Getting it right - keeping it right. Registration Systems for 21st Century Chemistry

Graham Lock
Research Data Management

Why do we “Register” compounds ?

- Track potential intellectual property
- Manage biological and physico-chemical data
- Produce a reliable structure-searchable database
- Allow calculation of properties
- Enable SAR work to be performed

The changing nature of Registration

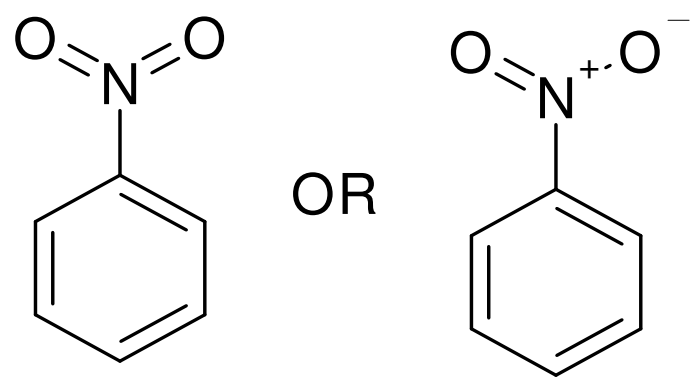
- High throughput chemistry
- Quantitative chiral chromatography
- Electronic submission / registration

What do we mean by “Registration” ?

- Standardise the representation of chemical substances
- Arrange data into a hierarchy
- Check novelty of a compound
- Assign different identifiers to different chemical structures

Standardise the representation of chemical substances - Nitro groups

- Commonly occurring
- No standardisation within the industry



(but neither is really correct - really need a delocalised bond type)

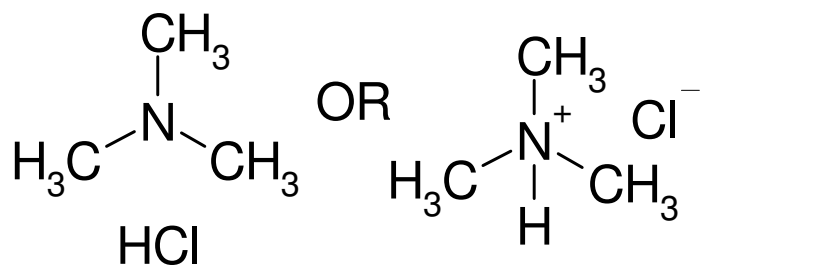
GSK Registry converts all nitro groups to the "pentavalent" form.

MODGRAPH

GlaxoSmithKline

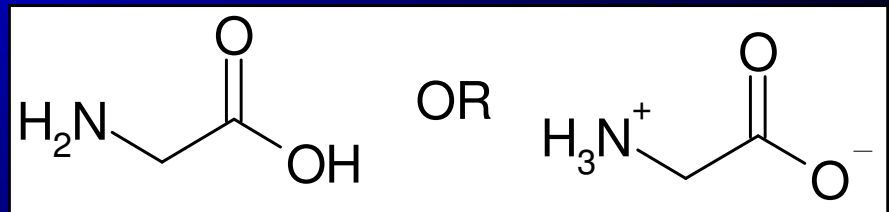
Standardise the representation of chemical substances - Salts and charges

- Salts - represent with proton transfer or not ?



(Do you standarise on an uncharged or proton transfer representation ?)

- Zwitterionic structures



(Do you standarise on an uncharged or proton transfer representation ?)

GSK Registry only accepts the uncharged representation.

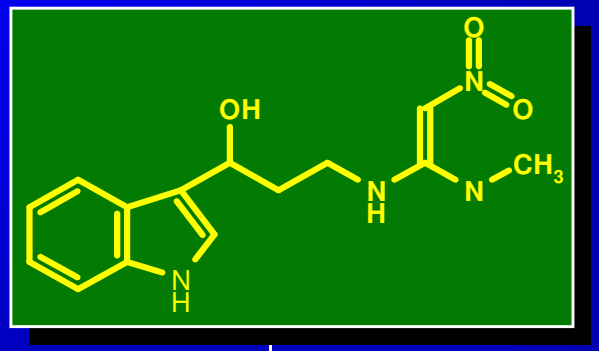
MODGRAPH

GlaxoSmithKline

What do we mean by “Registration” ?

- Standardise the representation of chemical structures
- Arranging data into a hierarchy
- Check novelty of a compound
- Assign different identifiers to different chemical structures

GSK Registration Hierarchy



Parent

Free base	Hydrochloride	Hydrate

Version

C1902/23/1 *C1422/12/1*	*C2034/2/1* *C1902/10/2*	*C1902/2/2*	**Preparation**

MODGRAPH

GlaxoSmithKline

The GSK Registry Number

- Different chemical compounds and stereoisomers get different PCNs
- Different salts, solvates and isotopic labels of the same compound get different version codes

Parent Compound Number (incremented number)

↓

GSK123456 A

↑

Version code (sequential letter, starting at "A")

MODGRAPH

GlaxoSmithKline

What do we mean by “Registration” ?

- Standardise the representation of chemical structures
- Arranging data into a hierarchy
- Check novelty of a compound
- Assign different identifiers to different chemical structures

Criteria for novelty (1)

The simple stuff is straight-forward :

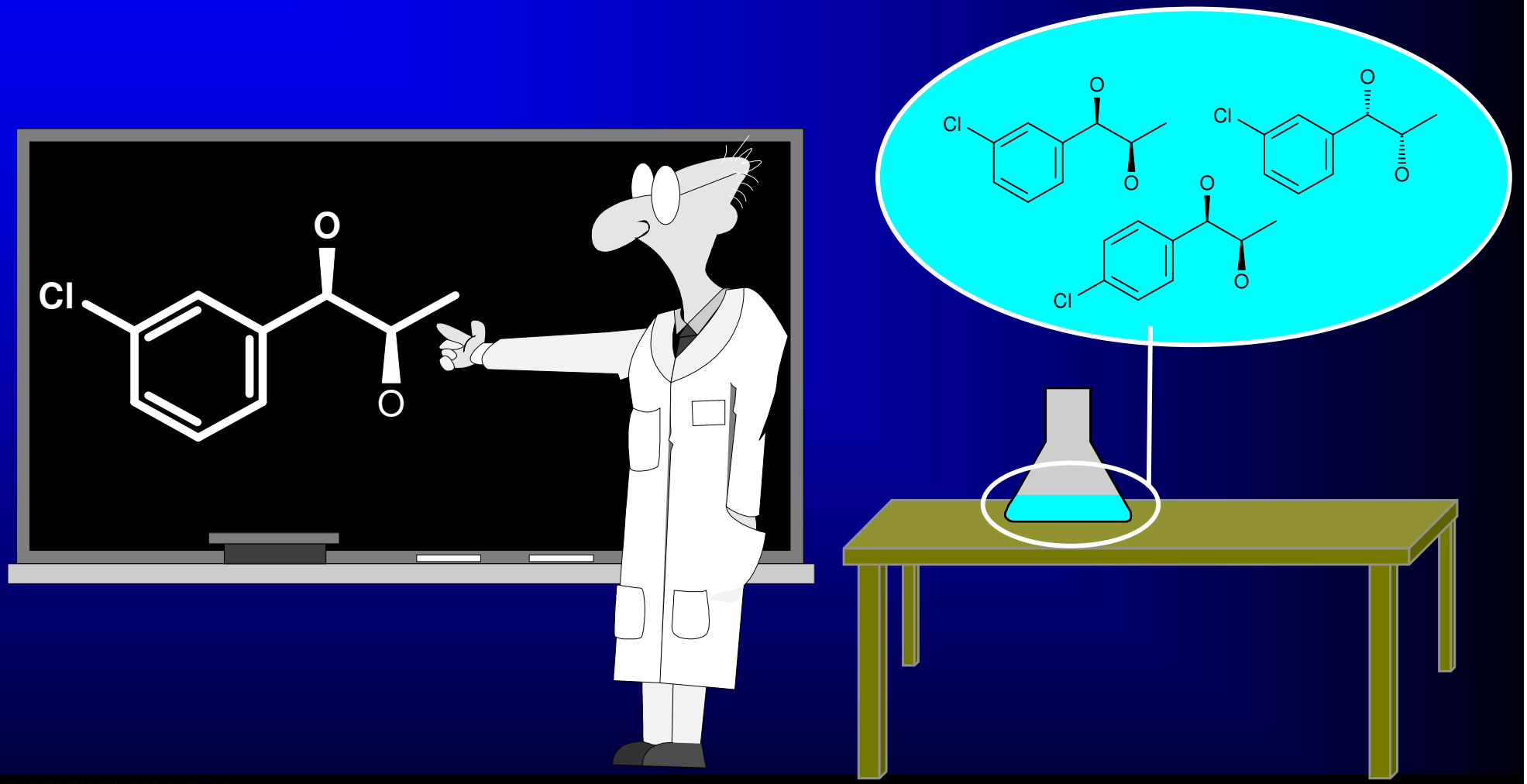
- Different chemical structure = novel
- Different enantiomer = novel
- Different geometric isomer = novel

Criteria for novelty (2)

But in reality, things are more complicated :

- Mixtures of different chemical structures = ?
- Mixtures of stereoisomers = ?
- Uncertainty in chemical structures = ?
- Uncertainty in stereoisomerism = ?

Chemical and Stereochemical Uncertainty



MODGRAPH

GlaxoSmithKline

Chemical and Stereochemical Uncertainty (2)

"It's mainly the 'R' isomer"

"The groups are relatively trans to each other"

"It's one isomer - I don't know which one"

"It's something I scraped off the bench"

The answer = Business Rules

- The foundation of any registration system
- Must be fully defined early on
- To be decided by the scientists (as they have to live with them !)

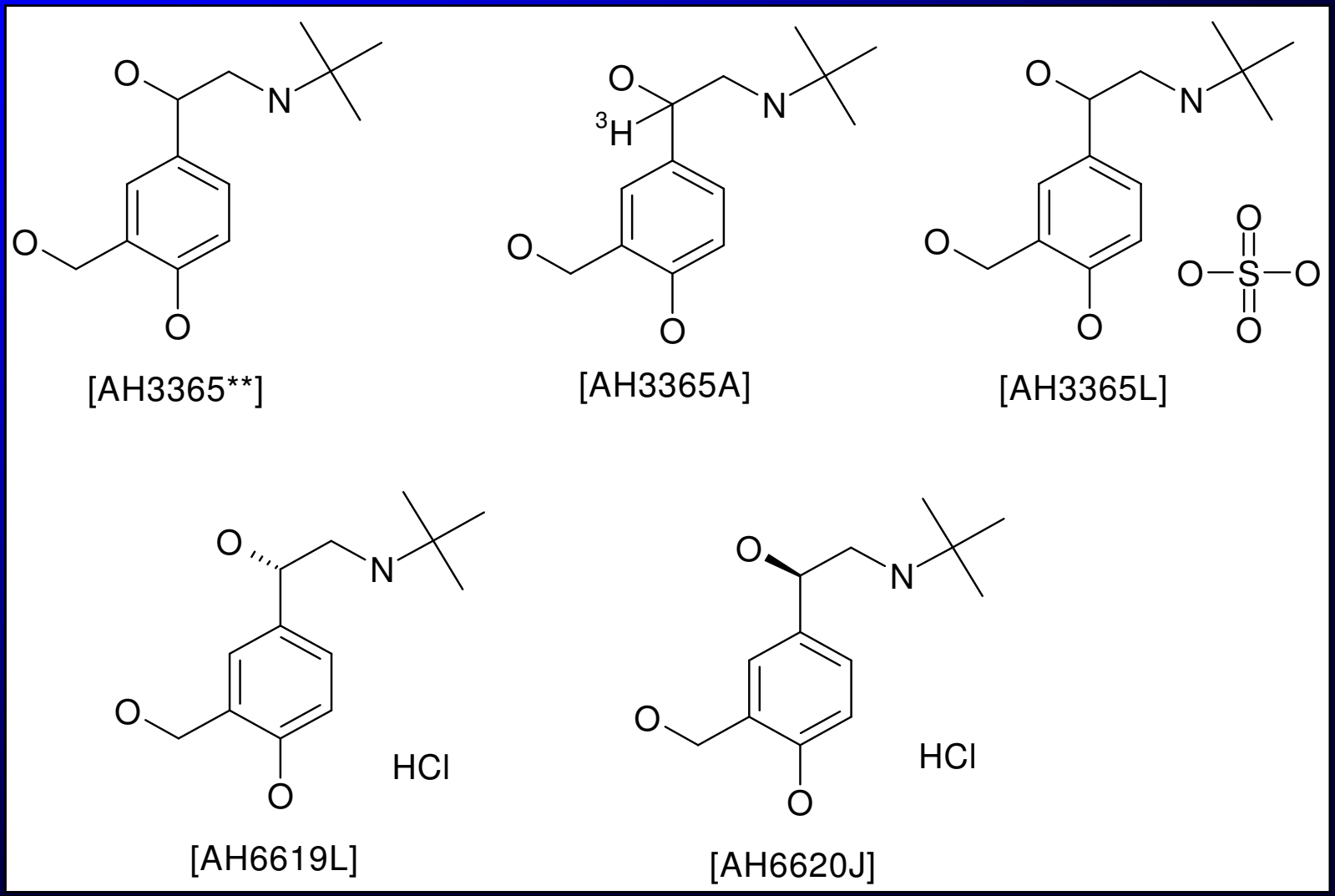
GSK Business Rules

- Defined >20 years ago
- Integral part of the GSK Registry system
- Defined ranges for mixtures of different structures or stereoisomers
- Handles uncertainty in structure or stereochemistry

What do we mean by “Registration” ?

- Standardise the representation of chemical structures
- Arranging data into a hierarchy
- Check novelty of a compound
- Assign different identifiers to different chemical structures

Assigning identifiers to chemical substances



MODGRAPH

GlaxoSmithKline

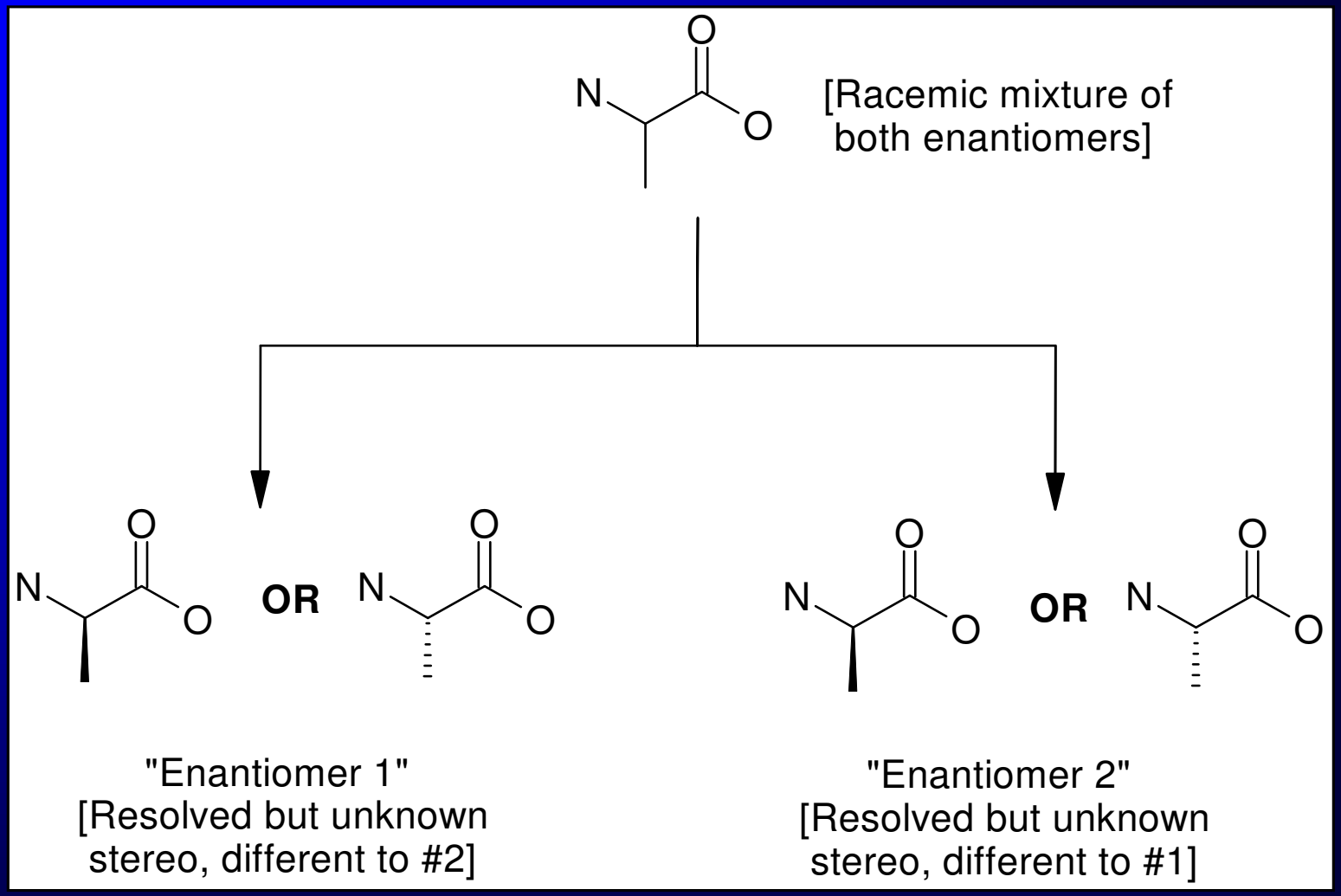
High Throughput Chemistry

- High volumes
- Lower quality ?
- Should these be “registered” ?

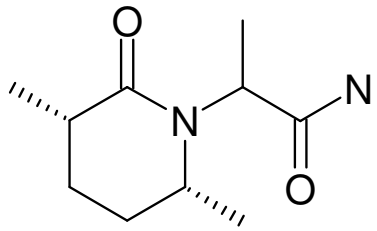
Quantitative chiral chromatography

- Used to separate stereoisomers
- Gives “pure” isomers of unknown stereochemistry
- How are these to be registered ?

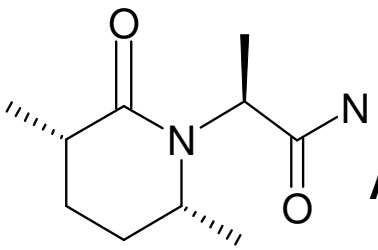
Separation of stereoisomers (1)



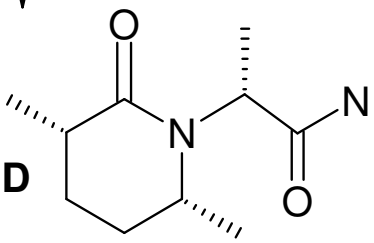
Separation of stereoisomers (2)



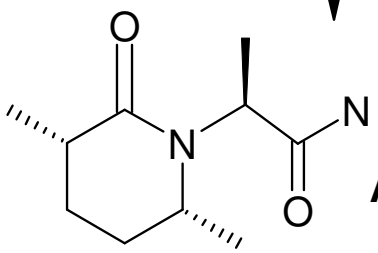
[racemic at side-chain, all other stereocentres resolved as drawn]



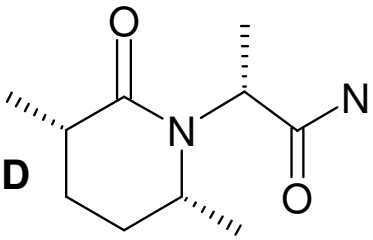
AND



[ratio of isomers 3:7 at unassigned stereocentre, major isomer unknown]

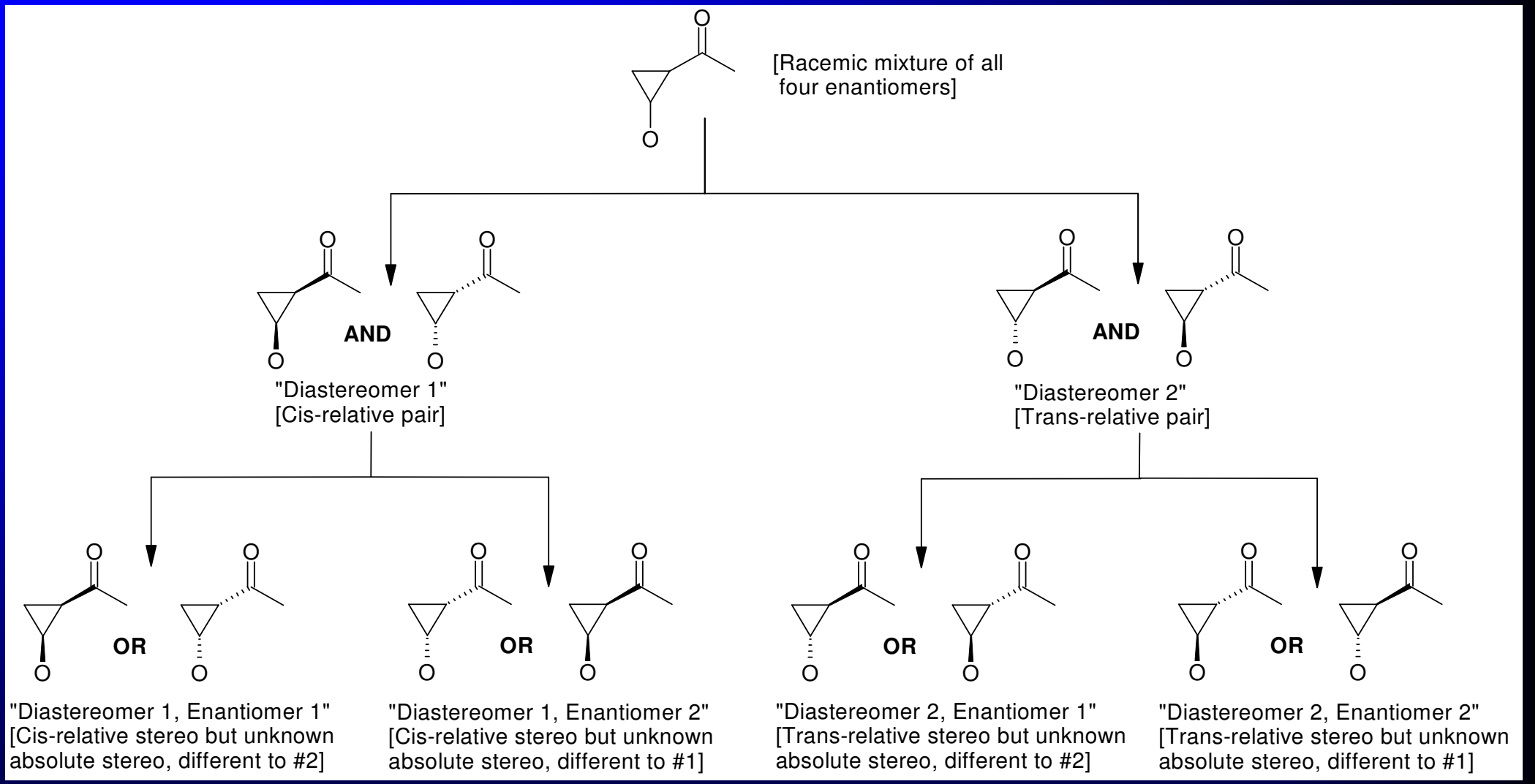


AND



[ratio of isomers 7:3 at unassigned stereocentre, major isomer unknown]

Separation of stereoisomers (3)



MODGRAPH

GlaxoSmithKline

“End-User Registration” - electronic submission or fully automated registration ?

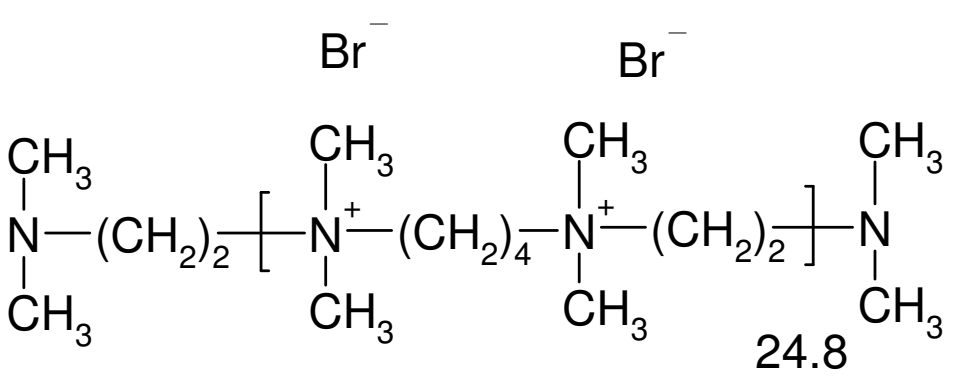
GSK approach :

- All compounds submitted electronically
- Simple compounds registered automatically
- All others registered by full-time registrars

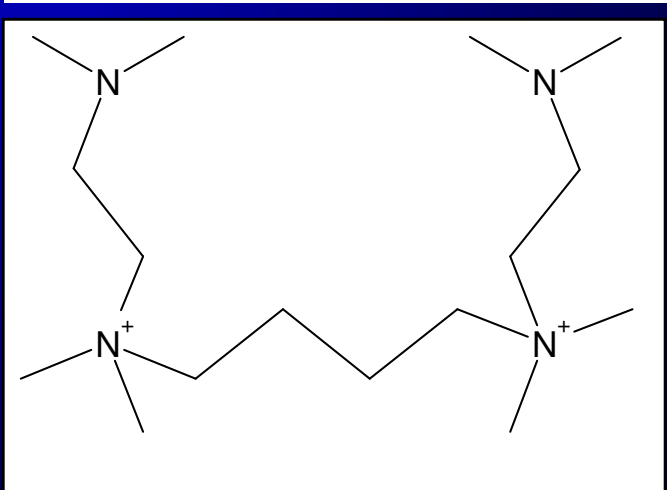
Electronic data capture

Need better methods of representing uncertainly in chemical structures (more like the way chemists draw things)

Structure as drawn in chemists hard-copy lab notebook



Structure as represented in a database (not GSK registry !)



Text comments = "DI-BROMIDE N=24.8"

MODGRAPH

GlaxoSmithKline

Amendments - “Everything must change”

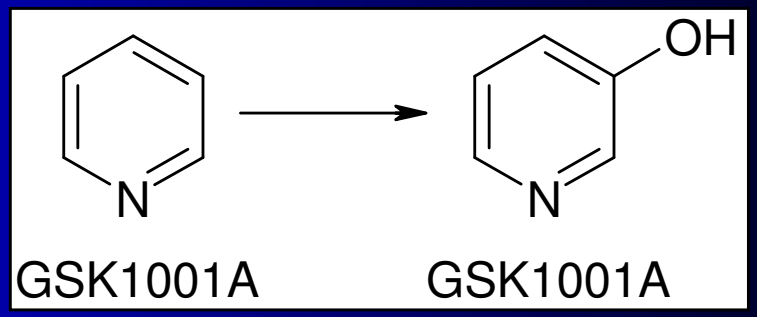
- Registration is a sub-set of “Amendments”
- A record can be amended many times
- Full auditing is required
- Have to handle collisions

Types of amendments

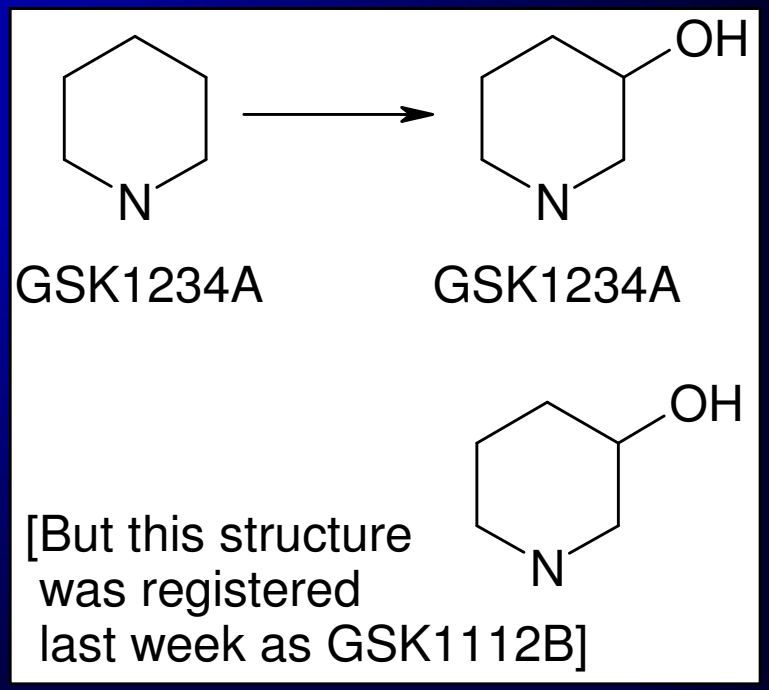
- Change the lab-notebook reference
- Modify a version record
- Modify a parent structure
- Move a preparation
- “Delete” data

Modifying structures (1)

- Simple (no collision)



- Amended structure already registered

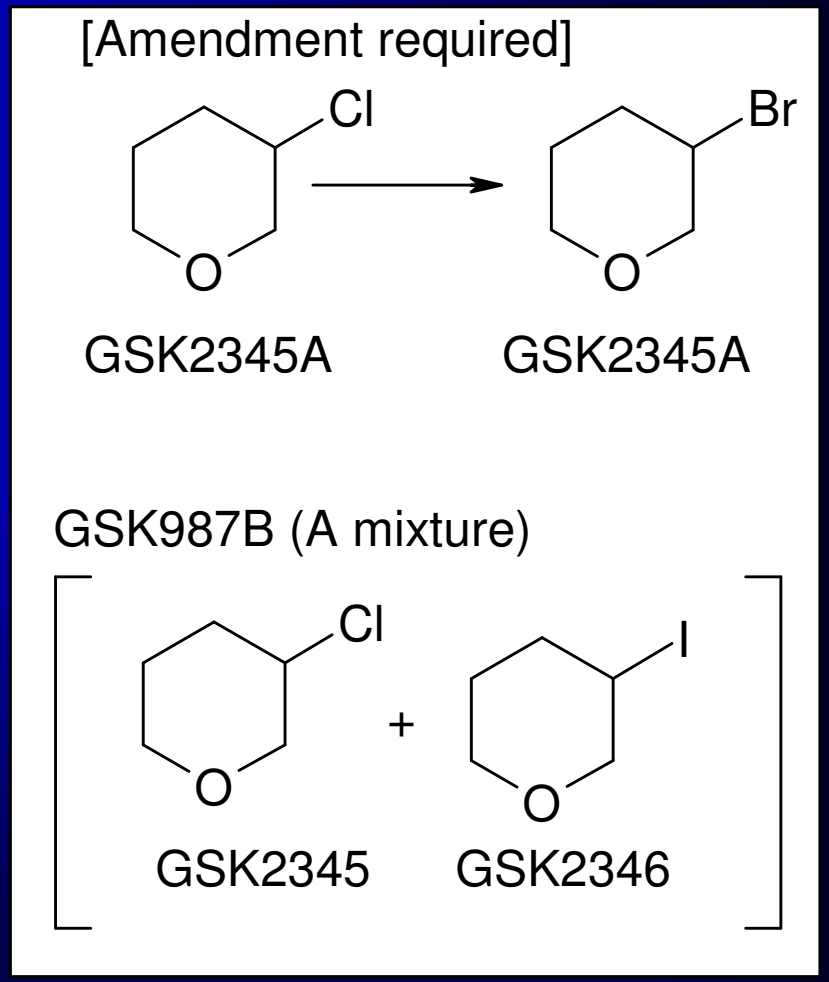


MODGRAPH

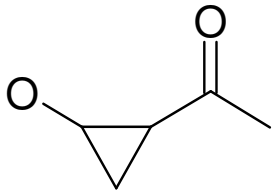
GlaxoSmithKline

Modifying structures (2)

- Amending components of “mixture” records

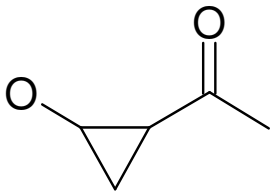


Repeated analysis and amendments (1)

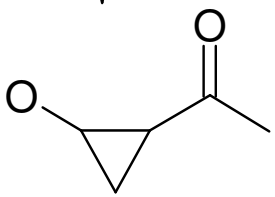


"Racemate" = GSK123
[all four possible enantiomers]

[Chiral chromatography
and registration of isomers]

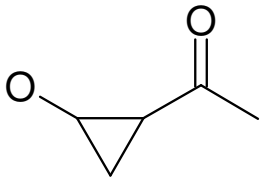


"Enantiomer 1" = GSK456
[Single enantiomer of unknown absolute stereochemistry, different to #2]

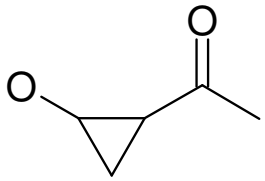


"Enantiomer 2" = GSK457
[Single enantiomer of unknown absolute stereochemistry, different to #1]

Repeated analysis and amendments (2)

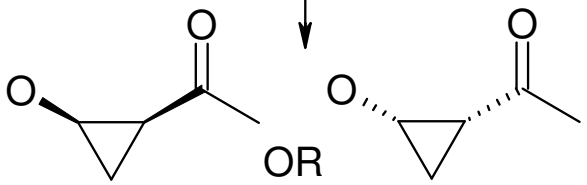


"Enantiomer 1" = GSK456
[Single enantiomer of unknown absolute stereochemistry, different to #2]

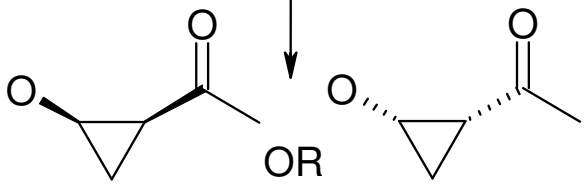


"Enantiomer 2" = GSK457
[Single enantiomer of unknown absolute stereochemistry, different to #1]

[NMR analysis and amendment]

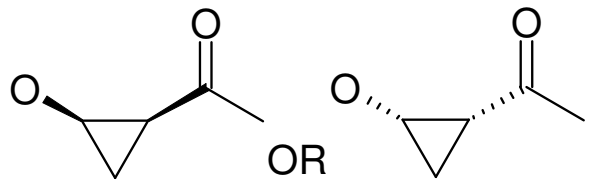


"Enantiomer 1" = GSK456
[Single Cis enantiomer of unknown absolute stereochemistry, different to #2]

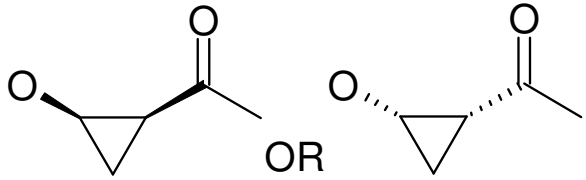


"Enantiomer 2" = GSK457
[Single Cis enantiomer of unknown absolute stereochemistry, different to #1]

Repeated analysis and amendments (3)

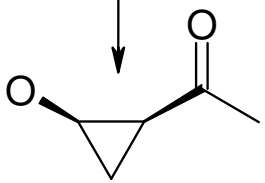


"Enantiomer 1" = GSK456
[Single Cis enantiomer of unknown absolute stereochemistry, different to #2]

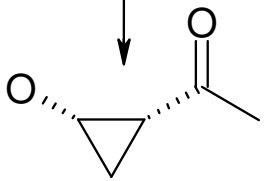


"Enantiomer 2" = GSK457
[Single Cis enantiomer of unknown absolute stereochemistry, different to #1]

[X-ray analysis and amendment]



"R,S - Enantiomer" = GSK456
[Single enantiomer with absolute stereochemistry as drawn]



"S,R - Enantiomer" = GSK457
[Single enantiomer with absolute stereochemistry as drawn]

Some Aspects of the solution

Rashmi Mistry

MODGRAPH

GlaxoSmithKline

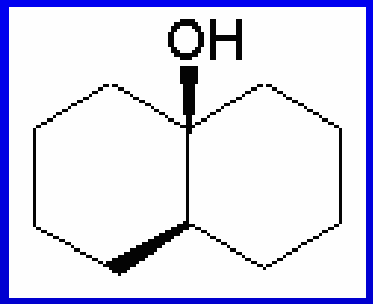
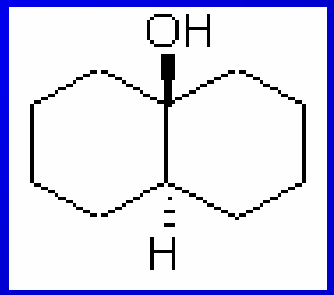
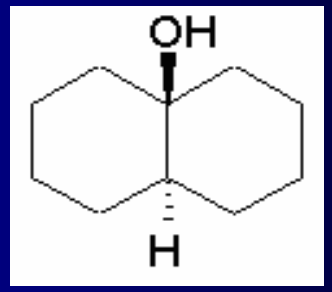
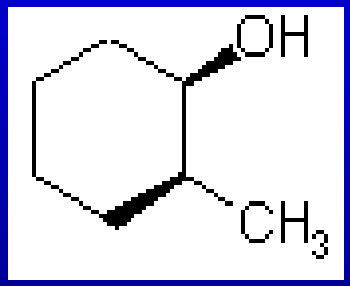
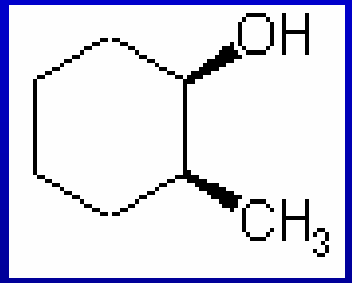
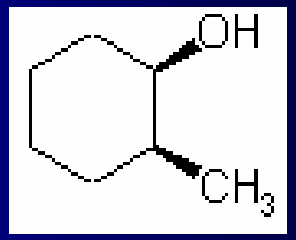
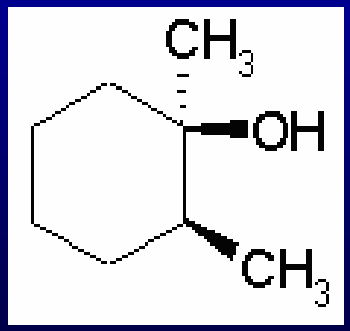
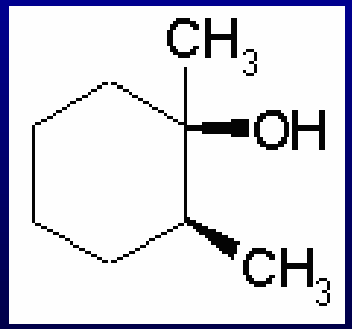
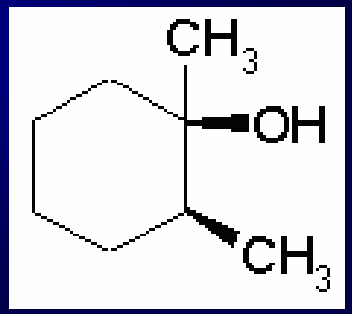
Some Aspects of the solution

- Normalisation and Standardisation/ Representation of chemical substances
- Data Model to represent the hierarchy (with examples)
- Rules for atom/bond centred data
- Data Model to represent Atom and Bond centred information (with examples)
- Data Model to represent Data and Structure Amendments
- Representation of Data and Structure on Collisions
- Extensions to SMILES to hold Structure Related information

Normalisation and Standardisation / Representation of chemical substances

- Charge connectivity table for defining the charge connectivity of each atom in the P-Table (A default set is provided with the system)
- User defined Salt/Solvate dictionary (A default set is provided)
- Data normalisation (like Nitro groups), defined as SMARTS data
- Normalisation of drawing conventions (i.e. citing of stereo bonds, drawing bridge structures)

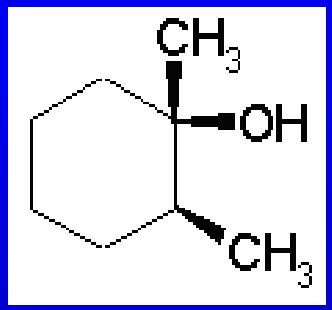
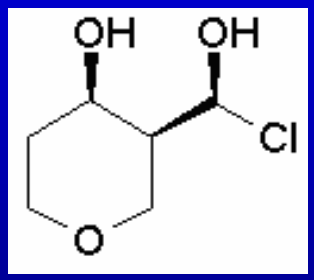
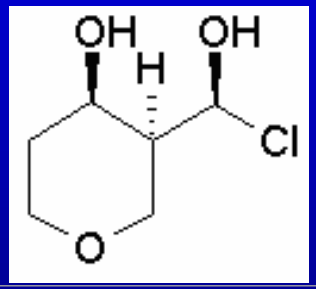
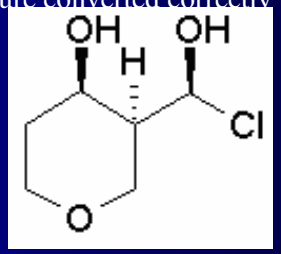
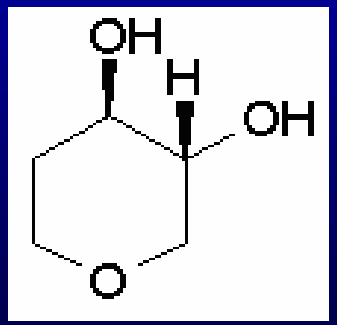
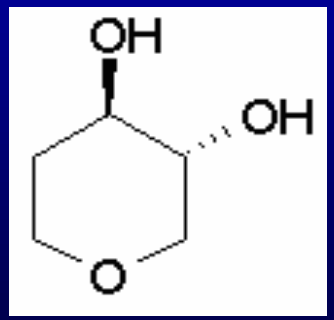
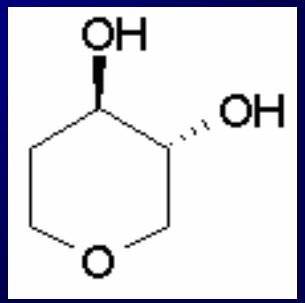
Normalisation of drawing conventions

Incorrect representation	Correct representation	Comments	Modgraph conversion module
		Stereochemistry is cited within a ring when it should be cited by adding a hydrogen with a hash/wedge bond	Structure converted correctly to give: 
		Stereochemistry is cited within a ring when it should be cited on an acyclic substituent	Structure converted correctly to give: 
		There are two thin ends of Hash/Wedge bonds at one atom. It should be cited with the minimum number of hash/wedge bonds.	Structure converted correctly to give: 

MODGRAPH

GlaxoSmithKline

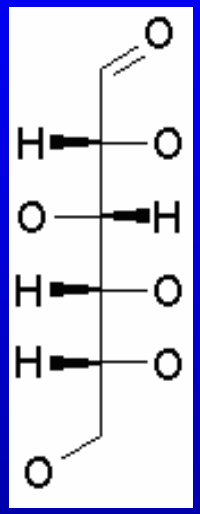
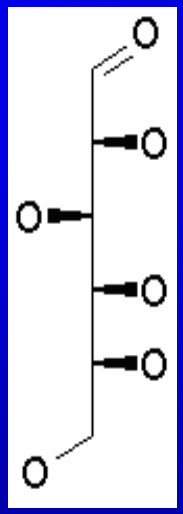
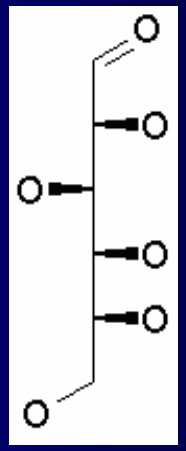
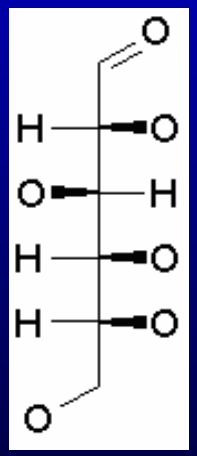
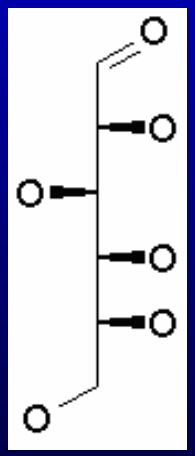
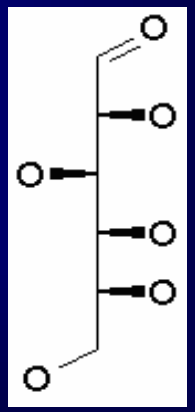
Normalisation of drawing conventions

Incorrect representation	Correct representation	Comments	Modgraph conversion module
	The inter-conversion should fail.	There are two thin ends of a Hash bond at one atom. This does not make sense so the conversion should fail	As required, the inter-conversion failed.
		Stereochemistry is cited with both the thin end and the thick end of two hash/wedge bonds at the same atom. It should be cited by adding a hydrogen with a hash/wedge bond	Structure converted correctly to give: 
		There are two thin ends of Hash/Wedge bonds at one atom. It should be cited with the minimum number of hash/wedge bonds.	Structure converted correctly to give: 

MODGRAPH

GlaxoSmithKline

Normalisation of drawing conventions

Incorrect representation	Correct representation	Comments	Modgraph conversion module
		Stereochemistry is cited with a hydrogen with a hash/wedge bond when it should be cited on an acyclic substituent. N.B. Here the hydrogen has the same configuration to the acyclic substituent due to the bond angles	Structure converted correctly to give: 
		A non-stereo bond to hydrogen has been added when it is not required. N.B. The configuration of the bond to the acyclic substituent should not be changed when the hydrogen is dropped.	Structure converted correctly to give: 

MODGRAPH

GlaxoSmithKline

Data Model to represent the hierarchy

Table:R_PARENT

Name	Null?	Type
DB_NO	NOT NULL	NUMBER
SMILES	NOT NULL	VARCHAR
PCN (Parent Compound Number)	NOT NULL	VARCHAR

Table:R_VERSION

Name	Null?	Type
DB_NO	NOT NULL	NUMBER
SMILES	NOT NULL	VARCHAR
PARENT_DB_NO	NOT NULL	NUMBER
REGNO (Registration Number)	NOT NULL	VARCHAR
SALT_INFO		VARCHAR

Table:R_PREP

Name	Null?	Type
DB_NO	NOT NULL	NUMBER
VERSION_DB_NO	NOT NULL	NUMBER
LNB	NOT NULL	VARCHAR

Example 1a

Table:R_PARENT

Name	Value
DB_NO	2311
SMILES	“CC(C)(C)NCC(O)c1ccc(O)c(CO)c1”
PCN (Parent Compound Number)	“AH3365”

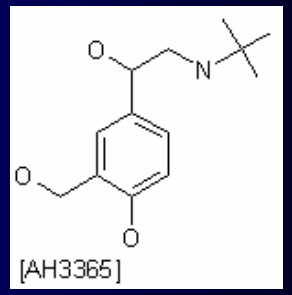
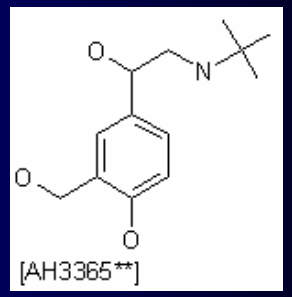
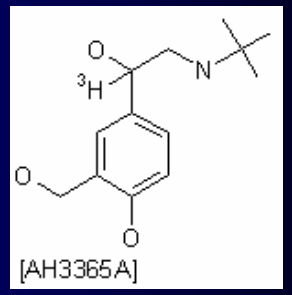


Table:R_VERSION

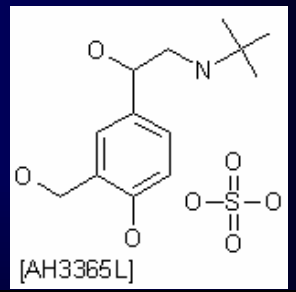
Name	Value
DB_NO	2312
SMILES	“CC(C)(C)NCC(O)c1ccc(O)c(CO)c1”
PARENT_DB_NO	2311
REGNO (Registration Number)	“AH3365**”
SALT_INFO	“”



DB_NO	2313
SMILES	“[3H]C(O)(CNC(C)(C)C)c1ccc(O)c(CO)c1”
PARENT_DB_NO	2311
REGNO (Registration Number)	“AH3365A”
SALT_INFO	“”



DB_NO	2314
SMILES	“CC(C)(C)NCC(O)c1ccc(O)c(CO)c1.OS(=O)(=O)O”
PARENT_DB_NO	2311
REGNO (Registration Number)	“AH3365L”
SALT_INFO	“9,1”



MODGRAPH

GlaxoSmithKline

Rules for atom/bond centred data

- Allows the user to annotate the molecule with specific symbols that are defined by the business rules
- Atom symbols are:-
  - "*" To represent a fully resolved atom centre
  - "M" To represent a mixture of enantiomers in a specified range A%-B%
  - "U" To represent resolved chiral centres (A%-B%) at which the absolute stereochemistry is unknown, with an Isomer N text.
- Bond symbols are:-
  - "M" To represent a E/Z mixture about a double bond in a specified range A%-B%
  - "U" To represent unknown configurations
  - "X" To represent mixture of geometric isomers <> A%-B%

Data Model to represent Atom and Bond centred information

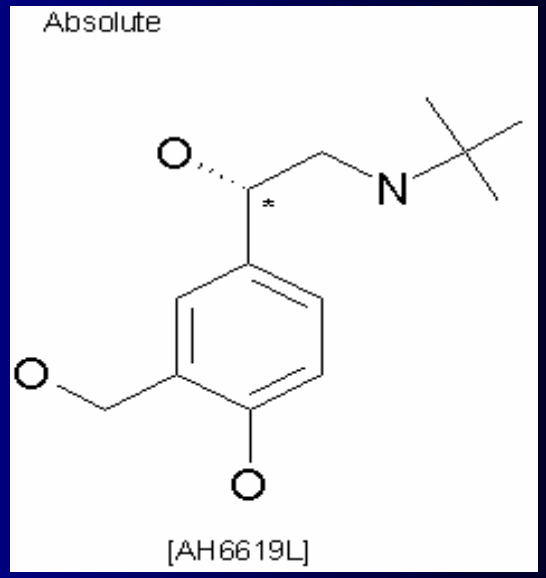
Table:R_PARENT

Name	Null?	Type
DB_NO	NOT NULL	NUMBER
SMILES	NOT NULL	VARCHAR
PCN (Parent Compound Number)	NOT NULL	VARCHAR
ABSOLUTE_FLAG		VARCHAR
ISOMER_ID		NUMBER
ISOMER_RANGE		NUMBER
ACD (Atom Centred Data)		VARCHAR
BCD (Bond Centred Data)		VARCHAR

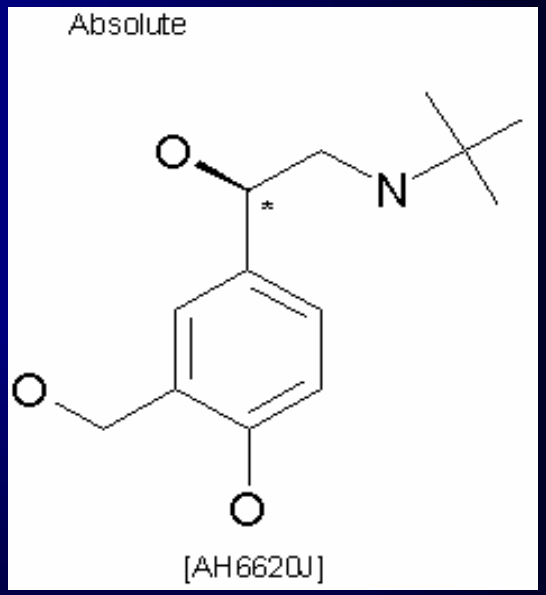
Example 1a

Table:R_PARENT

Name	Value
DB_NO	2320
SMILES	“CC(C)(C)NC[C@@H](O)c1ccc(O)c(CO)c1”
PCN (Parent Compound Number)	“AH6619”
ABSOLUTE_FLAG	“T”
ISOMER_ID	0
ISOMER_RANGE	0
ACD	“6,1 ”
BCD	“ ”



Name	Value
DB_NO	2321
SMILES	“CC(C)(C)NC[C@H](O)c1ccc(O)c(CO)c1”
PCN (Parent Compound Number)	“AH6620”
ABSOLUTE_FLAG	“T”
ISOMER_ID	0
ISOMER_RANGE	0
ACD	“6,1 ”
BCD	“ ”



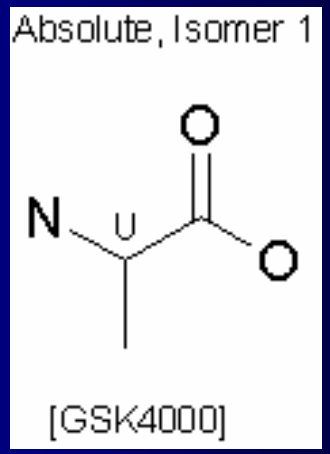
MODGRAPH

GlaxoSmithKline

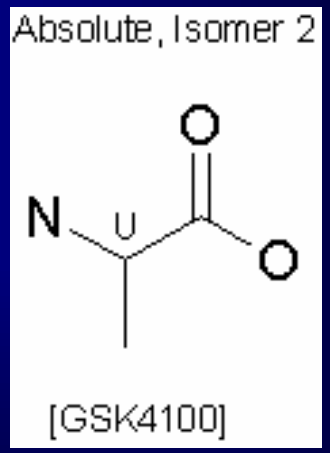
Example 2

Table:R_PARENT

Name	Value
DB_NO	3000
SMILES	“CC(N)C(=O)O”
PCN (Parent Compound Number)	“GSK4000”
ABSOLUTE_FLAG	“T”
ISOMER_ID	1
ISOMER_RANGE	0
ACD	“1,3”
BCD	“ ”



Name	Value
DB_NO	3010
SMILES	“CC(N)C(=O)O”
PCN (Parent Compound Number)	“GSK4100”
ABSOLUTE_FLAG	“T”
ISOMER_ID	2
ISOMER_RANGE	0
ACD	“1,3 ”
BCD	“ ”



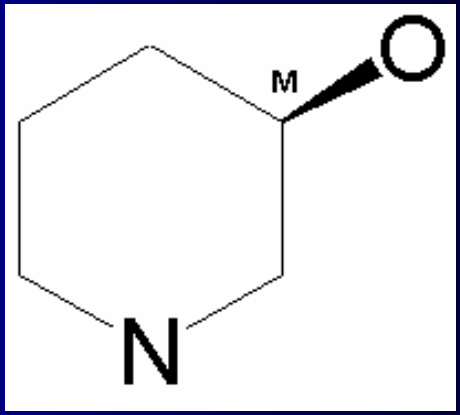
MODGRAPH

GlaxoSmithKline

Example 3

Table:R_PARENT

Name	Value
DB_NO	5340
SMILES	“O[C@@H]1CCCNC1”
PCN (Parent Compound Number)	“GSK5000”
ABSOLUTE_FLAG	“”
ISOMER_ID	0
ISOMER_RANGE	0
ACD	“1,2”
BCD	“ ”

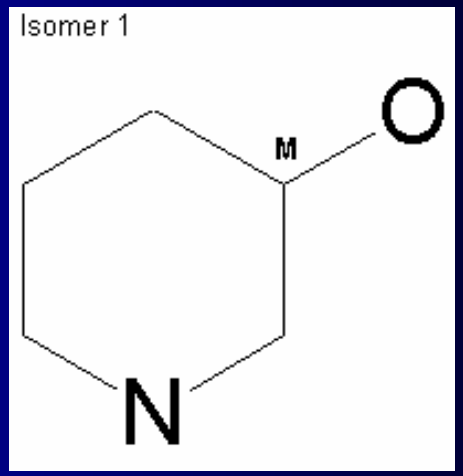


This display represents a mixture of enantiomers in which the predominant isomer is present in the range A%-B%. The configuration of the predominant isomer is known and is therefore drawn.

Example 4

Table:R_PARENT

Name	Value
DB_NO	5610
SMILES	“OC1CCCNC1”
PCN (Parent Compound Number)	“GSK6070”
ABSOLUTE_FLAG	“”
ISOMER_ID	1
ISOMER_RANGE	0
ACD	“1,2”
BCD	“ ”



A mixture of enantiomers is present, the major of which falls within the range A%-B%. Since the identity of the predominant isomer is unknown normal bonds are used in the display with a compulsory text field descriptor ISOMER n.

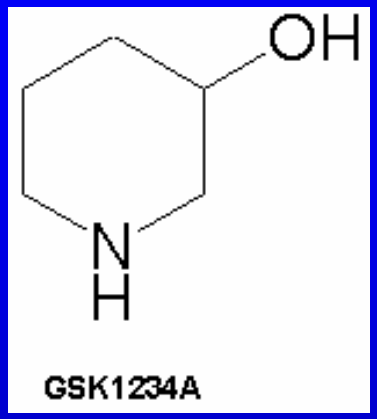
Data Model to represent Data and Structure Amendments

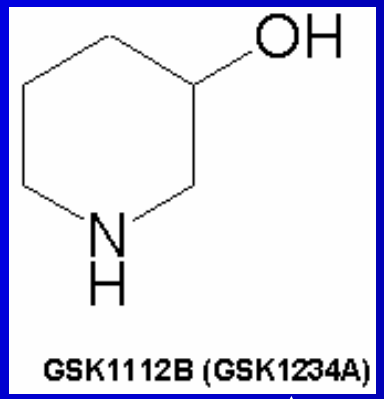
- Each of the Parent, Version and Preparation tables are duplicated as audit tables
- The system performs Oracle row level auditing
- Any data/structure change at any level in the hierarchy is simply copied into the audit table

Representation of Data and Structure on Collisions

- Collisions need to be handled both at Parent, Version
- The system has record status that enables this to be managed, this status is known as ‘Preferred/’Non Preferred’
- The assignment of ‘Preferred/’Non Preferred’ is a business process.
- A ‘Preferred’ structure may have many ‘Non Preferred’ structures

Preferred Non Preferred Example



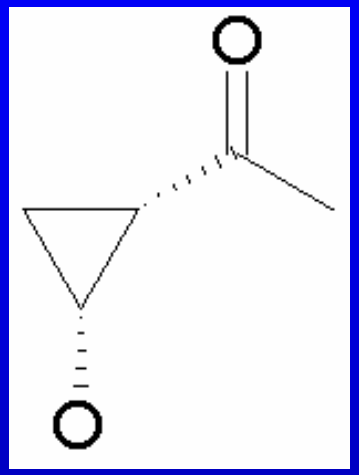


This is the 'Non Preferred' compound

Has reference to the preferred compound (GSK1234A)

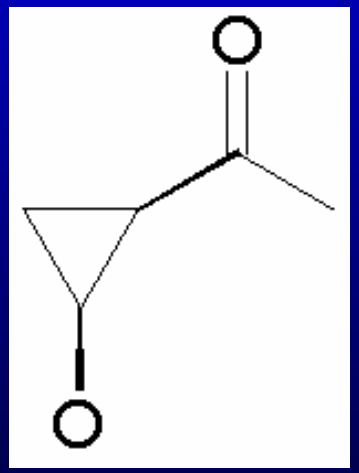
This is the 'Preferred' compound

Extensions to SMILES to hold Structure Related information



Fully resolved molecule (hash bonds)

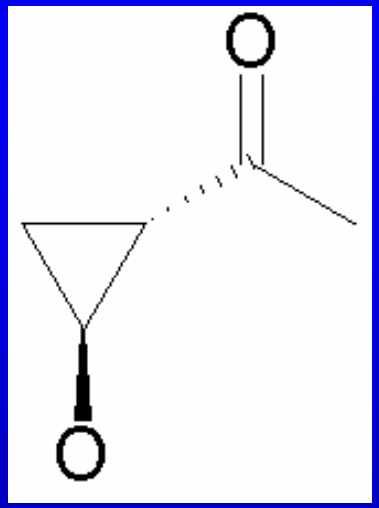
SMILES: **CC(=O)[C@H]1C[C@H]1O**

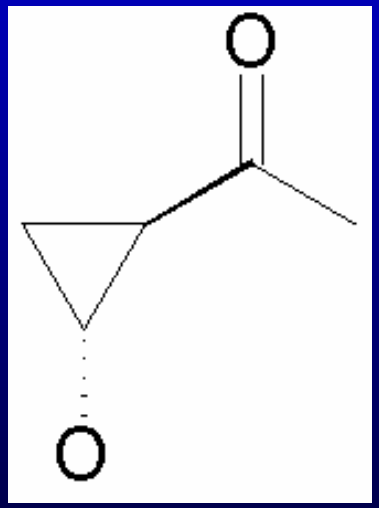


Fully resolved molecule (wedge bonds)

SMILES: **CC(=O)[C@@H]1C[C@@H]1O**

Extensions to SMILES to hold Structure Related information





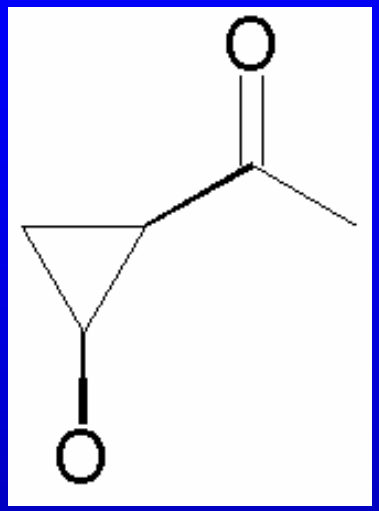
Trans-Relative stereo

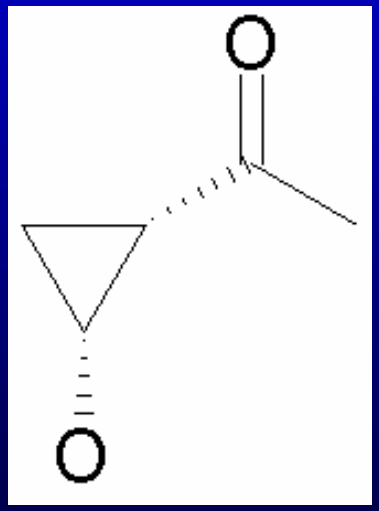
SMILES: **CC(=O)[C#H:1]1C[C##H:1]1O**

(User the '#' symbol ?)

The other enantiomer would generate the same canonical SMILES

Extensions to SMILES to hold Structure Related information





Cis-Relative stereo

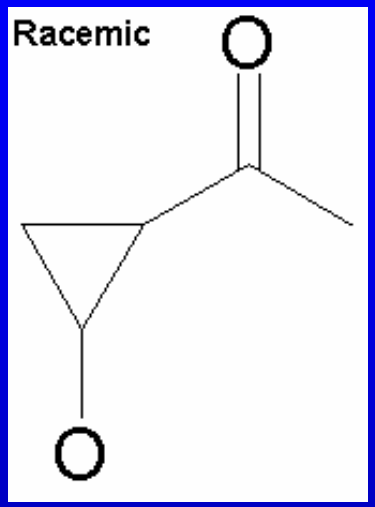
SMILES: **CC(=O)[C#H:1]1C[C#H:1]1O**

The other enantiomer would generate the same canonical SMILES

MODGRAPH

GlaxoSmithKline

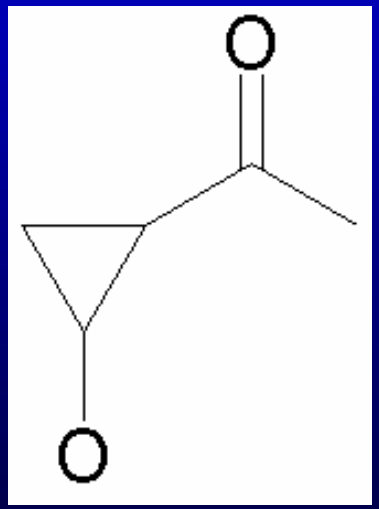
Extensions to SMILES to hold Structure Related information



Racemic mixture of all four enentimers

SMILES: **CC(=O)[C^H]1C[C^H]1O**

(User the '^' symbol?)



Unknown stereo would continue to be represented by the flat canonical SMILES

SMILES: **CC(=O)C1CC1O**

Sulfur and Phosphourus Stereochemistry

- When can we have this please?

Acknowledgements

- John Hollerton, GSK
- John Bradshaw, Daylight
- Daylight team

MODGRAPH

GlaxoSmithKline