



What makes Chemical Information different?

David Walsh, Pfizer

Agenda:

Chemical information is timeless

Chemical Information is large

Chemistry is about making new substances

How do we describe molecules? (*nomenclature, structure and more...*)



Chemistry is an old science



- Chemistry Information does not go out of date
- Properties of chemicals reported in 1900 are still valid
- Methods of making chemicals based on chemical transformations reported in 1900 are still used.
- Information services covering chemistry date back to 1817 (Gmelin) and 1880 (Beilstein) and 1907 (Chemical Abstracts)



The size of chemical information

To be up to date in all branches of chemistry, a chemist would have to read 2,000 papers a day

If the chemist read abstracts only they would need to read 70,000 pages per year

The information would be published in 8000 journals

Then there are patents to consider and conferences to attend!



The size of chemical information

As this is impossible, the process of keeping up to date in chemistry is delegated to abstracting databases.

Notably Chemical Abstracts and the end-user versions of Scifinder (Scholar)



Making new substances



- Chemistry is about the generation of new substances
- The majority of chemists are engaged in making new substances
- On average, every paper in Chemical Abstracts reports 2 new substances
- ~25 million substances have been reported in Chemical Abstracts to date



Growth of chemical information



- ◆ Chemical Information is growing at exponential rates
- ◆ It is anticipated that 80 million substances will have been reported by 2025
- ◆ And 300 million substances by 2050
- ◆ There is no limit to the potential number of chemical substances



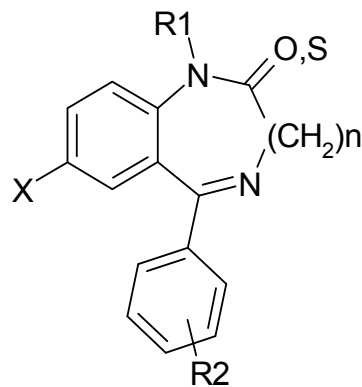
Making new substances

Single compounds can be intrinsic or exemplified

Especially in patents, compounds can be generically claimed in **Markush** structures

The number of potential compounds defined by a single Markush structure can be infinite

Making new substances



R1=alkyl, alkenyl,alkoxyalkyl(c1-c4)

R2=optionally substituted with OH,NH₂,-O-Ph,X (1-4 halogens)

n=0-4

Examples of Markush claims of a molecule in a patent:
The numbers of potential compounds are infinite



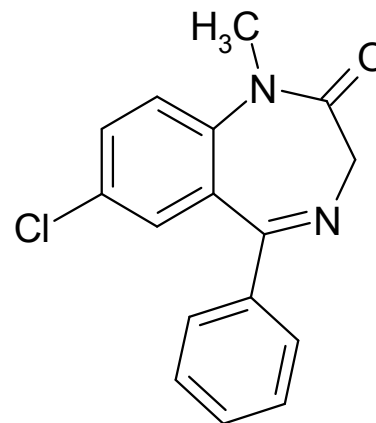
Why are so many different ways to define chemicals necessary?

- ◆ **How do you define a chemical?**

- ◆ Structure
 - Systematic Names
 - Generic Name
 - Trade Names
 - Laboratory codes
 - Registry Numbers
 - Molecular formulae
 - Notations
 - Fragment Codes
 - Connection Tables

Why are so many different ways to define chemicals necessary?

Chemicals are not amenable to linear representation





Why are so many different ways to define chemicals necessary?

As chemicals are invented and developed, additional names are associated with them.

First the chemist draws a structure

Then a laboratory code is assigned

Then a systematic name

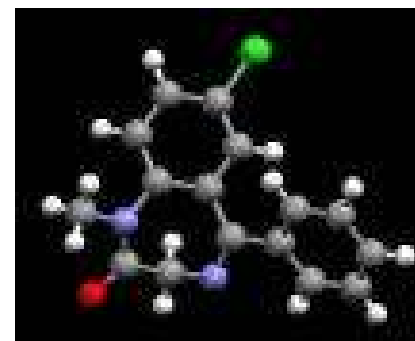
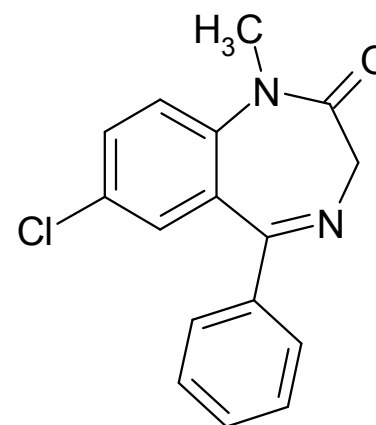
Then a generic name

Then a trade name

Why are so many different ways to define chemicals necessary?

Chemicals may be described at various levels of precision

They can be defined implicitly or generically,
by substructure,
by functional groups,
by chemical classes,
by stereochemistry,
by shape



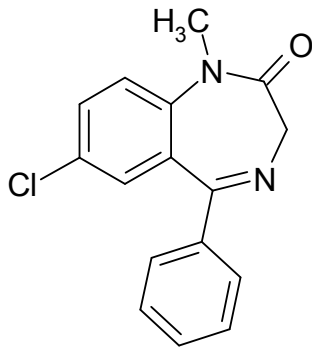


Why are so many different ways to define chemicals necessary?

- Historically, nomenclature changes.
- Chemists may describe a molecule in relation to chemical classes to emphasise a particular point.
- Chemicals are articles of commerce, hence trade names proliferate

Structure

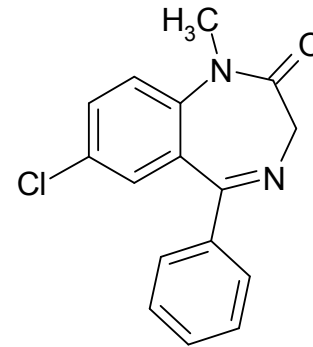
- ◆ Diazepam, (Valium)



Laboratory Codes

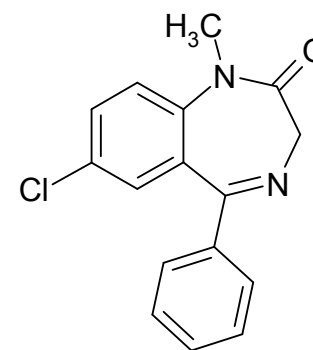
Laboratory codes are the Registry numbers of the inventors of the compounds

- La III
- Ro 5-2807
- Wy-3467
- NSC-77518



Systematic Names

- ◆ Systematic Names
- ◆ IUPAC and CAS
- ◆ Unambiguous name defined to structure but complex!
- ◆ Also, the rules of nomenclature can be applied with different dialects giving different strings which are valid
- ◆ 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one
- ◆ 7-chloro-1,3-dihydro-1-methyl-5-phenyl-3H-1,4-benzodiazepin-2(1H)-one





Generic Name

Diazepam

- ◆ USAN
- ◆ BAN
- ◆ INN

Few molecules achieve Generic names (only those destined for pharmaceutical or Agrochemical applications)

Sometimes, generic names differ geographically, e.g. Acyclovir, Aciclovir

Trade Names

Trade Names are associated with particular companies, or particular formulations containing that molecule.

Different trade names may be associated with different country markets.

- Alboral, Aliseum, Amiprol, Alupram, An-Ding, Ansioisina, Anisiolin, Apaurin, Apozepam, Assival, Atensine, Atilen, Bialzepam, Calmocitene, Calmpose, Cercine, Cereglart, Cristalia, Diapam, Dialar, Diazemuls, Dipam, Diazetard, Dienpax, Dipezona, Domalium, Duxen, Eridan, Eurosan, Evacalm, Faustan, Freudal Gewacalm, **Horizon**, Kiatrium, Lamra, Lembrol, Levium, Liberetas, Mandrozep, Morosan, Neurolytril, Noan, Novazam, Paceum, Pacitran, Paxate, **Paxel**, Quievita, Relaminal, Relanium, Renborin, Saromet, Sedipam, Seduxen, Serenamin, Serenzin, Setonil, Sibazon, Solis, Sonacon, Stesolid, Stesolin, Tensopam, Tranimul, Tranquase, Tranquirit, Tranquo-Puren, Tranquo-Tabliten, Umbrium, Unisedil, Usempax AP, Valaxona, Valiquid, Valitran **Valium**, Valrelease, Atran, Vival, Vivol, Zipan



Registry Numbers



- Chemical Abstracts 439-14-7 (11100-37-1, 53320-84-6)
- Beilstein Registry Numbers BRN 0754371
- RTECS DF1575000
- Derwent Compound Number R01255
- Derwent Chemical Resource 18536-0-0-0

Registry numbers are sequential numbers assigned by computer with little information relating to their identity.



Molecular formula



C₁₆H₁₃ClN₂O

Provides an ambiguous definition of the molecule.

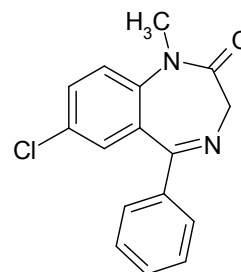
Hill Formula defines the order of the atoms present:

Carbon followed by Hydrogen; then other atoms in alphabetical order

Hydrogen atoms not normally drawn in chemical structures

Notations

- Computer readable or generated linear strings
- Smiles strings
- ClC1(=CC2(=C(C=C1)N(C)C(=O)CN=C2C3(=CC=CC=C3)))
- Wiswesser Line Notation
- T67 GNV JN IHJ CG G1 KR



InChI

An open source method of describing molecules has recently been developed for IUPAC by Dmitrii Tchekhovskoi, Steve Stein and Steve Heller at NIST. Chemical structures are expressed in terms of five layers of information – connectivity, tautomeric, isotopic, stereochemical and electronic. The unique connectivity layer is essential.

Possible future applications of InChI include: ordering chemicals from suppliers; finding compounds in the literature via text-based search engines like Google; or passing the identity of a substance to a colleague for use in these or other applications.

Diazepam

InChI=1/C16H13ClN2O/c1-19-14-8-7-12(17)9-13(14)16(18-10-15(19)20)11-5-3-2-4-6-11/h2-9H,10H2,1H3



Notations for Peptides, Polypeptides, Proteins and polymers

B-ENDORPHIN

Tyr-Gly-Gly-Phe-Met-Thr-Ser- Glu Lys-Ser-Gln-Thr-Pro- Leu-
Val-Thr-Leu-Phe-Lys-Asn- Ala-Ile-Ile-Lys-Asn Ala-His- Lys-Lys-
Gly-Glu-OH)

YGGFMTSEKSQTPLVTLFKNAIKNAKKGE-OH

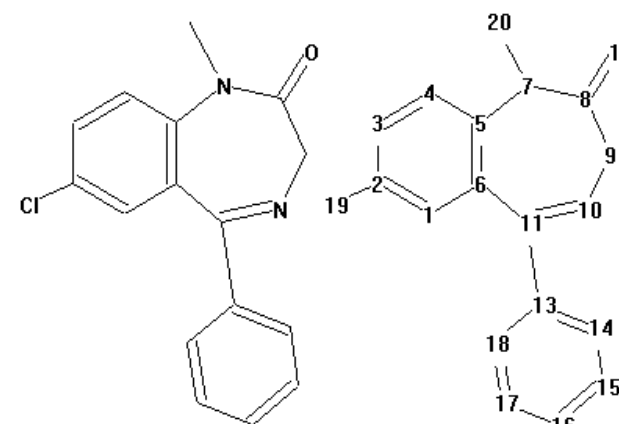
Linear notations are useful for proteins and DNA

Synthetic polymers also have notations to determine repeating
units

Connection Tables

*****CONNECTIONS*****

NOD	SYM	ROL	NOD/BON/SIT/STE	NOD/BON/SIT/STE	NOD/BON/SIT/STE
1	C		2 RN	6 RU	
2	C		3 RN	1 RN	
3	C		7 RSE	4 RN	2 RN
4	C		11 RU	5 RN	3 RN
5	C		6 RN	4 RN	
6	C		1 RU	5 RN	
7	N		12 CSE	8 RSE	3 RSE
8	C		13 CDE	9 RSE	7 RSE
9	C		10 RSE	8 RSE	
10	N		11 RDE	9 RSE	
11	C		14 CSE	4 RU	10 RDE
12	C		7 CSE		
13	O		8 CDE		
14	C		17 RN	15 RN	11 CSE
15	C		16 RN	14 RN	
16	C		19 RN	15 RN	
17	C		18 RN	14 RN	
18	C		19 RN	17 RN	
19	C		16 RN	18 RN	



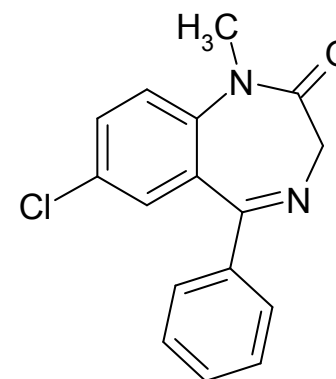
Fragment Codes

- Fragment Codes define a fixed feature of a molecule which may be found in thousands of other molecules.
- Combining fragment codes reduces the number of molecules that share some common structural similarity.
- This is a useful process for identifying molecules which have been described in a Markush structure in a patent

D780 Fused heterocycle with 2 rings
and 2 Nitrogens (C6-C5N2)

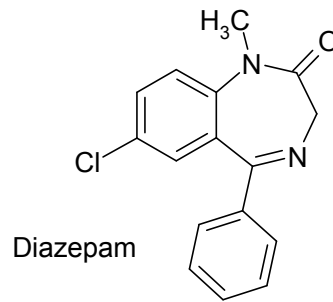
H641 1 Halogen bonded to carbocyclic aromatic
ring (from 1981 only)

J521 Carbonyl group where carbon is in a
heterocyclic ring

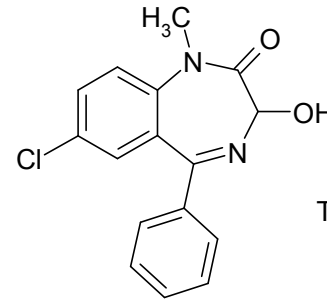


Chemical and activity classes

- Chemical Class
- Benzodiazepines

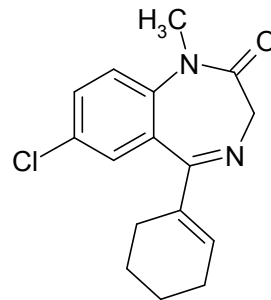


Diazepam

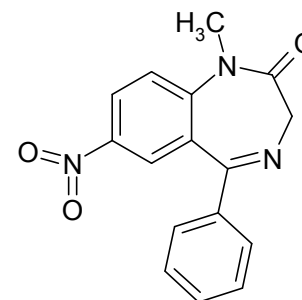


Temazepam

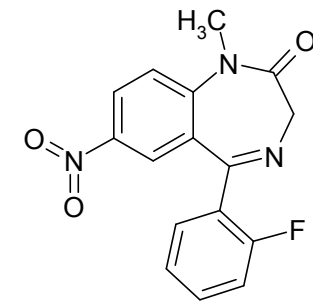
- Activity class
- Anxiolytic
- Muscle relaxant
- Hypnotic



Tetrazepam

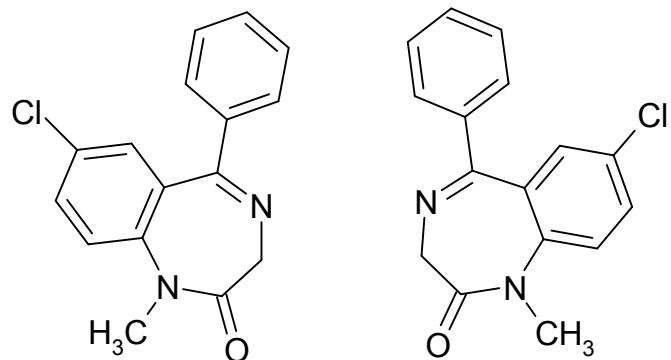
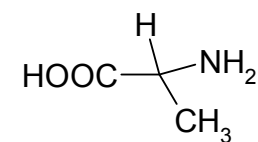
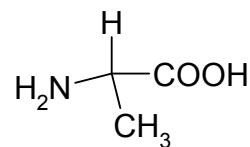
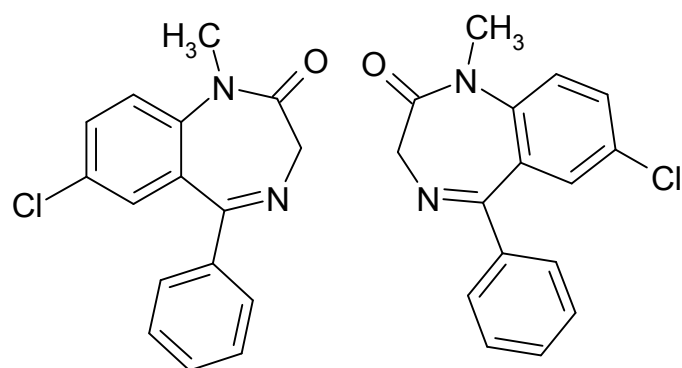


nitrazepam



flunitrazepam

Stereochemistry





Salt forms and mixtures

Drug compounds often have associated salt forms to aid solubility and other physical properties.

Drug forms are co-formulated for synergistic properties and ease of use.

Each salt form and mixture has a different Registry number and may be considered as a family.

Family searches can be made by structure search or by the use of component registry numbers (/CRN)



Crystal Structure and Polymorphs

Chemicals are not 2-dimensional but have a 3-dimensional shape

This is important for their mode of action.

Crystal structure databases allow the 3-dimensional coordinates of the molecule to be precisely identified

Polymorphs are variants of the crystal structure and give the molecule different physical and dissolution properties



Why Chemical Information is different

Many thanks for your attention

David Walsh
Information Management,
Pfizer, Sandwich