

Chemical Noise Handling in Patent Chemistry Mining

Review of ChemAxon's Structure
Checker & Standardizer

Chemaxon EUGM 2012

Andrew Hinton Ph.D.

SureChemOpen

SureChem Data Stats

Patents

- 20 million annotated US,EP, PCT full text documents and Japio abstracts
- 70 million DOCDB records

MEDLINE

- 19 million abstracts

Chemistry

- 12 million unique chemical structures



Science is better when data is opened up
Welcome to **SureChemOpen**

SureChemPro

Web-based app with all the tools and functionality for professional patent chemistry search

SureChemDirect

Enterprise API and Data Feed solutions for easy integration of structure and patent data into internal workflows and applications

Structure Search within Patents

Enter your SureQuery™

Example: blah blah blah blah Fielded Search

SEARCH FOR KEYWORD(S)...

"Pi3k inhibitors"

Example: cancer, erectile dysfunction, cancer AND face

... IN DOC SECTIONS

- All
- Title
- Abstract
- Claims
- Description

BIBLIOGRAPHIC FIELDS ⓘ

Assignee(s)/Applicant(s) MERCK SERONO SA AND

IPCR C07D AND

Select bibliographic field... AND

Select bibliographic field... AND

Select bibliographic field...

PATENT AUTHORITIES ⓘ

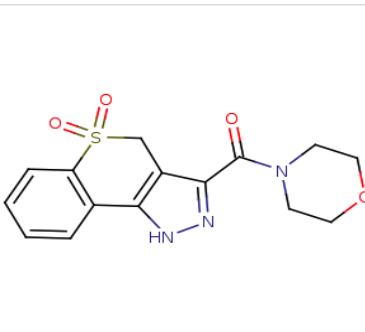
- All (inc. DocDB)
- US Applications
- US Granted
- EP Applications
- EP Granted
- WO

PUBLICATION DATE

Example: YYYYMMDD; YYYY; [YYYYMMDD TO YYYYMMDD]; [YYYY to YYYY]; [YYYY to *]

Search

Input structure



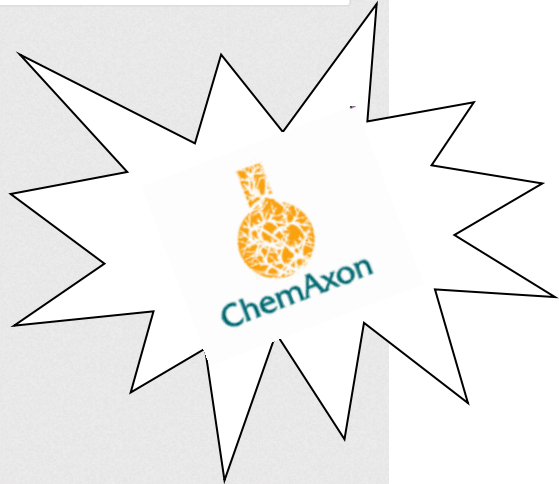
SELECT STRUCTURE SEARCH

- Substructure
- Duplicate
- Exact
- Similarity

SEARCH FOR STRUCTURE IN DOC SECTION(S)

- All
- Title
- Abstract
- Claims
- Description

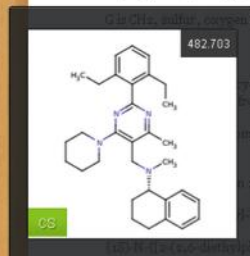
Structure searches in document sections are coming soon



Chemistry from Text & Images

(ii) phenyl and pyridyl, each of which is substituted with from 0 to 4 substituents independently chosen from halogen, hydroxy, amino, cyano, C₁-C₄alkyl, C₁-C₄alkoxy, (C₃-C₇cycloalkyl)C₀-C₄alkyl, C₁-C₂haloalkyl, C₁-C₂haloalkoxy and mono- and di-(C₁-C₄alkyl)amino; and

for NRE; wherein RE is:



(alkyl)C₀-C₄alkyl, phenyl or a 5- or 6-membered heteroaryl ring, each of which is substituted with from 0 to 3 substituents from R_x.

position comprising at least one compound or salt according to claim 1, in combination with a physiologically acceptable

chosen from:

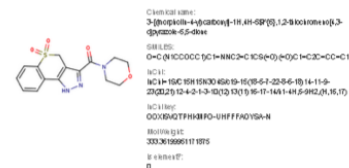
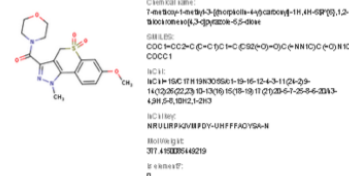
- (2,6-diethylphenyl)-6-methylpyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (2,6-diethylphenyl)-4-methyl-6-morpholin-4-ylpyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([2-(2,6-diethylphenyl)-4-methyl-6-piperidin-1-yl]pyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([2-(2,6-diethylphenyl)-4-methyl-6-pyrrolidin-1-yl]pyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([4-azepan-1-yl-2-(2,6-diethylphenyl)-6-methylpyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([2-(2,6-diethylphenyl)-4-methyl-6-thiomorpholin-4-yl]pyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([2-(2,6-diethylphenyl)-4-methyl-6-(4-methylpiperidin-1-yl)pyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- 1-[2-(2,6-diethylphenyl)-6-methyl-5-((methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino)methyl)pyrimidin-4-yl]piperidin-4-ol;
- (1S)-N-([2-(2,6-diethylphenyl)-4-(3,3-dimethylpiperidin-1-yl)-6-methylpyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([2-(2,6-diethylphenyl)-4-(3,5-dimethylpiperidin-1-yl)-6-methylpyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1S)-N-([2-(2,6-diethylphenyl)-4-methyl-6-pyridin-4-yl]pyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- ethyl 1-[2-(2,6-diethylphenyl)-6-methyl-5-((methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino)methyl)pyrimidin-4-yl]piperidine-4-carboxylate;
- {1-[2-(2,6-diethylphenyl)-6-methyl-5-((methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino)methyl)pyrimidin-4-yl]piperidin-4-yl}methanol;
- 2-[1-[2-(2,6-diethylphenyl)-6-methyl-5-((methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino)methyl)pyrimidin-4-yl]piperidin-4-yl]ethanol;
- 1-[2-(2,6-diethylphenyl)-6-methyl-5-((methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino)methyl)pyrimidin-4-yl]piperidine-4-carboxylic acid;
- 1-[2-(2,6-diethylphenyl)-6-methyl-5-((methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino)methyl)pyrimidin-4-yl]-D-proline;
- (1R)-N-([2-(2,6-diethylphenyl)-4-(1H-imidazol-1-yl)-6-methylpyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- (1R)-N-([2-(2,6-diethylphenyl)-4-methyl-6-(1H-pyrazol-1-yl)pyrimidin-5-yl)methyl)-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;
- 1-([2-(2,6-diethylphenyl)-4-methyl-6-piperidin-1-yl]pyrimidin-5-yl)methyl)piperidin-4-ol;

* 1, 2, 3, 4, and 5, 10, 12, and 13 are CH₃.

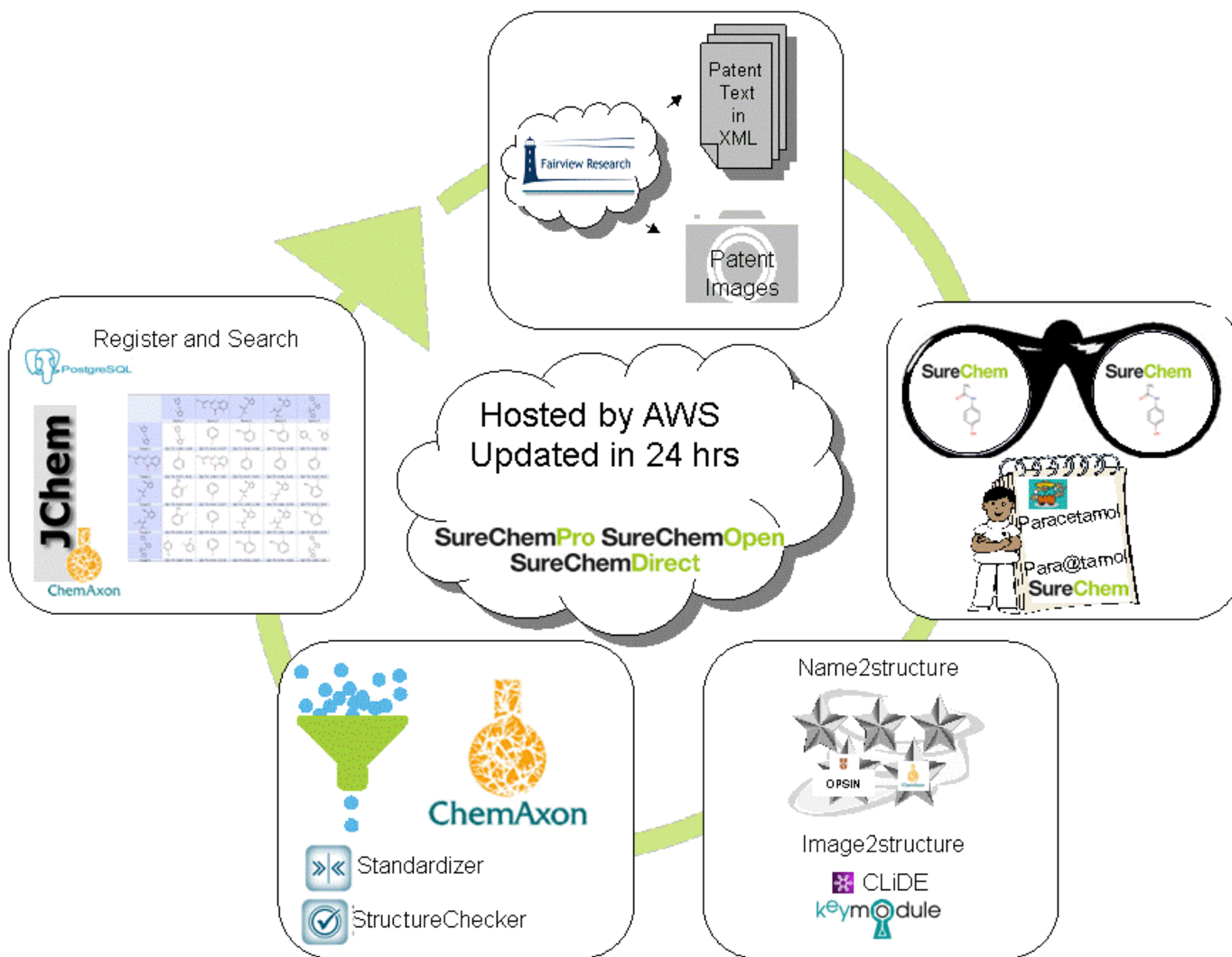
More preferably, the invention relates to compounds of Formula (I), (I*) and related Formulas, selected from the following group. In the table below, in case the structure contains one or more stereogenic centers, the respective structure is depicted in a arbitrary absolute configuration. These structures also include the respective enantiomers having the opposite stereochemistry and the corresponding enantiomers.

Example No	structures	Example No	structure
1		2	
3		4	

STRUCTURE(S) EXTRACTED FROM THIS IMAGE:



Automated Chemical Mining of Patents



Chemical Noise

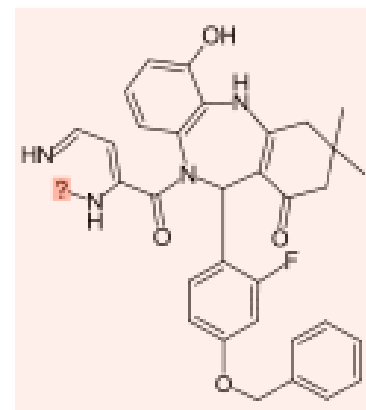
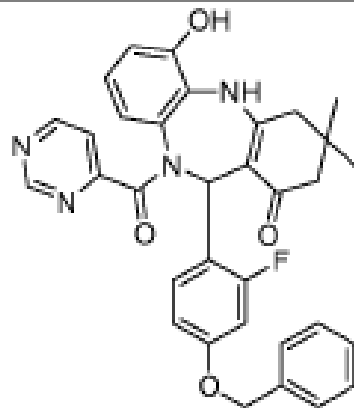


Gabriel Sko

Chemical Noise, Patents & Chemistry Mining

- **Annoying variation**
 - Abbreviations
 - Charge group
 - Stereo and tautomeric isoforms
- **Unnecessary Chemistry**
 - Common simple element
 - General ambiguous groups
- **Nuisance Fragments (N2S and I2S)**
 - Markush elements with no context
 - Partial name recognition due to whitespace
 - Truncation generating unlikely chemical species
 - Non-chemical artefacts from I-2-S process

2-Chloro-4-morpholinothienc [3,2-d]pyrimidin-6-yl)-N-methyl,N- methanesulfonylmethanamine (General Procedure C-2,0.22 mmol) was reacted using



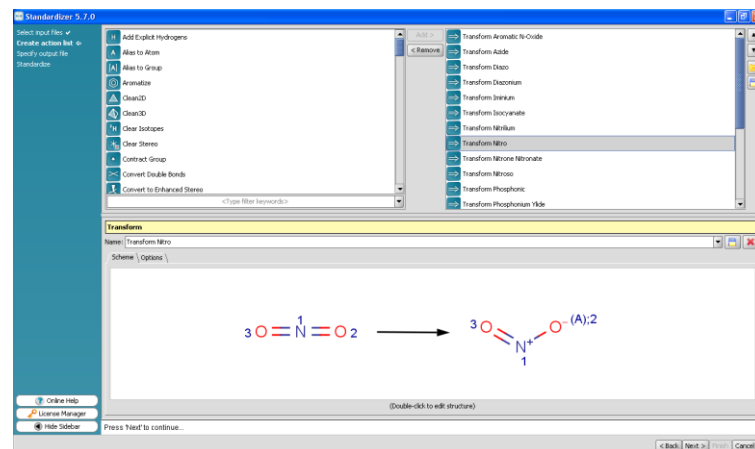
Robust Structure Handler Required



Structure Checker & Standardizer

- Features

- Transform using SMIRKS
- Detect empty or erroneous atom types
- Normalize aromaticity, charge & hydrogens



Robust Structure Filtration and Normalization Workflow

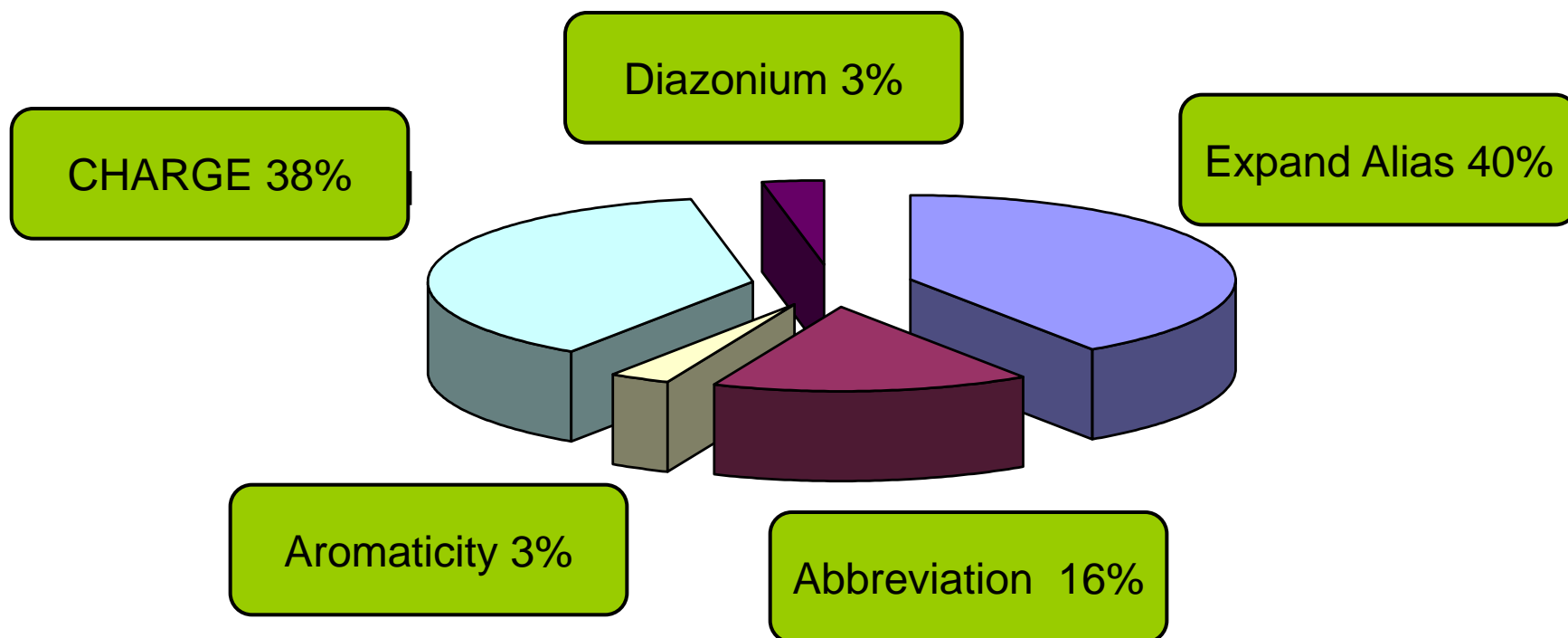
- Used GUI in development
- Production implementation uses Java
- Utilized at several points
 - Image-2-Structure
 - Name-2-Structure
 - Storage and Searching
- Filter and transform
- Most “noise” from Image-2-Structure is removed
- Fewer duplicate structures stored once changed into canonical form



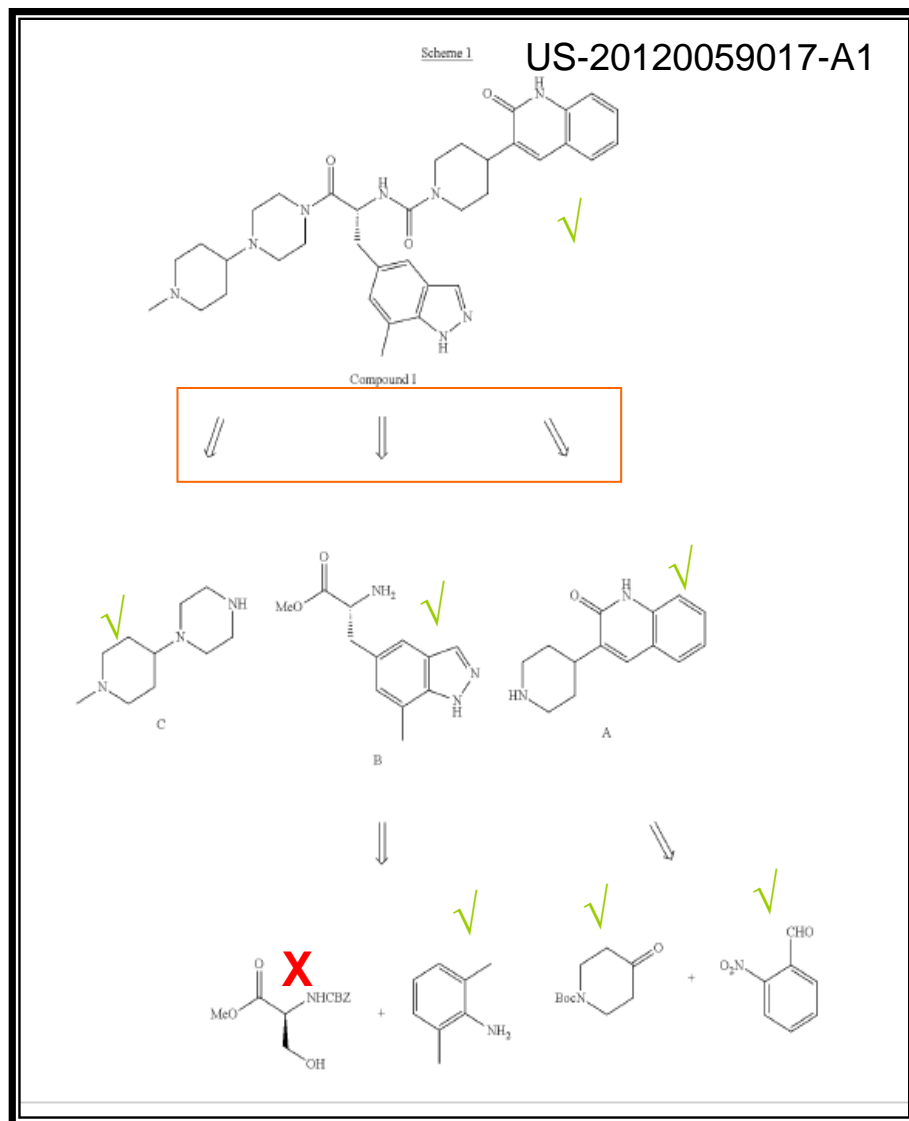
	Empty Structure Checker
	Star Atom Checker
	Valence Error Checker
	Abbreviated Group Checker
	Alias Checker
	Aromaticity Error Checker
	Coordination System Error Checker
	Covalent Counterion Checker
	Metallocene Error Checker
	Molecule Charge Checker
	Explicit Hydrogen Checker

	Transform Aromatic N-Oxide
	Transform Azide
	Transform Diazo
	Transform Diazonium
	Transform Iminium
	Transform Isocyanate
	Transform Nitrilium
	Transform Nitro
	Transform Nitrene Nitronate
	Transform Nitroso
	Transform Phosphonic
	Transform Phosphonium Ylide
	Transform Selenite
	Transform Silicate
	Transform Sulfine
	Transform Sulfon
	Transform Sulfonium Ylide
	Transform Sulfoxide
	Transform Sulfoxonium Ylide
	Transform Tertiary N-Oxide
	Wedge Clean

Top Transformations



An Observable Improvement



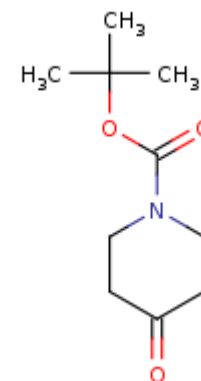
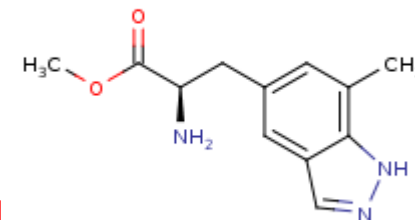
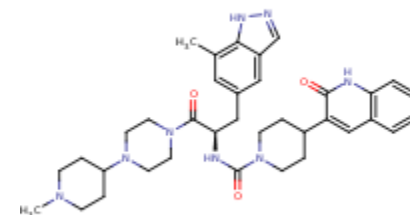
Whole structures retained

Artefacts removed

Wedge bonds cleaned

Structures with irresolvable issues removed

Protecting groups expanded



Perspectives

- Positives
 - GUI easy to use
 - Straightforward to implement in Java API
 - Offers most of the functionalities we need
 - Examples of best-practice available (though limited)
- Negatives
 - Error handling
 - Tautomerization not suitable for unguided automated use
 - Documentation and features overlap between products

Create a FREE account at www.surechem.com

Elisabeth Piveteau
e.piveteau@digital-science.com
+44 7801 133 928