Identification of chemical structures in literature sources using semantic analysis and automatic structure generation

Dr William G Town

bill.town@kilmorie.com

and

Dr Jeffery Press

jpress@cambridgesoft.com

ICIC Nimes 2006

CambridgeSoft

Overview

- Personal view of the history of text processing
- Context – why is text processing important?
- Why is chemistry a problem?
- How the problem is solved
- Some examples of the name to chemical structure technology in action

Text processing – a personal history

- Chomsky (University of Sheffield - 1960s)
- Systran (European Commission - 1970s and 80s)
- Text and data mining (Derwent/IBM (Charles Huot) 1995-)
- Syntactic and semantic analysis (Liz Liddie 1995-)
- Funding boost from security and defence agencies' interest in automatic text analysis
- Internet search engines

A journey through 40 years of text processing history

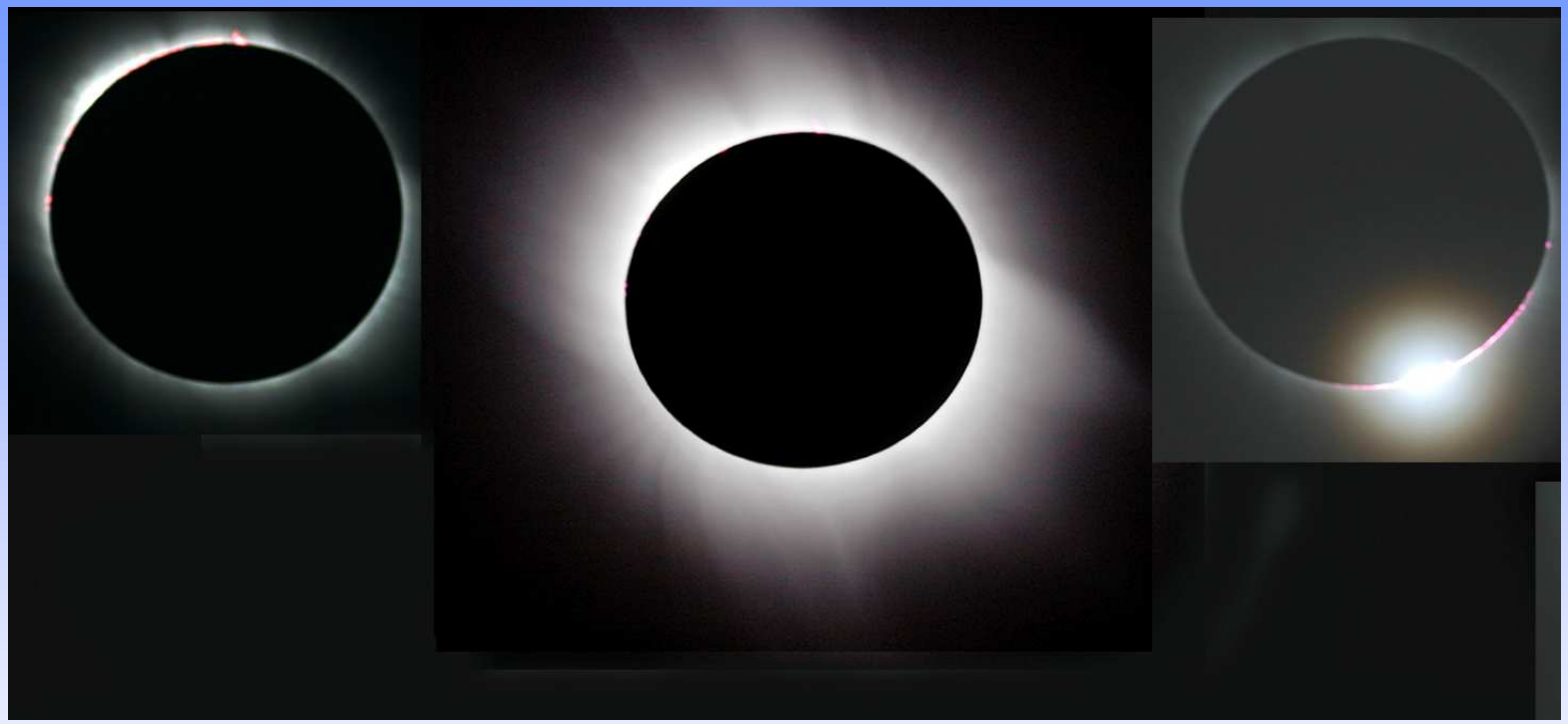
1960s – setting out with hope and enthusiasm

1970s – on the road with first successes

1970s – the first set backs

1990s – on the road again with more funding/vehicles

2010 – Google eclipses the information world



Context – why is text processing important?

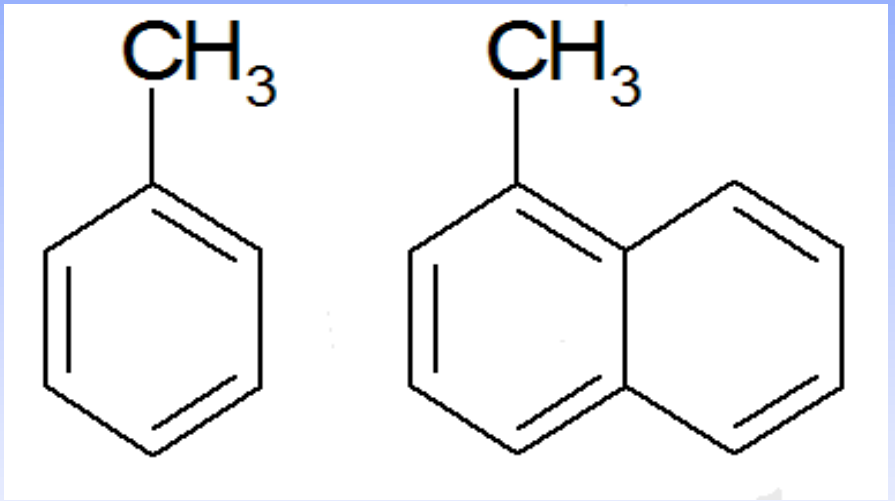
- >20 million machine readable journal articles
- >10m machine readable patents
- Private documents
- Corporate archives
- Electronic Laboratory Notebooks
- World Wide Web

Why is chemistry a problem?

- Chemical substances have a wide variety of trivial and official names.
- No simple text search can find chemical substances which are named using one of the alternative names.
- Synonym expansion is insufficient.
- Searching by structure solves the problem in such cases.

Substructure search

We would like to find compounds which are supersets of the given structure (e.g. toluene and methylnaphthalene)



Text searches won't find documents with similar structures

Objective

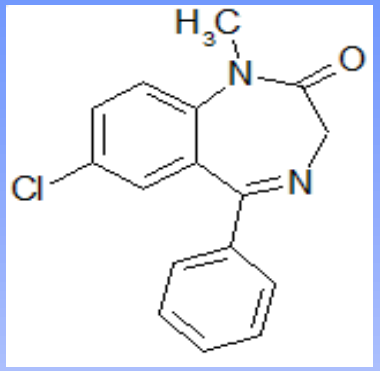
Enable - structure / substructure searches of chemical substances mentioned in documents. (e.g. in the chemical literature)

where the chemical structures were not created with a chemical structure editor, and are only mentioned as “text”, (for example. via their common name or IUPAC names).

i.e. there may be no ‘connection table’ associated with those structures or they are there as bit maps only.

Example – I need to find information about Valium

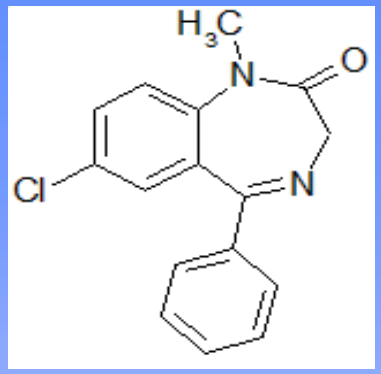
Valium (Trade Name) = Diazepam (Generic Name) = CAS # 439-14-5 (Chemical ID #) =



Valium has > 149 “names”

ALBORAL, ALISEUM, ALUPRAM , AMIPROL ,ANSIOLIN , ANSIOLISINA , APAURIN, APOZEPAM, ASSIVAL , ATENSINE , ATILEN , BIALZEPAM , CALMOCITENE, CALMPOSE , CERCINE, CEREGULART, CONDITION, DAP, DIACEPAN, DIAPAM , DIAZEMULS , DIAZEPAN , DIAZETARD , DIENPAX, DIPAM , DIPEZONA, DOMALIUM , DUKSEN, DUXEN, E-PAM, ERIDAN, EVACALM, FAUSTAN, FREUDAL , FRUSTAN, GIHITAN, HORIZON, KIATRIUM, LA-III , LEMBROL, LEVIUM, LIBERETAS , METHYL DIAZEPINONE, MOROSAN , NEUROLYTRIL NOAN NSC-77518 PACITRAN PARANTEN PAXATE PAXEL PLIDAN QUETINIL QUIATRIL QUIEVITA RELAMINAL RELANIUM RELAX RENBORIN RO 5-2807 S.A. R.L. SAROMET SEDAPAM SEDIPAM SEDUKSEN SEDUXEN , SERENACK SERENAMIN SERENZIN SETONIL SIBAZON SONACON STESOLID STESOLIN , TENSOPAM TRANIMUL TRANQDYN TRANQUASE TRANQUIRIT , TRANQUO-TABLINEN , UMBRIUM UNISEDIL USEMPAX AP VALEO VALITRAN VALRELEASE VATRAN VELIUM, VIVAL VIVOL WY-3467

There are many different systematic chemical names for Valium



Valium = Diazepam = CAS # 439-14-5 =

7-chloro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one

7-chloro-1-methyl-5-phenyl-3H-1,4-benzodiazepin-2(1H)-one

7-chloro-1-methyl-5-phenyl-1,3-dihydro-2H-1,4-benzodiazepin-2-one

7-chloro-1-methyl-2-oxo-5-phenyl-3H-1,4-benzodiazepine

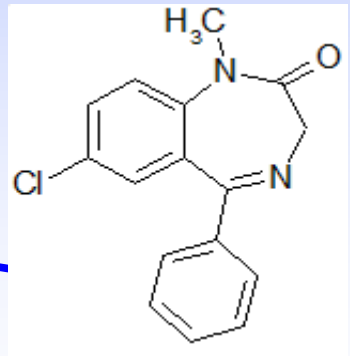
1-methyl-5-phenyl-7-chloro-1,3-dihydro-2H-1,4-benzodiazepin-2-one

7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one

7-chloro-1-methyl-5-3H-1,4-benziodiazepin-2(1H)-one

An example of different chemical names for the same molecule: however, they share common chemical fragments!

1-METHYL-5-PHENYL-7-CHLORO-1,3-DIHYDRO-2H-1,4-BENZODIAZEPIN-2-ONE

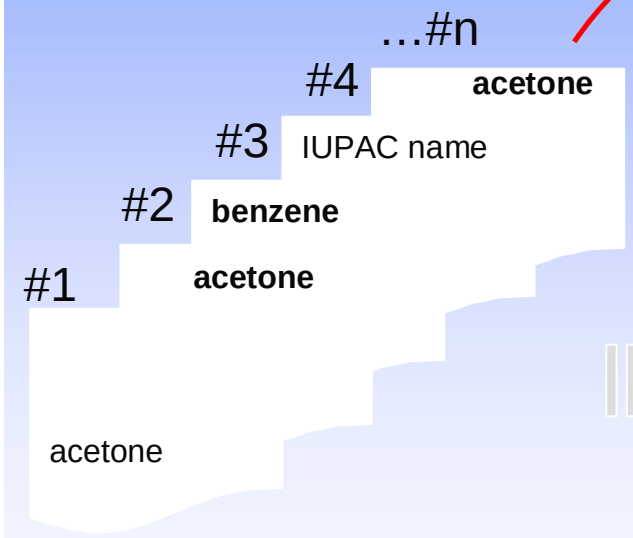


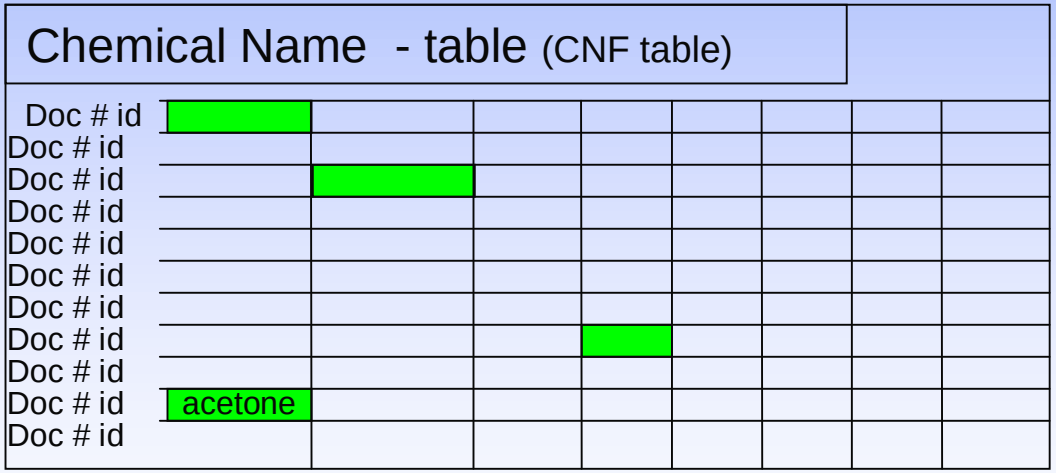
7-CHLORO-1-METHYL-5-PHENYL-1,3-DIHYDRO-2H-1,4-BENZODIAZEPIN-2-ONE

Chemical names are identified & extracted into a chemical name table

Identify & Extract Chemical names & chemical name fragments (CNF's)

Create CNF table
for indexing & analysis



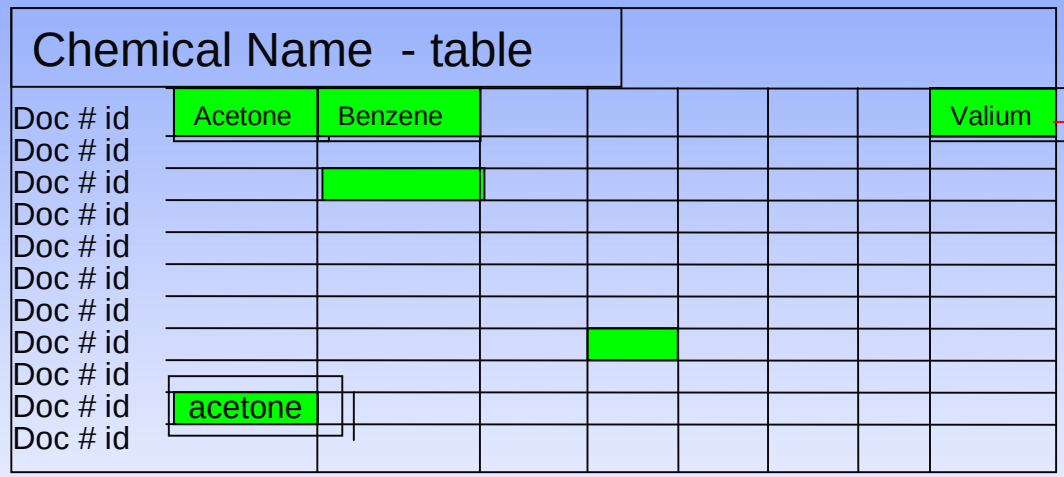


Documents
(containing words, references, etc.)

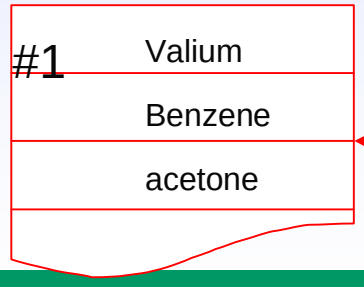
▪ ▪ ▪ **Every chemical name in every document**

Step 3

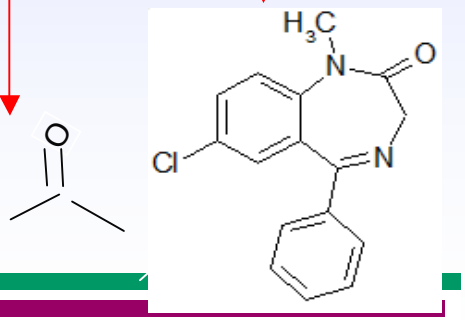
Process chemical names (& CNF's) into → chemical structures via the [name = struct] program

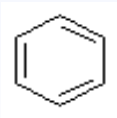


Name = Struct Program



Associate Structure with original documents

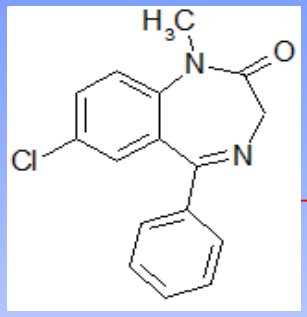




CambridgeSoft

Result

Draw This



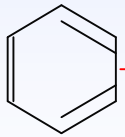
Retrieve

Valium

Diazepam

Retrieve

acetone



Retrieve

#2 benzene

Automated chemical entity extraction

- TEMIS & Elsevier/MDL
  - Chemical Entity Relationship Skill Cartridge
- IBM
  - Business Information Workbench
- ReelTwo
  - SureChem
- CambridgeSoft
  - *ChemFinder/Text (Document Manager)*

Example

Step 1 Identify the chemical entities

United States Patent

Shindo et al.

[15] **3,692,776**

[45] **Sept. 19, 1972**

[54] **PROCESS FOR PREPARING 7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE**

[72] Inventors: **Minoru Shindo**, Tokyo; **Kanji Moro**, Tokyo; **Teizo Shinozaki**, Chiba-ken, all of Japan

[73] Assignee: **Chugai Seiyaku**, Kabushiki Kaisha, Tokyo, Japan

[22] Filed: **Oct. 27, 1970**

[21] Appl. No.: **84,549**

Related U.S. Application Data

[62] Division of Ser. No. 841,611, Oct. 27, 1970.

[52] **U.S. Cl.** .....**260/239.3 D, 260/562 N**

[51] **Int. Cl.** **C07d 53/06**

[58] **Field of Search**.....260/239.3 D

[56] **References Cited**

UNITED STATES PATENTS

3,371,085 2/1968 Reeder et al.....260/239.3 D

Primary Examiner—Henry R. Jiles

Assistant Examiner—Robert T. Bond

Attorney—Otto John Munz

[57] **ABSTRACT**

The known 7-chloro-1,3-dihydro-l-methyl-5-phenyl-2H-1, 4-benzodiazepin-2-one which was found to be pharmacologically effective in neuro-psychic disorders is prepared easily and in high yield by an improved process comprising the reaction of a novel intermediate, N-aminoacetyl-5-chloro-N-methylanthranilic acid, with phosphorus pentachloride followed by reacting with benzene in the presence of aluminum chloride, the intermediate which is also found to have valuable pharmacological activities being, in turn, prepared by the reaction of 5-chloro-N-methyl-N-phthalimidoacetylanthranilic acid with hydrazine.

15 Claims, No Drawings

CambridgeSoft

Example

Step 2 Extract the chemical names and load into tables

United States Patent

Shindo et al.

[15] **3,692,776**

[45] **Sept. 19, 1972**

[54] **PROCESS FOR PREPARING 7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE**

[72] Inventors: **Minoru Shindo**, Tokyo; **Kanji Moro**, Tokyo; **Teizo Shinozaki**, Chiba-ken, all of Japan

[73] Assignee: **Chugai Seiyaku**, Kabushiki Kaisha, Tokyo, Japan

[22] Filed: **Oct. 27, 1970**

[21] Appl. No.: **84,549**

Related U.S. Application Data

[62] Division of Ser. No. 841,611, Oct. 27, 1970.

[52] **U.S. Cl.** .....**260/239.3 D, 260/562 N**

[51] **Int. Cl.** **C07d 53/06**

[58] **Field of Search**.....260/239.3 D

[56] **References Cited**

UNITED STATES PATENTS

3,371,085 2/1968 Reeder et al.....260/239.3 D

Primary Examiner—Henry R. Jiles
Assistant Examiner—Robert T. Bond
Attorney—Otto John Munz

[57] **ABSTRACT**

The known 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1, 4-benzodiazepin-2-one which was found to be pharmacologically effective in neuro-psychic disorders is prepared easily and in high yield by an improved process comprising the reaction of a novel intermediate, N-aminoacetyl-5-chloro-N-methylanthranilic acid, with phosphorus pentachloride followed by reacting with benzene in the presence of aluminum chloride, the intermediate which is also found to have valuable pharmacological activities being, in turn, prepared by the reaction of 5-chloro-N-methyl-N-phthalimidoacetylanthranilic acid with hydrazine.

15 Claims, No Drawings

Chemical Entities Extracted from page

- 7-chloro-1.3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one
- 7-chloro-1.3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one
- N-aminoacetyl-5-chloro-N-methylanathranilic acid
- Phosphorus pentachloride
- benzene
- aluminum chloride
- 5-chloro-N-methyl-N-phthalimidoacetylanthranilic acid
- hydrazine

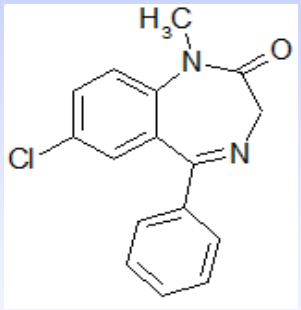
CambridgeSoft

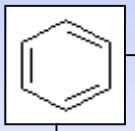
Step 3 Chemical names are processed into chemical structures (connection tables)

USP # 3,692,776

7-chloro-1.3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one	benzene	aluminum chloride	phosphorus pentachloride	hydrazine

Name = Structure Processing





SMILES:
c1ccccc1

Connection tables

```
  6  6  0  0  0  0  0  0  0  0  0999 V2000
    6.7092    5.6087    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.7076    4.5056    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    7.6607    3.9551    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    8.6160    4.5062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    8.6121    5.6136    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    7.6583    6.1591    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  2  3  1  0  0  0  0
  3  4  2  0  0  0  0
  4  5  1  0  0  0  0
  5  6  2  0  0  0  0
  6  1  1  0  0  0  0
M  END
```

CambridgeSoft

Example of text

General. All reactions were carried out under an argon atmosphere i

glassware. Elemental analyses were performed by E & R Microana

Inc., Parsippany, N.J. or by Atlantic Microlabs Inc, Norcross, GA. Toluene was

distilled under nitrogen from molten sodium. THF was distilled under argon from

sodium benzophenone ketyl. Unless stated otherwise, commercially obtained materials

Log of example analysis

- C:\CambridgeSoft\ChemFinderText\TagChemicalNames\in\example.html
- argon
- Toluene
- nitrogen
- sodium
- argon
- sodium
- benzophenone
- ketyl
- 2-bromobiphenyl
- Chlorodicyclohexylphosphine, palladium acetate
- n-butyllithium
- Di-tert-butylchlorophosphine
- potassium
- fluoride

Organic Syntheses *article*

B. *N,N-Dimethyl-N'-p-tolylformamidine*. A 250-mL, three-necked, round-bottomed flask equipped with a reflux condenser and a magnetic stirring bar coated with Teflon is placed under a positive nitrogen pressure and charged with 100 mL of methanol (Note 9). Sodium metal (1.4 g, 0.06 mol) (Note 10) is then added in small portions. After all of the sodium has reacted, *p*-toluidine (6.4 g, 0.06 mol) (Note 11) is added and the resulting solution is stirred for 5 min. The iminium salt (10.6 g, 0.065 mol) produced in Step A is added in one portion and the resulting mixture is refluxed with stirring overnight. The reaction mixture is cooled to room temperature and the solvent is removed on a rotary evaporator. The residue is taken up in chloroform (100 mL) and extracted twice with a saturated, aqueous solution of sodium bicarbonate (2 × 30 mL). The chloroform phase is dried over anhydrous magnesium sulfate and filtered, and the solvent is

Procedure A

Notes

1. Tschantz, M. A.; Burgess, L. E.; Meyers, A. I. Org. Synth., Coll. Vol. IX 1998, 530.
2. (S)-(+)-2-Phenylglycinol is commercially available, or may be prepared by reducing (S)-(+)-2-phenylglycine as follows: A 1-L, three-necked, round-bottomed flask is charged with 200 mL of dry THF under an argon atmosphere. Portionwise, 10.01 g of sodium borohydride (264.6 mmol) is added, followed by the dropwise addition of 65.1 mL of boron trifluoride etherate (528.7 mmol). The colorless suspension is stirred for 15 min, followed by portionwise addition of 20.00 g of (S)-(+)-2-phenylglycine (132.2 mmol). (CAUTION: exotherm and gas evolution!). The resulting suspension is heated at reflux for 12 hr and then is allowed to cool to room temperature followed by quenching with methanol until gas evolution ceases. The reaction mixture is concentrated under reduced pressure to yield a colorless solid that is taken up in 400 mL of 20% aqueous sodium hydroxide solution. The basic solution is extracted three times with 200-mL portions of dichloromethane, the combined organic layers are dried over sodium sulfate, and concentrated under reduced pressure to yield 14.72 g (81%) of (S)-(+)-2-phenylglycinol as a colorless solid: mp 72–74°C, [a]D +32.2°; 1H NMR (300 MHz, CDCl3) d: 3.52 (dd, 1 H, J = 10.7, 8.3), 3.71 (dd, 1 H, J = 10.7, 4.4), 4.01 (dd, 1 H, J = 8.2, 4.4), 7.23–7.35 (m, 5 H).
3. Merck 951 grade silica gel was used. Chromatography was performed in the manner described by Still2 using 20:1 (w/w) of silica gel to crude product.

Procedure A 2

Procedure

A 2-l. three-necked flask, equipped with a sealed stirrer, a dropping funnel and a reflux condenser provided with a calcium chloride drying tube, is charged with 100 g. (0.625 mole) of diethyl malonate (Note 1) and 400 ml. of commercial absolute ethanol. Stirring is started, and a solution of 35 g. of potassium hydroxide pellets (Note 2) in 400 ml. of commercial absolute ethanol is added at room temperature during a period of 1 hour. A white crystalline precipitate forms during the addition, and, after all the hydroxide has been added, stirring is continued for an additional 2 hours. After the mixture has stood overnight, it is heated to boiling on the steam bath and filtered while hot with suction (Note 3). Precipitation of the potassium ethyl malonate is completed by cooling the filtrate in an ice bath. The salt is collected by suction filtration, washed with a small amount of ether, and dried under reduced pressure at room temperature. An additional amount of the potassium salt is obtained by concentrating the mother liquors on the steam bath to about 100–125 ml. The total yield is 80–87 g. (75–82%).

Compounds referenced

ethyl alcohol (64-17-5)

hydrochloric acid, HCl (7647-01-0)

Benzene (71-43-2)

methyl alcohol (67-56-1)

ether, diethyl ether (60-29-7)

hydrogen (1333-74-0)

acetonitrile (75-05-8)

sodium bicarbonate (144-55-8)

Benzoic acid (65-85-0)

Angewandte Chemie *journal article supporting data*

distilled under nitrogen from molten sodium. THF was distilled under argon from sodium benzophenone ketyl. Unless stated otherwise, commercially obtained materials were used without purification. Amines were purified by passing them through a short plug of alumina. Aryl halides were purchased from Aldrich Chemical company with the exception of 2-bromobiphenyl which was purchased from Lancaster Synthesis Inc. Chlorodicyclohexylphosphine, palladium acetate, and *n*-butyllithium were purchased from Strem Chemical company. Di-*tert*-butylchlorophosphine, potassium

Hamlet Act III Scene 1

Read on this book;
That show of such an exercise may colour
Your loneliness. We are oft to blame in this,--
'Tis too much proved--that with devotion's visage
And pious action we do sugar o'er
The devil himself.

KING CLAUDIUS

[Aside] O, 'tis too true!
How smart a lash that speech doth give my conscience!
The harlot's cheek, beautied with plastering art,
Is not more ugly to the thing that helps it
Than is my deed to my most painted word:
O heavy burthen!

LORD POLONIUS

I hear him coming: let's withdraw, my lord.

Chemical name to structure conversion

- CambridgeSoft – ‘name=struct’
- ACD/Labs – ACD/Name to structure batch
- OpenEye Scientific Software – Nam2mol

Name=Struct: Features

- Converts names to chemical structures
  - IUPAC names
  - CAS names
  - common usage names (largest fraction of names)
- Cahn-Ingold-Prelog stereochemistry (R, S, E, Z, r, s)
- cis and trans relative stereochemistry
- Typo recognition and correction
- Flexible – handles variation well
- Batch version – 30,000 names processed per minute

Name=Struct: Six Principles

- **Anything allowed by the rules is acceptable**
  - IUPAC recommendations
  - CAS
- **Anything forbidden by the rules is acceptable**
  - Alphabetical ordering of substituents
    - » 3-bromo-2-chlorobenzoic acid
    - » 2-chloro-3-bromobenzoic acid
  - **Older rules that have been deprecated should still work**
    - » Dated works cannot be modified retroactively
    - » Both current and obsolete systems should be supported

Name=Struct: Six Principles

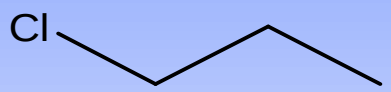
- **Any rule that can be extended should be**
  - Hantsch-Widman nomenclature for heterocycles
    - » CAS – 14 heteroatoms
    - » IUPAC – 19 heteroatoms

- **If it looks like it ought to be a rule, it is a rule**
  - Neither IUPAC or CAS rules are complete
  - Acetone, butyrone, valerone imply rule: (trivial acid name) + “one”: = di(number of carbons in trivial acid – 1)yl ketone

    stearic acid = C18 acid -> stearone = di(C17) ketone

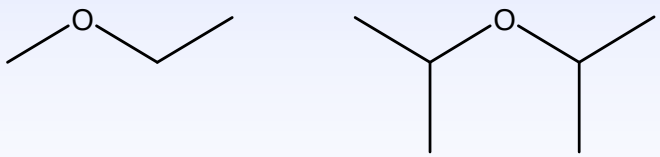
Name=Struct: Six Principles

- **The meaning of logically ambiguous names is determined by common usage**
  - e.g. propyl chloride is ambiguous, but common usage assumes that undifferentiated alkanes are in the unbranched forms



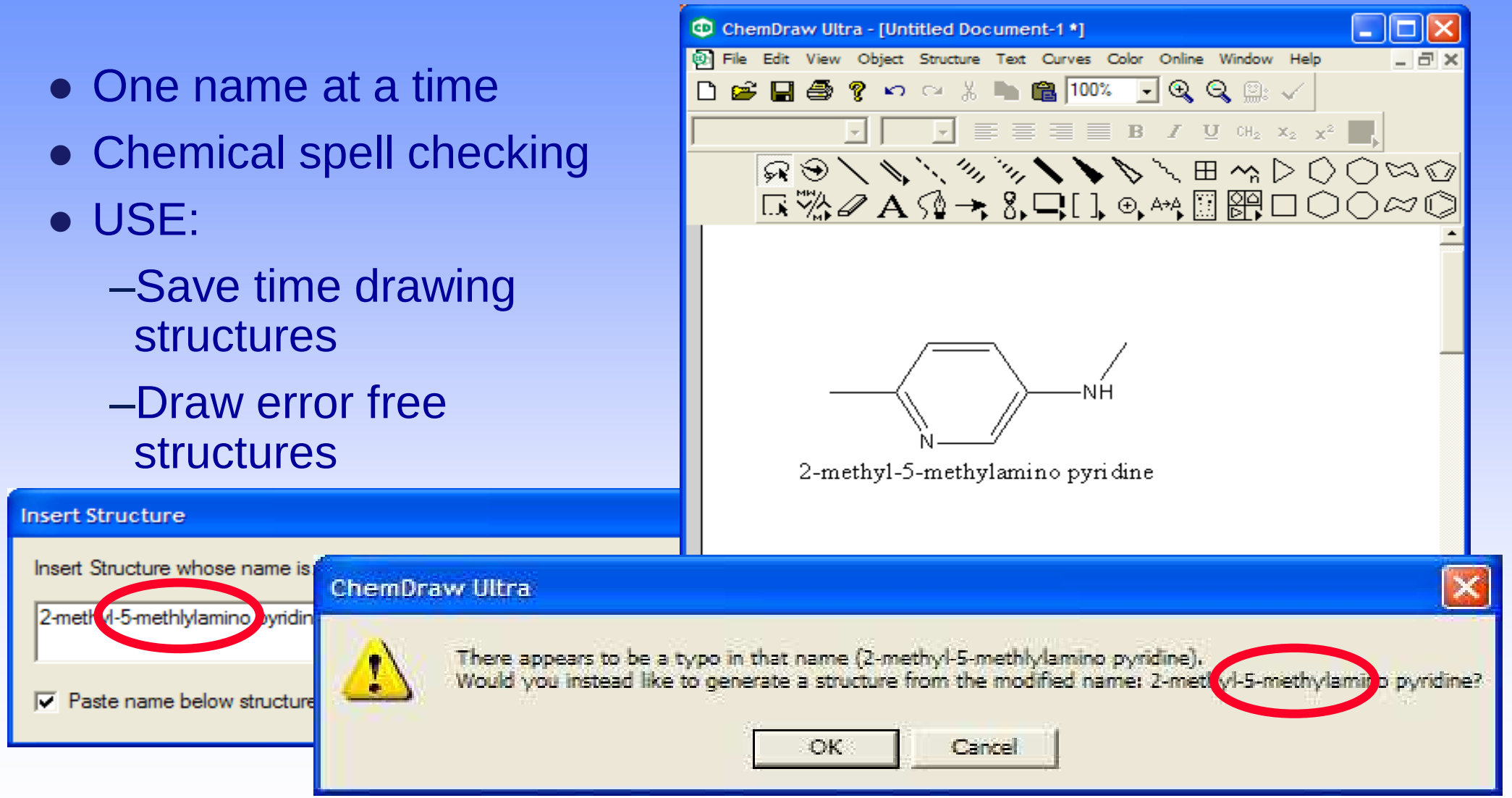
propyl chloride

- **Punctuation and capitalization do not matter; spacing matters as little as possible**
  - *p*-Dinitrobenzene vs. P-DINITROBENZENE
  - methyl ethyl ether vs. methylethyl ether



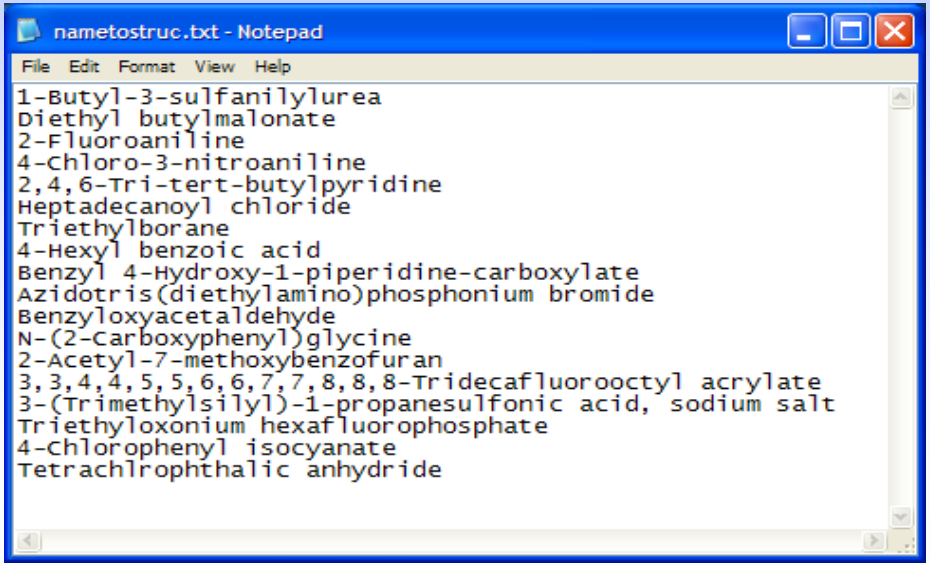
Name=Struct: ChemDraw

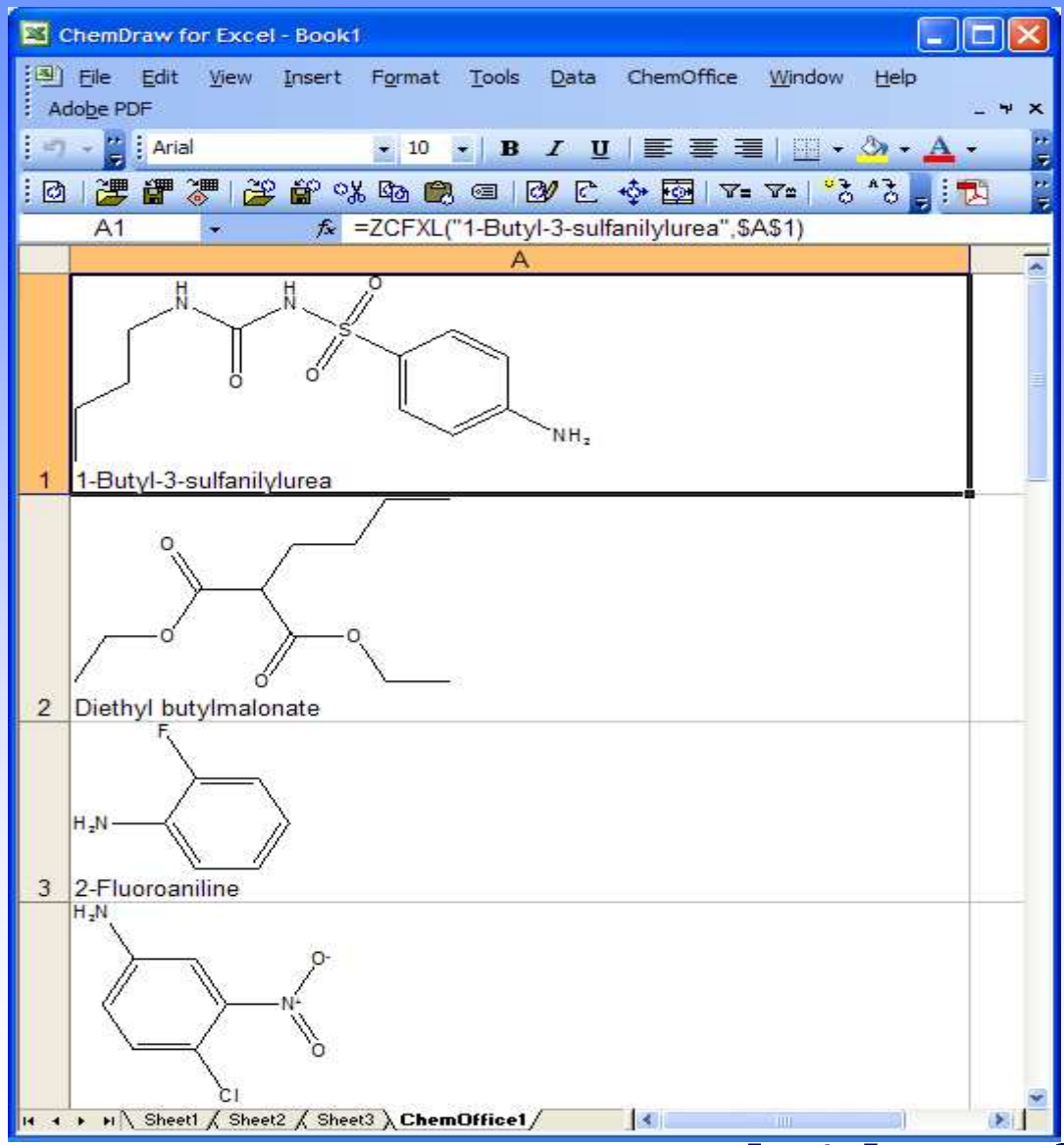
- One name at a time
- Chemical spell checking
- USE:
  - –Save time drawing structures
  - –Draw error free structures



Name=Struct: ChemDraw for Excel

- Multiple names
- Export as SD file
- USE: create a chemical spreadsheet or small database





CambridgeSoft

Name=Struct: Batch

- Multiple names
- Text in -> SD file out
- Command line mode
- Programmable
- USE: database building and cleaning

Summary

- Personal view of the history of text processing
- Context – why is text processing important?
- Why is chemistry a problem?
- How the problem is solved
- Some examples of the name to chemical structure technology in action

Any questions?