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

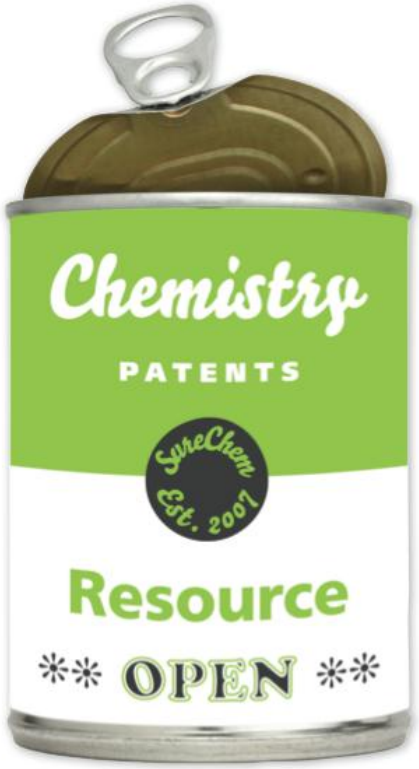
- 20 million annotated US,EP, PCT full text documents and Japio abstracts
- 70 million DOCDB records

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

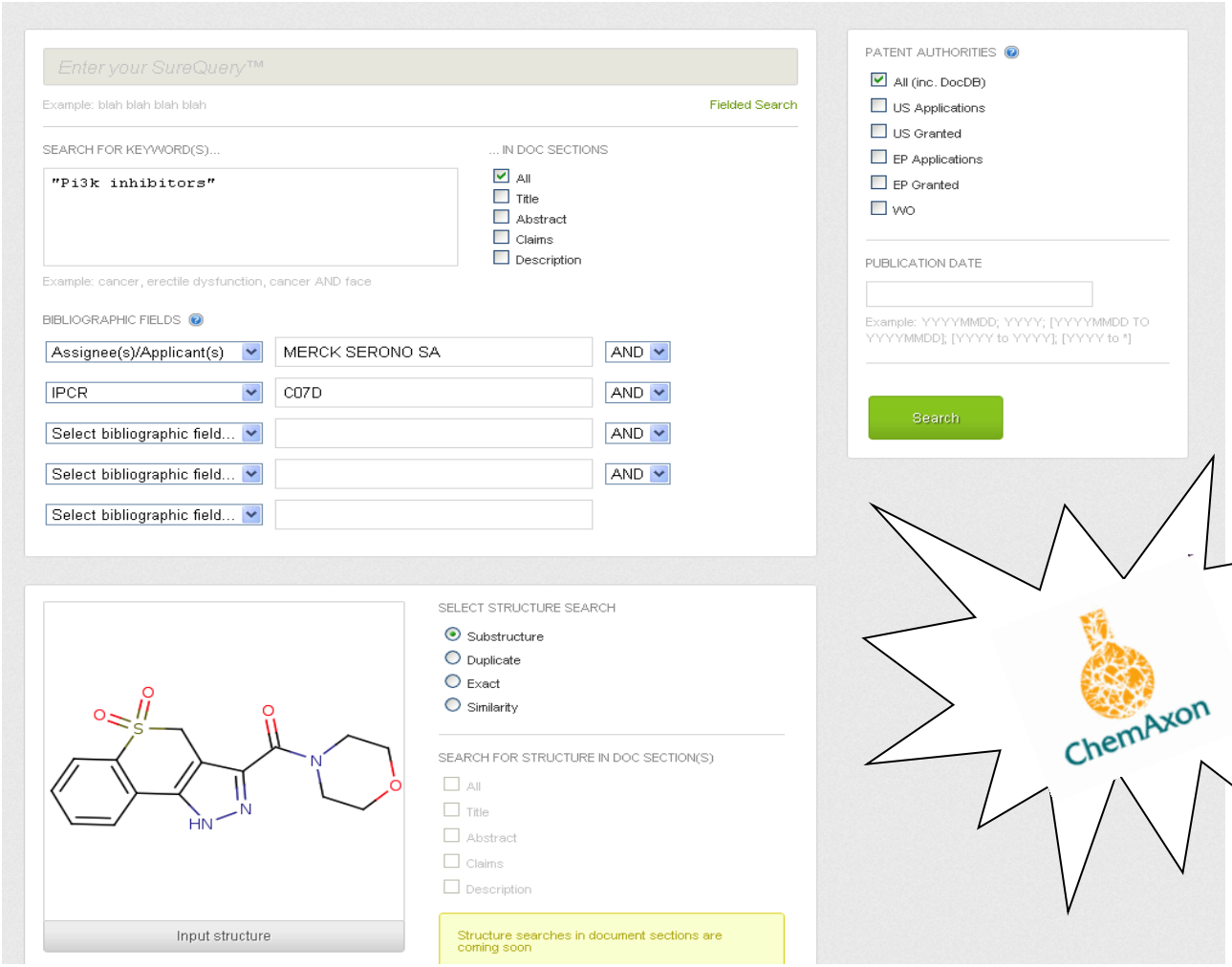
SureChemDirect

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

DIGITAL science

www.surechem.com

Structure Search within Patents



DIGITAL science

www.surechem.com

SureChem

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)Co-C4alkyl, C1-C2haloalkyl, C1-C2haloalkoxy and mono- and di-(C1-C4alkyl)amino; and

r NRE; wherein RE is:

loalkyl)Co-C4alkyl, phenyl or a 5- or 6-membered heteroaryl ring, each of which is substituted with from 0 to 3 substituents m Rx.

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;

enyl)-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-ylpyrimidin-5-yl]methyl}-N-methyl-1,2,3,4-tetrahydronaphthalen-1-amine;

(1S)-N-{[2-(2,6-diethylphenyl)-4-methyl-6-pyrrolidin-1-ylpyrimidin-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-ylpyrimidin-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-ylpyrimidin-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;

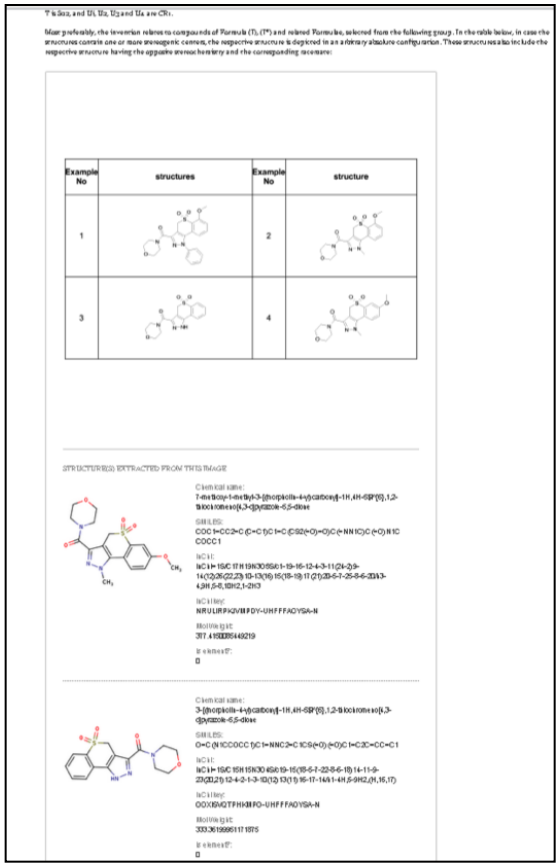
1-[2-(2,6-diethylphenyl)-6-methyl-5-({methyl[(1S)-1,2,3,4-tetrahydronaphthalen-1-yl]amino}methyl)pyrimidin-4-yl]piperidine-4-carboxylic acid;

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;

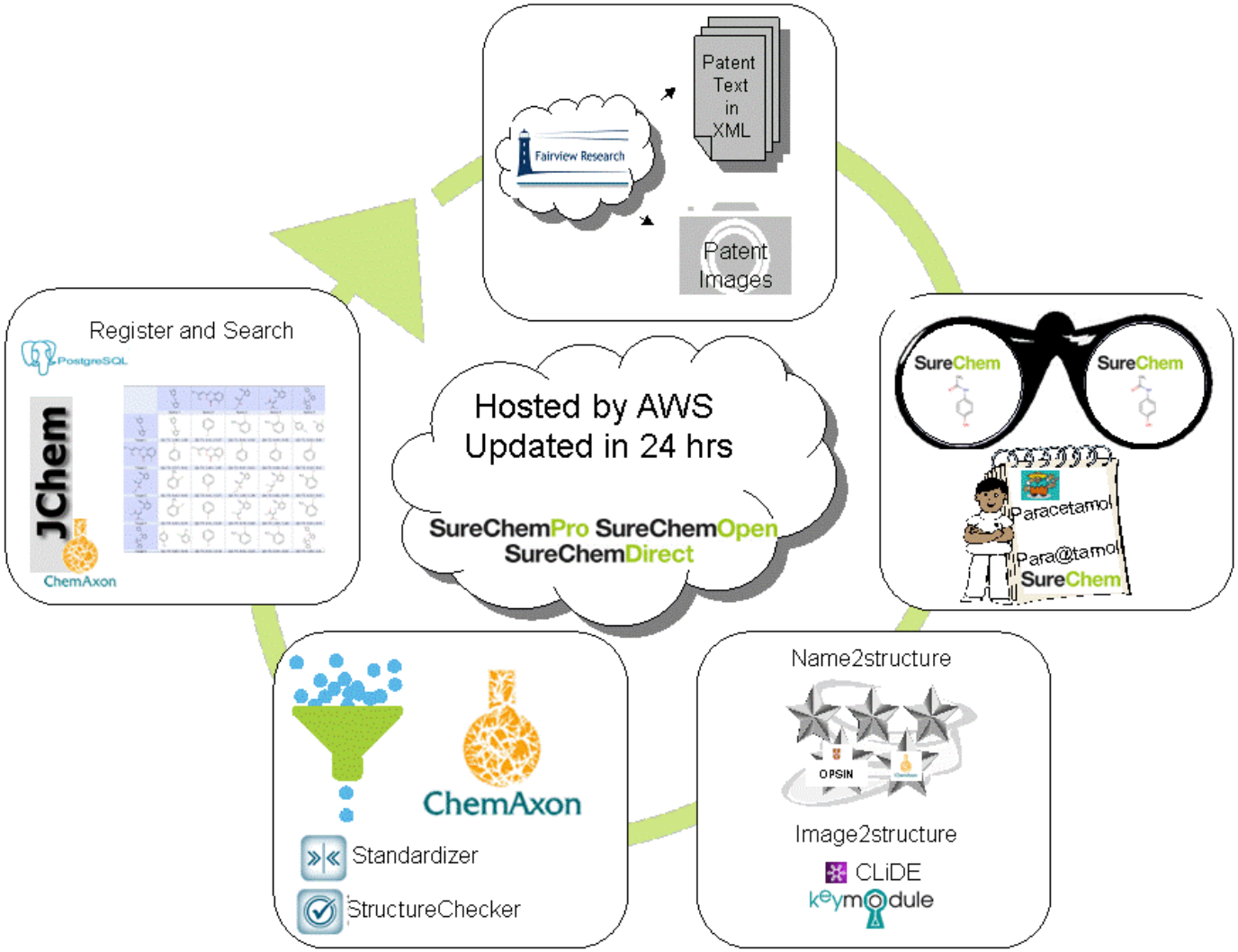
1-{[2-(2,6-diethylphenyl)-4-methyl-6-piperidin-1-ylpyrimidin-5-yl]methyl}piperidin-4-ol;



Example No	structures	Example No	structure
1		2	
3		4	

STRUCTURE(S) EXTRACTED FROM THIS IMAGE

Automated Chemical Mining of Patents



DIGITAL science

www.surechem.com

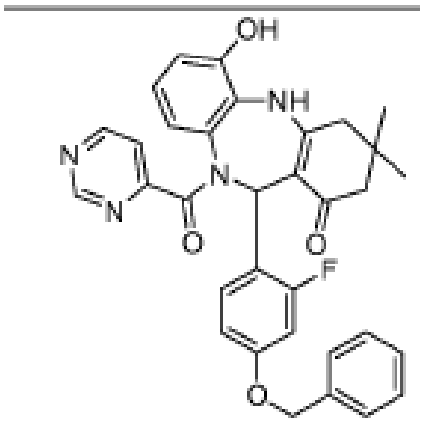
SureChem

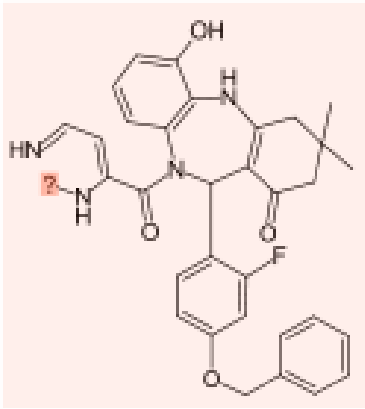
Chemical Noise

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

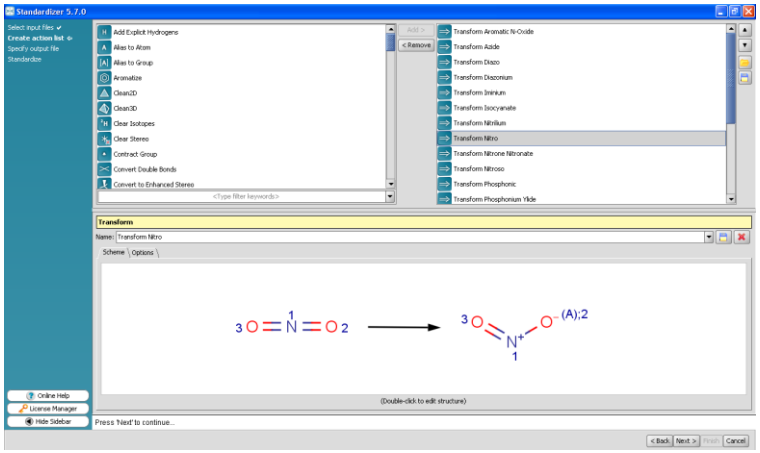




Robust Structure Handler Required

Structure Checker & Standardizer

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



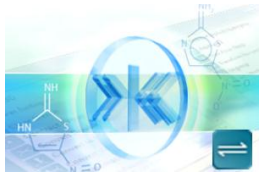
DIGITAL science

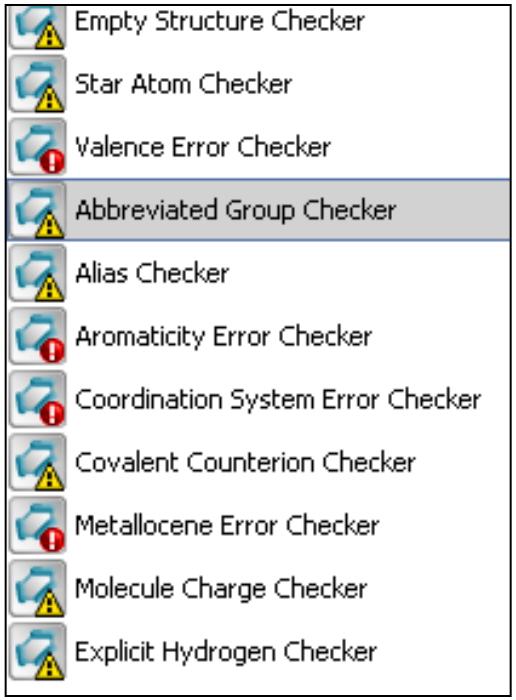
www.surechem.com

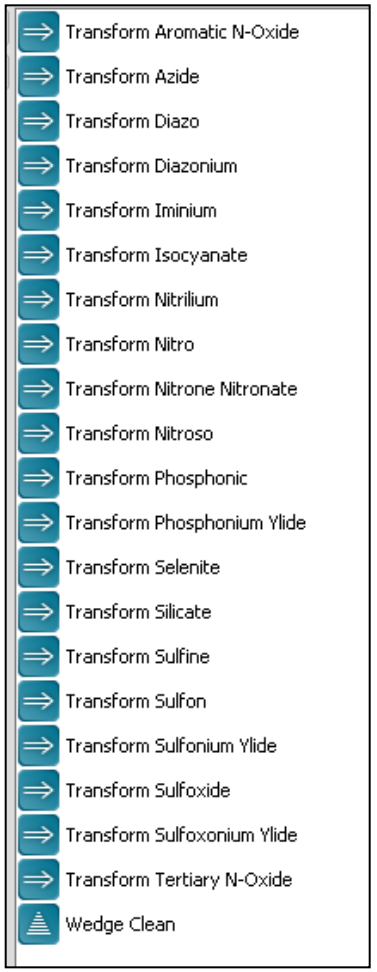
SureChem

Robust Structure Filtration and Normalization Workflow

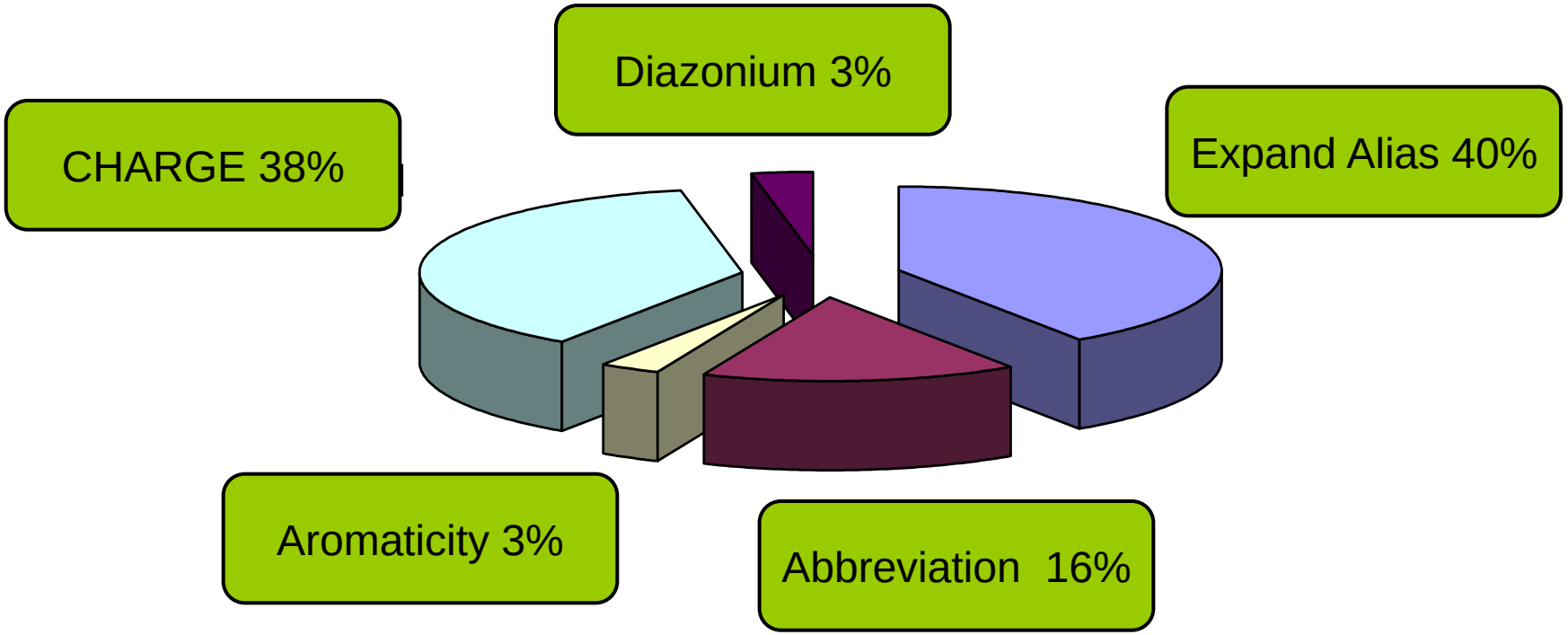
- Used GUI in development
- Production implementation uses Java
- Utilized at several points
  - Image-2-Structure
  - Name-2-Structure
  - Storage and Searching
- Filter and transform
- Most “noise” from Image-2-Structure is removed
- Fewer duplicate structures stored once changed into canonical form



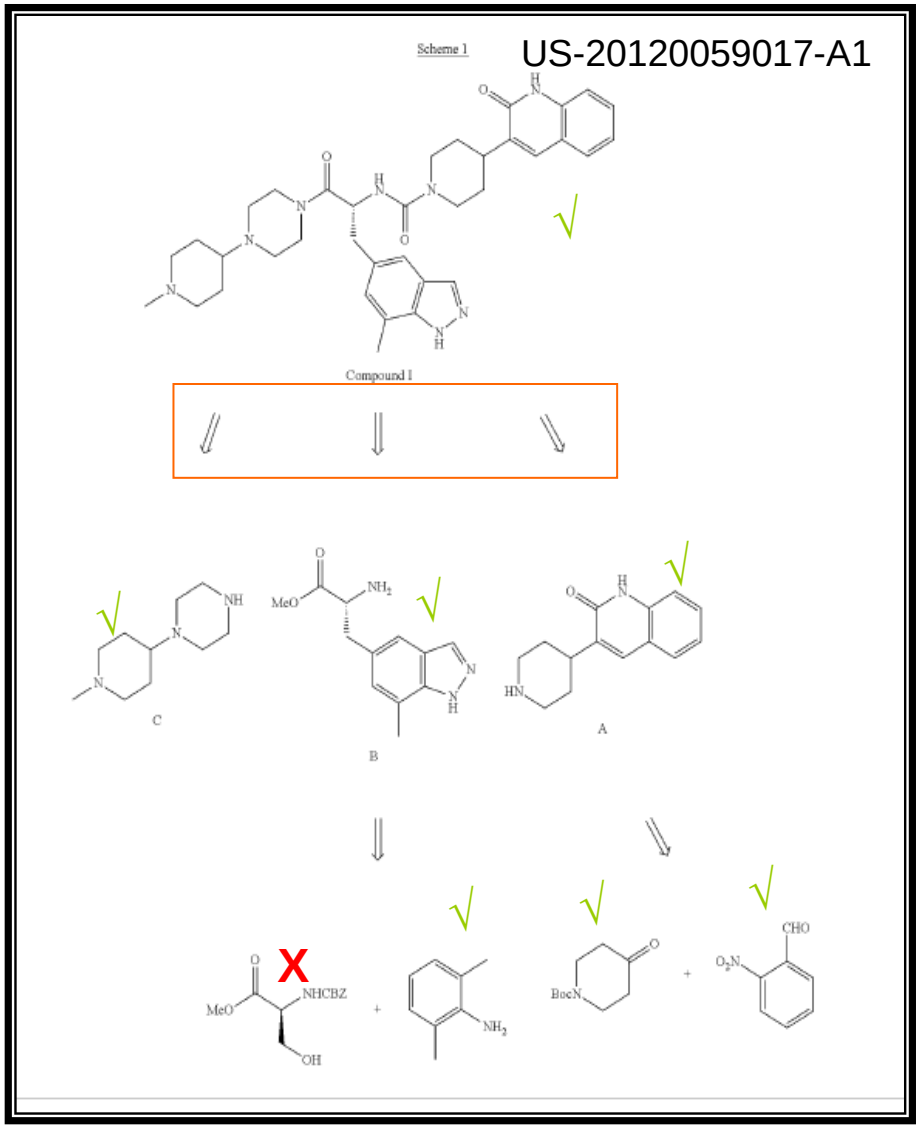




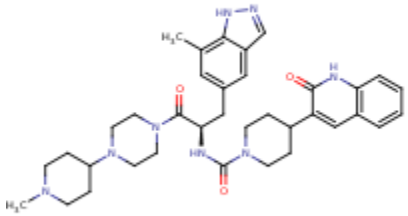
Top Transformations



An Observable Improvement

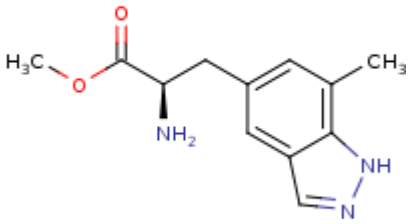


Whole structures retained

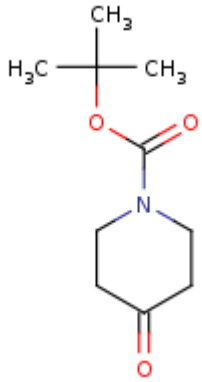


Artefacts removed

Wedge bonds cleaned



Structures with irresolvable issues removed



Protecting groups expanded

DIGITAL science

www.surechem.com

SureChem

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