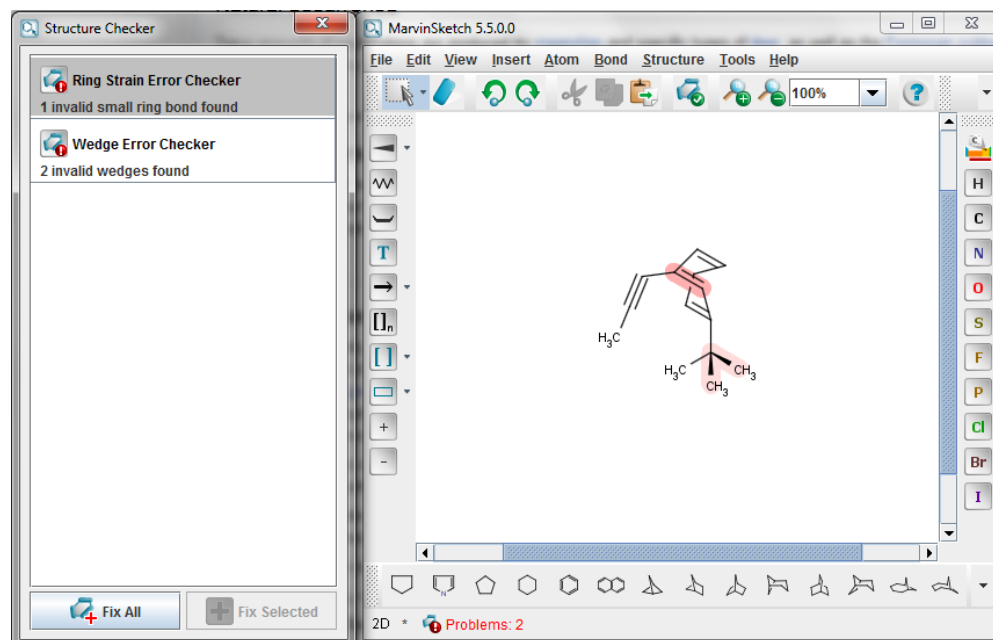
# Structure Checker Overview

- Reporting and/or fixing structure issues
- Manual or automatic
- Configurable checker modules for different issues
  - atom issues (chirality, aliases, overlap, valence...)
  - bond issues (length, angle, crossing...)
  - substructure issues
  - reaction map errors
  - aromaticity, attached data...
- Configurable fixing protocols














# Integration

- Accessibility:
  - standalone GUI and batch mode
  - MarvinSketch
  - KNIME and Pipeline Pilot
  - via *Chemical Terms* in
    - Instant JChem
    - JChem for Excel
    - JChem Cartridge
    - JChem Web Services



# Checkers – error and issue reporting

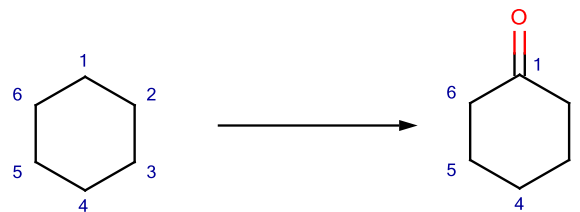
-  Aromaticity error checker
-  Chiral flag error checker
-  Coordination system error checker
-  **Metalocene error checker**
-  **OCR error checker**
-  Ring strain error checker
-  Reaction map error checker
-  Valence error checker
-  Wedge error checker

-  Abbreviated group checker
-  Alias checker
-  Atom map checker
-  **Atom query property checker**
-  Atom value checker
-  Attached data checker
-  Bond angle checker
-  Bond length checker
-  Covalent counterion checker
-  Crossed double bond checker
-  **Empty structure checker**
-  Explicit hydrogen checker
-  **Explicit lone pair checker**
-  Isotope checker
-  Missing atom map checker
-  **Molecule charge checker**
-  Multicenter checker
-  Multicomponent checker
-  Overlapping atoms checker
-  Overlapping bonds checker
-  Pseudo atom checker
-  Query atom checker
-  Query bond checker
-  **Racemate checker**
-  Radical checker
-  **Rare element checker**
-  Solvent checker
-  **Star atom checker**
-  **Substructure checker**
-  Three dimension checker
-  **Unbalanced reaction checker**
-  **Valence property checker**
-  Wiggly double bond checker

# New Checkers 1 (in 5.6)

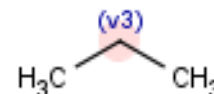
- Unbalanced reaction checker

- Check if the reaction scheme has more atoms on one side than the other
- Fix: manual fixing only



- Valence property checker

- Detect valence property
- Fix: remove



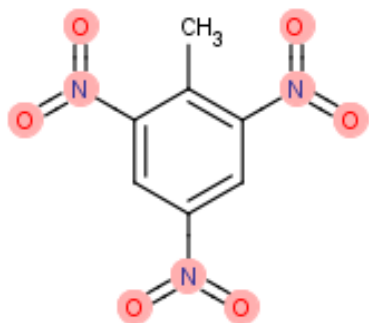
- Atom query property checker

- Detect valence property
- Fix: remove

(a);1 C — H 2

# New Checkers 2

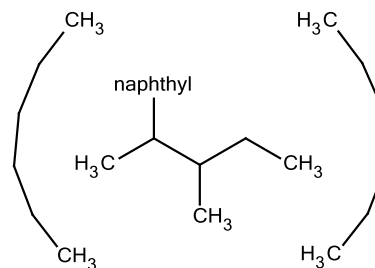
- Substructure checker
  - Find SMARTS-defined structural elements
  - Fix: automatic fix using SMIRKS or manual fix available



The screenshot shows the 'Substructure Checker' interface. It has a yellow header bar with the title 'Substructure Checker'. Below the header, there are two main sections: 'Fixer Options' and 'Checker Options'.  
The 'Fixer Options' section contains three radio buttons: 'Manual fix', 'Do not fix', and 'Fix'. The 'Fix' option is selected. To the right of these options is a 'Transform' button with a dropdown arrow.  
The 'Checker Options' section contains a text box labeled 'SMARTS' with the following text: [\*:1][N:2](=[O:3])=[O:4]>>[\*:1][N+:2](=[O:3])[O-:4]

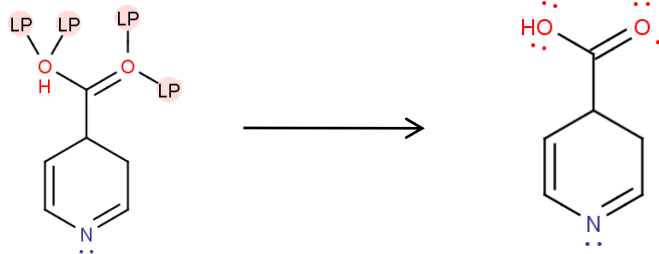
# New Checkers 3

- OCR error checker
  - Errors stemming from misinterpretation of characters in OCR processes



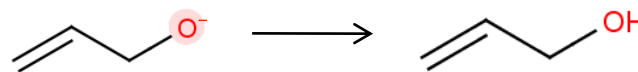
- Empty structure checker
  - Find and handle empty structure fields in multiple structure files

- Explicit lone pair checker
  - Find explicitly drawn lone pairs
  - Fix: remove

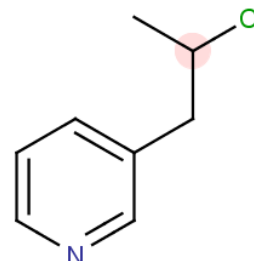


# New Checkers 4

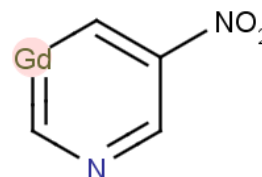
- Molecule charge checker
  - Finds molecules with non zero net charge
  - Fix: remove charge by addign/removing hydrogens



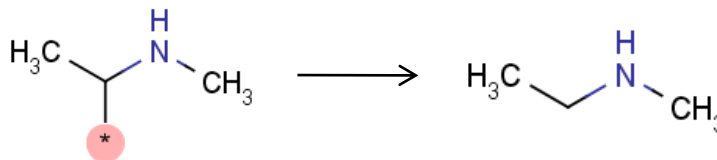
- Racemate checker
  - Finds molecules with chiral centers with no specific stereo configuration
  - Fix: only manual fixing available



- Rare element checker
  - Finds rare elements in structures
  - Fix: only manual fixing available



- Star atom checker
  - Finds star atoms
  - Fix: remove



# Configuration options

- Checker configuration
  - individual configuration for each checker
- Fixer actions
  - Manual fix
  - Do not fix
  - Fix by rules
- Checker list and configuration options can be saved and reloaded

The screenshot shows a configuration window with two columns of checkers. The left column contains: Bond Length Checker, Chiral Flag Error Checker, Coordination System Error Checker, Covalent Counterion Checker, Crossed Double-Bond Checker, Empty Structure Checker, Explicit Hydrogen Checker, Explicit Lone Pair Checker, Isotope Checker, Metallocene Error Checker, and Missing Atom Man Checker. The right column contains: Abbreviated Group Checker, Alias Checker, Aromaticity Error Checker, Atom Map Checker, Atom Value Checker, Attached Data Checker, Bond Angle Checker, Bond Length Checker, Chiral Flag Error Checker, Coordination System Error Checker, Covalent Counterion Checker, and Crossed Double-Bond Checker. Between the columns are 'Add >' and '< Remove' buttons. Below the columns is a search box labeled '<Type filter keywords>'. At the bottom, the 'Abbreviated Group Checker' configuration is shown with two sections: 'Select a fix action upon a problem' (with radio buttons for Manual fix, Do not fix, and Fix) and 'Configure checker' (with radio buttons for Detect Expanded Groups, Detect Contracted Groups, and Detect All Groups). A dropdown menu is set to 'Ungroup'.

# Configuration options – examples

**Bond Length Checker**

Select a fix action upon a problem

Manual fix

Do not fix

Fix

Clean ▾

Configure checker

Detect short bonds

Detect long bonds

Detect both bonds

**Explicit Hydrogen Checker**

Select a fix action upon a problem

Manual fix

Do not fix

Fix

Remove Explicit Hydrogen ▾

Configure checker

Lonely  Charged  Mapped

Isotopic  Radical  Wedged

**Pseudo Atom Checker**

Select a fix action upon a problem

Manual fix

Do not fix

Fix

Convert to Carbon ▾

Convert to Carbon

Delete Atom

Convert Pseudo Atom to Group

**Abbreviated Group Checker**

Select a fix action upon a problem

Manual fix

Do not fix

Fix

Ungroup ▾

Ungroup

Contract Group

Expand Group

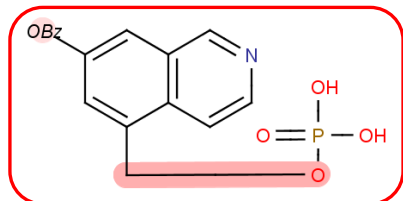
Configure checker

Detect Expanded Groups

Detect Contracted Groups

Detect All Groups

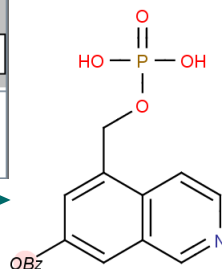
# Series of checkers and fixers: example



**Bond Length Checker**  
1 bond found with wrong length

Clean

**Pseudo Atom Checker**  
1 Pseudo atom found

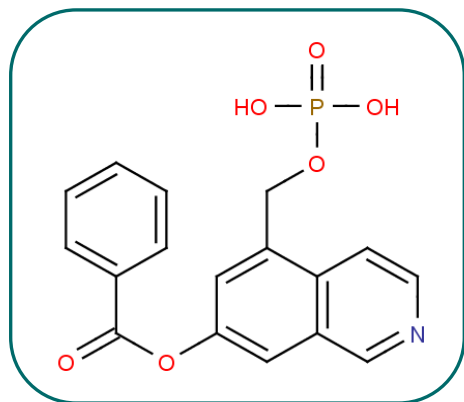
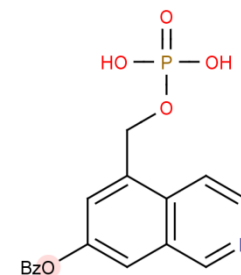


**Pseudo Atom Checker**  
1 Pseudo atom found

Convert to Carbon

Delete Atom

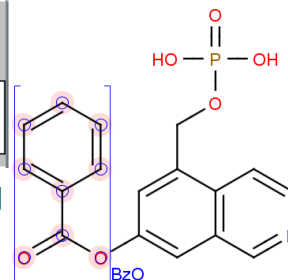
Convert Pseudo Atom to Group



**Abbreviated Group Checker**  
1 abbreviated group found

Ungroup

Contract Group



**Abbreviated Group Checker**  
1 abbreviated group found

Ungroup

Expand Group

# Reporting or fixing

## Set options

Choose a run method determining how to handle the identified issues, and select a report generation option.

### Operation Mode

- Check**  
Validates the structures without fixing the detected issues.
- Manual**  
Ignores all configured fixers. Users are always prompted to resolve any issues detected.
- Automatic**  
Fixes the structures with the selected fixers if possible. Users are never prompted.
- Fix**  
Solve problems with the fixer configured for each checker. Users are prompted to resolve any unfixable or conflicting problems.

### Report Options

- No Report**  
No report is created.
- File Report**  
Report is saved to a text file.
- Output Report**  
Report is saved to the output as property field. Note: Not all file format supports this option.

# Output options

- Optionally separate output files for accepted and discarded structures
- “Fail safe” mode
- Automatic discarding of OCR errors

## Specify output

Specify the location and name of the output files. You can save all structures in a single file, or you can separate the accepted and discarded ones in two files.

**Single Output**

Both fixed and unfixed structures will be saved to a single output file.

**Separated Output**

Fixed and Unfixed structures will be saved to separate output files.

Accepted	<input type="text"/>	<input type="button" value="X"/>	<input type="button" value="Browse..."/>
Discarded	<input type="text"/>	<input type="button" value="X"/>	<input type="button" value="Browse..."/>

Ignore errors and continue with next structure

Discard OCR errors

# Command-line Usage

---

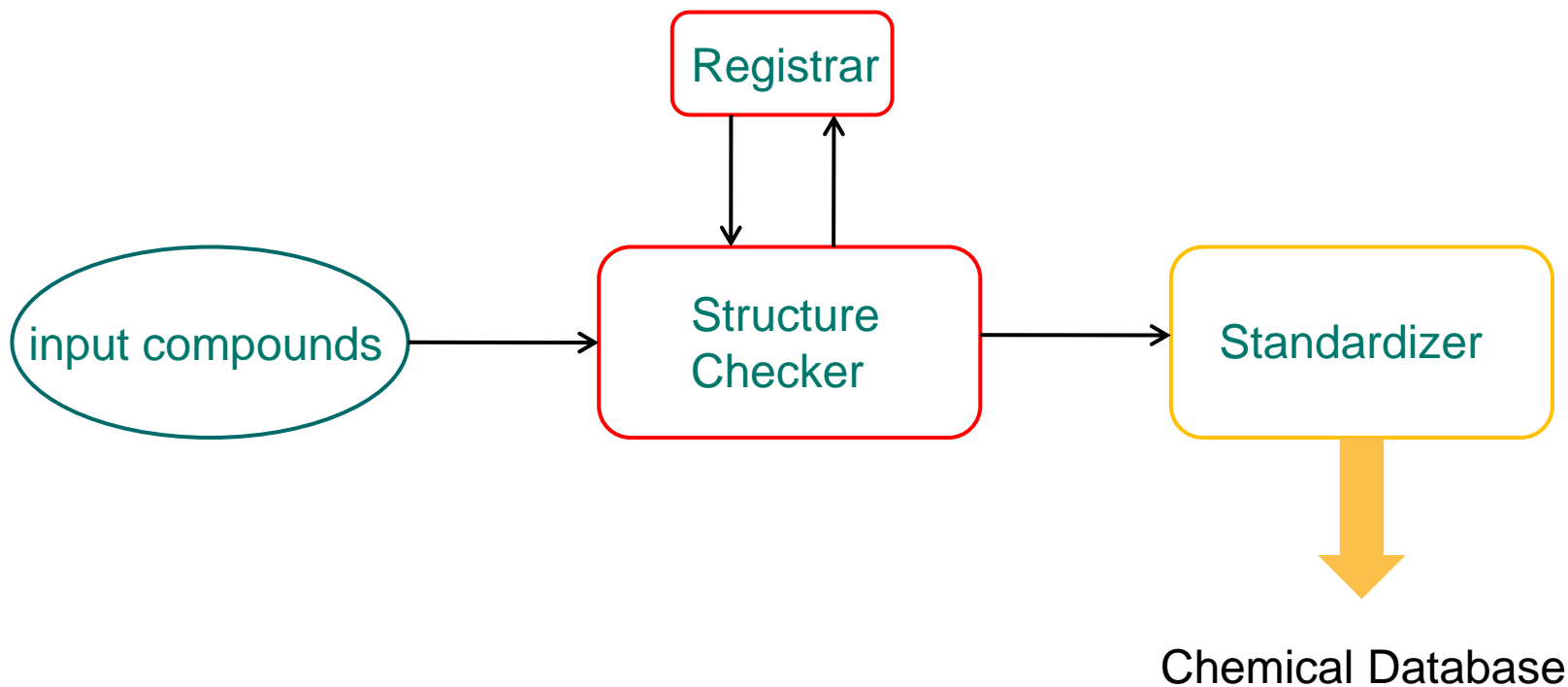
```
structurecheck -c "aromaticity..valence" -m fix -f sdf  
-o out.sdf in.sdf
```

```
-c      checkers, separated by ".."  
-m      [fix|check]  
-t      [single|separate|accepted|discarded]  
-o      output path  
-d      discarded path  
-f      format  
-ocr    discard molecules from OCR error  
-rp     write report to the property of output  
-l      write software error messages to log files
```

# Registration system

---

- Key component for registration systems
- Combined with Standardizer



# Structure Checker in KNIME (and Pipeline Pilot)

The image shows the KNIME software interface with a workflow titled "Multi-step synthesis". The workflow consists of two nodes: "MollImporter" (labeled "Structures with errors") and "Structure Checker" (labeled "Node 1"). A dialog box titled "Dialog - 0:1 - Structure Checker" is open, showing the configuration options for the Structure Checker node. The dialog has three tabs: "Structure Checker Options", "Flow Variables", and "Memory Policy". The "Structure Checker Options" tab is active, displaying a list of structure checking options:

- Abbreviated Group Checker
- Alias Checker
- Aromaticity Error Checker
- Atom Map Checker
- Atom Value Checker
- Attached Data Checker
- Bond Angle Checker

Below the list is a section titled "Build Your Configuration" with the following text:

Select a checker from the list on the left, and press 'Add >' button to append it to the checking queue on the right. You can change the order of checkers with the up and down arrow buttons on the right side. The current configuration can be saved for future use.

Some checkers have options that can be set if you select the corresponding checkers in the checking queue on the right. You can also select what action to perform in case of an issue. Automatic fixers are provided for many cases, but manual modification is available as an option as well.

At the bottom of the dialog, there is a "Structure column" dropdown menu set to "Molecule". The dialog has "OK", "Apply", and "Cancel" buttons.

# Future plans

---

- New checkers:
  - Reaction Checker
  - ~~Unbalanced Reaction Checker~~
  - ~~Valence Property Checker~~
  - R-group Checker
  - R-group Error Checker
  - Invalid R-group Checker
  - Polymer Checker
- Improve Structure Checker command line application
- Customizable fixers for Substructure Checker
- Welcome any suggestion for new checks

# Thank you

---



György Pirok



Zsolt Mohácsi



Attila Szabó



Imre Barna



István Rábel