

## SCY1102 Unit – 5 Cheminformatics

### Chemistry of Electronic Materials

Introduction: Computer representation of chemical compounds: Line notations - Wiswesser line notation - ROSDAL notation - SMILES coding - Advantages and disadvantages of different types of notations. Standard structure exchange formats: Structure of Mol files and SD files. Chemical structure drawing softwares. Molecule editors: CACTVS molecule editor - Chemdraw - ChemSketch - Chemwindow. Searching chemical structure: Similarity search.

#### Cheminformatics – terminology

Cheminformatics also referred as Chemoinformatics, cheminformatics, chemical informatics, molecular informatics and even chemo-bioinformatics. The term chemoinformatics was probably first introduced in the literature in 1998 by Frank Brown.

Chemoinformatics is the mixing of the information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the areas of chemical, Pharma industries, Agrochemical discover and research organizations.

Chemical informatics is the application of information technology to help chemists investigate new problems and organize, analyze, and understand scientific data in the development of novel compounds, materials, and processes.

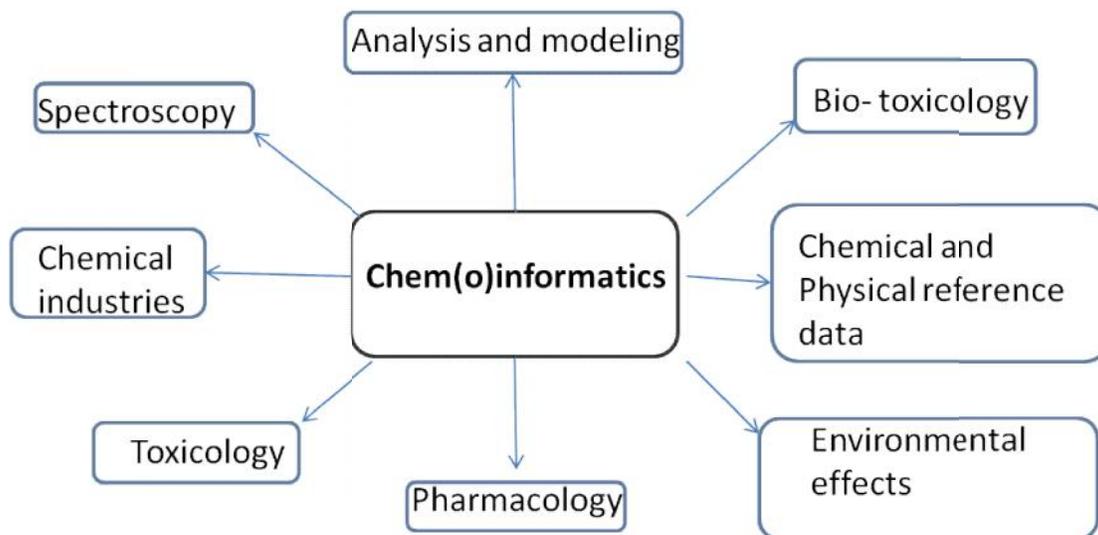
#### Importance of Cheminformatics:

- 1) Advances in theoretical and computational chemistry now allow chemists to model chemical compounds “in silico” with ever-increasing accuracy.
- 2) Molecular properties now becoming accessible through computation include molecular shape, electronic structure, physical properties, chemical reactivity, protein folding, structures of materials and surfaces, catalytic activity, and biochemical activities
- 3) Information Acquisition: Methods for generating and collecting data empirically (experimentation) or from theory (molecular simulation)
- 4) Information Use: Data Analysis, correlation, and application to problems in the chemical and biochemical sciences

#### Representation of chemical compounds

One of the major tasks in chemoinformatics is to represent chemical structures and transfer the various types of representation into application programs. A first basic step to "teaching" computer chemistry is to transform the molecular structure into language amenable to computer representation and manipulation. Basically computers can only handle bits of 0 and 1. Thus

coding is the basis for transferring data. In chemistry the chemical structures have to be represented in machine -readable from by scientific, artificial languages.



Chemical structures are usually stored in a computer as molecular graphs. Graph theory is a well-established area of mathematics that has found application not just in chemistry but in many other areas, such as computer science. A graph is an abstract structure that contains nodes connected by edges. The nodes and edges may have properties associated with them. For example, the atomic number or atom type may be associated with each node and the bond order with each edge. These atom and bond properties are important when performing operations with or upon the molecular graph. A graph represents the topology of a molecule only, that is, the way the nodes (or atoms) are connected.

The 2D graphical representation of chemical structures in structure diagrams can be considered to be the universal "natural language" of chemists. The computational representation of the molecular structures and the creation of structural databases usually follows the natural language. The procedures used to accomplish common tasks such as substructure searching are considered, including the use of graph theoretic methods, bit string screens and the use of canonical structure representations. The focus of on "2D representations" that are concerned with the chemical bonding between atoms rather than their 3D structures of the molecules

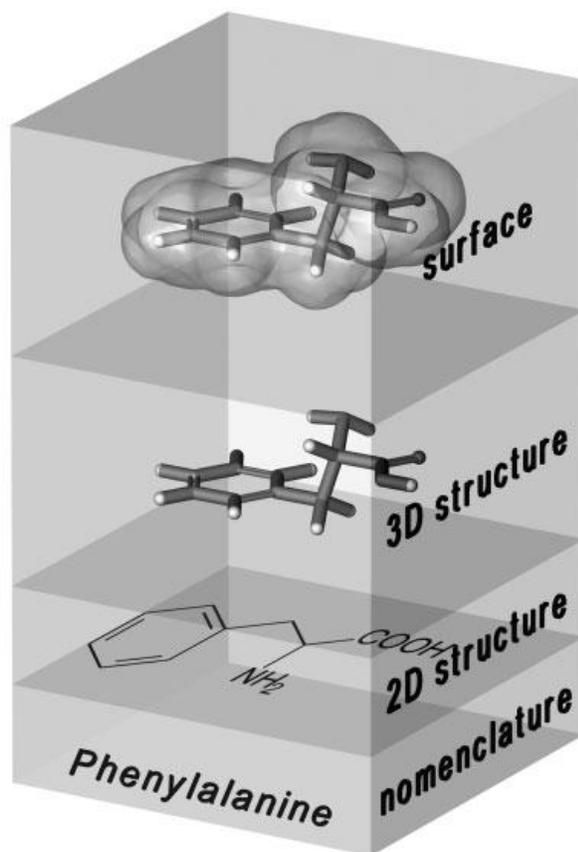


Figure 2: Hierarchical scheme and graphical representations of a Phenylalanine with different contents of structural information.

### Notations and coding

An alternative way to represent and communicate a molecular graph is through the use of a linear notation. The line notations could enter the code of large molecules faster than with a structure-edition program.

### Types of linear notations

- 1) Line notations
- 2) Wiswesser Line Notation(WLN)
- 3) Representation of Organic Structures Description Arranged Linearly (ROSDAL) notation
- 4) Simplified Molecular Input Line Entry Specification (SMILES) notation

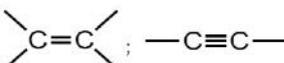
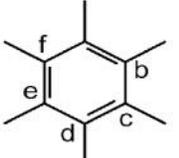
## Line notations

Line notations represent the structure of chemical compounds as a linear sequence of letters and numbers. The IUPAC nomenclature represents such kind of the line notations. However, the IUPAC nomenclature makes an alternative way to represent and communicate a molecular graph is through the use of a linear notation. A linear notation uses alphanumeric characters to encode the molecular structure. Linear notations are more compact than connection tables and so they can be particularly useful for storing and transmitting large numbers of molecules.

### Wiswesser Line Notation (WLN)

The Wiswesser line notation was introduced in 1946, in order to organize and to systematically describe the cornucopia of compounds in more concise manner. A line notation represents a chemical structure by the computer. In many cases the WLN uses the standard symbols from the chemical elements. Additionally, functional groups rings systems, positions of rings substituents, and positions of condensed rings are assigned to individual letters or combinations of symbols.

**Table.1: WLN coding of some important structural units**

<i>Class</i>	<i>Structural unit</i>	<i>WLN coding</i>
Hydrogen	H	H
Alkanes	$C_nH_{2n+2}$	n (e.g., $CH_3CH_2$ ; $CH_2CH_2$ )
Alkenes; alkynes		U; UU
Branched chains		X, Y
Aromatic rings Substituted derivatives		R R B, C, D, E, F
(Hetero)cyclic hydrocarbons		L.n.J; T.n.J L: beginning of a carbocyclic ring; T: beginning of a heterocyclic ring; n: number of atoms of the ring system; J: termination of the ring system
Alkyl halides	-X (X = F, Cl, Br, I)	F, G, E, I
Alcohols; ethers	-OH; -O-	Q; O
Ketones; aldehydes	-CO-; -CO-H	V; VH
Carboxylic acids; esters	-COOH; -CO-O-	VQ; VO

### Rules for WLN Notation:

WLN uses a very simple system of canonicalization based on alphanumeric order.

- 1) Capital Letter: A-Z are used for elements, atom groups, branches and ring positions
- 2) Numbers: 0-9 indicate the length of an alkyl chain or the ring number
- 3) Special characters: "&", "/", "-", and "" (blank) indicate rings and substitution positions.
- 4) Priority increases in the direction:
  - a. Symbols;
  - b. Numbers in numerical order; and
  - c. Letters in alphabetical order (with the exception of R which has lower priority than symbols).Coding generally begins at the substituent assigned the highest priority.

### Examples WLN Notation

WLN for CH<sub>4</sub> (Methane) : 1H

WLN for CH<sub>3</sub>-CH<sub>3</sub> (Ethane): 2H

WLN for CH<sub>3</sub>-CH<sub>2</sub>-CH<sub>3</sub> (Propane) :3H

WLN for C<sub>7</sub>HCl<sub>5</sub>O<sub>2</sub> (Pentachlorbenzoate) : QVR BG CG DG EG FG

### Acetone:



1V1

The two "1"s stand for saturated one-carbon chains, i.e. methyl groups. The "V" stands for a carbon doubly-bonded to oxygen.

### Diethyl Ether:

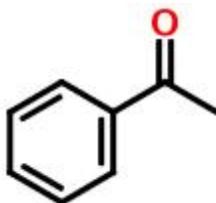


2O2

Given nothing more than the above example, the encoding of diethyl ether should be completely clear: "O" simply stands for a divalent oxygen atom.

### Acetophenone:

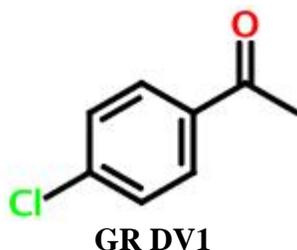
The benzene ring is one of the most ubiquitous functional groups in organic chemistry. Wiswesser knew this and wanted to make it easy to encode aromatic compounds. His solution is simplicity itself. Consider acetophenone:



1VR

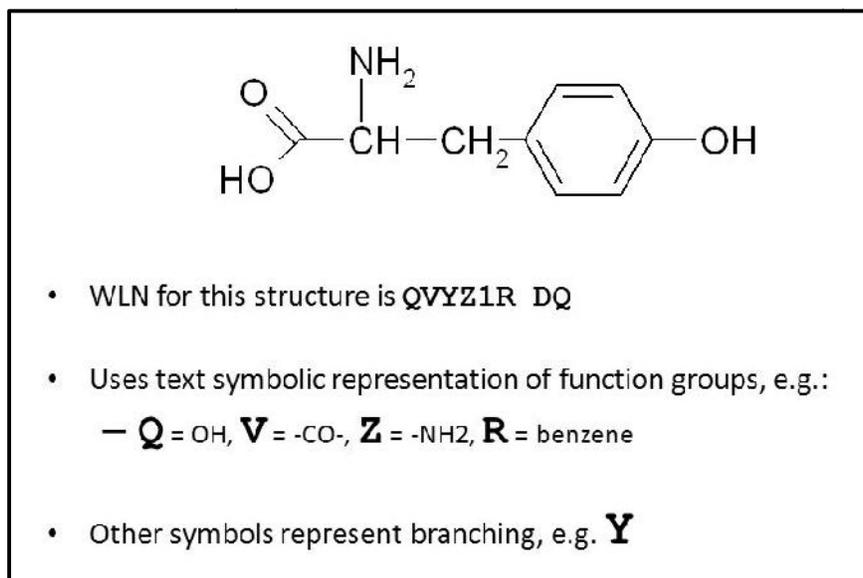
The "R" stands for a benzene ring. WLN canonicalization gives it the lowest priority and this is why it appears last.

#### 4-chloroacetophenone:



The "G" symbol stands for chlorine. The " DV1" stands for the 4-acyl substituent. Here, the "D" denotes the 4-position. The 3- position would result in " CV1", and the 2- position would give " BV1". The space character means that the character following it should be interpreted as ring locant.

#### Phenyl alanine



#### Advantages:

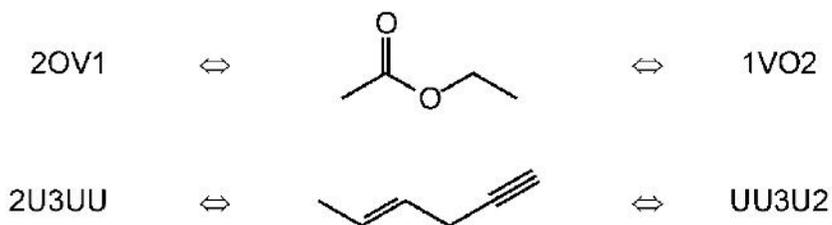
- 1) WLN is remarkably compact, especially when compared to SMILES and InChI.  
WLN for 4-chloroacetophenone : GRDV1  
InChI for 4-chloroacetophenone: InChI=1/C8H7ClO/c1-6(10)7-2-4-8(9)5-3-7/h2-5H,1H3
- 2) The functional group recognition is easy for humans, it's orders of magnitude easier for Machines.

- 3) WLN is concise linear code
- 4) Unambiguous, Simple substructure search
- 5) Includes stereochemistry and it is unique if rules are followed.

### Disadvantages:

- 1) Encoding WLN rules into a computer programme is difficult, and the rules for the canonicalization were computationally intractable.

Example:



- 2) Large number of complex rules and coding prone to errors
- 3) difficult to translate into a connection table and No support for coding reaction
- 4) Only those substructures contained in the coding can be retrieved in a substructure search

### Applications

The WLN was applied to indexing the chemical structure index (CSI) at the institute for scientific information (ISI) and the index chemicus registry systems (ICRS) as well as the crossbow systems of imperial chemical industries (ICI). With the introduction of commotion tables in the chemical abstracts service (CAS) in 1965 and the advent of molecular editors in the 1970S, which directly produced connection tables, the WLN lost its importance.

### ROSDAL

ROSDAL (Representation of organic structure description arranged linearly) syntax was developed by S, Welford, J.Barnard and M.F.Lynch n 1985 for the beilsten institute. This line notation was intended to transmit structure information between the user and the beilsten was DIALOG system (Beilsten-Online) during database retrieval queries and structure displays. This exchange of structure information by the ROSDAL ASCII character string is very fast.

### ROSDAL syntax is characterized by

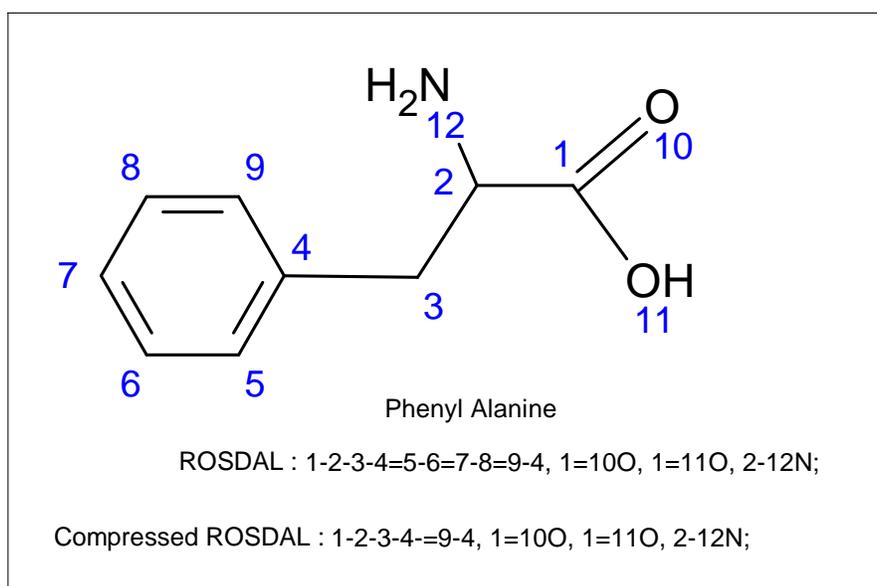
- a) A simple coding of a chemical structure using alphanumeric symbols which can easily b learned by a chemist.

- b) In the Linear structure representation, each atom of the structure is arbitrarily assigned a unique number, except for the hydrogen atoms.
- c) Carbon atoms are shown in the notation only by digits.
- d) The other types of atoms carry, in addition their atomic symbols.
- e) In order to describe the bonds between atoms, bond symbols are inserted between the atoms numbers.
- f) Branches are marked and separated from the other parts of the code by commas. The ROSDAL linear notation is unambiguous but not unique.

**The sequence for setting up a ROSDAL notation is**

- 1) The structure diagram is drawn and the atoms are arbitrarily numbered (each atom is assigned a unique number).
- 2) Atomic symbols are usually written directly behind the index of an atom.
- 3) Usually only the indices of the carbon atoms are written, not the symbols: hydrogen atoms can have, but do not need, an atom number.
- 4) Bond types are described as follows:
  - a. “-“ for a single bond
  - b. “=” for a double bond
  - c. “#” for triple bond
  - d. “?” for any connection
- 5) Simplifications are allowed, such as writing alternating bonds as “-=”.
- 6) Commas separate branches and substituents.
  - e. Examples:

**Phenylalanine:**



## ROSDAL Advantages and disadvantages

### Advantages:

- 1) Simple code, easy to learn
- 2) Fast data exchange format
- 3) Includes stereochemistry

### Disadvantages

- 1) No support for coding reactions
- 2) Not Unique

## SMILES Coding

In 1986 David Weininger the SMILES notation at the US Environmental Research Laboratory USEPA, Duluth, MN, for chemical data processing. The Chemical structure information is highly compressed and simplified in the notation. The flexible, easy to learn language describes chemical structure as a line notation. The SMILES language has found widespread distribution as a universal chemical nomenclature.

SMILES Definition: The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings.

The basic SMILES rules are:

### 1) Atoms represented by their symbols.

Atoms are represented by their atomic symbols: this is the only required use of letters in SMILES. Each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets, [ ].

[H <sup>+</sup> ]	proton
[Fe <sup>+2</sup> ]	iron (II) cation
[OH <sup>-</sup> ]	hydroxyl anion
[Fe <sup>++</sup> ]	iron (II) cation
[OH <sub>3</sub> <sup>+</sup> ]	hydronium cation
[NH <sub>4</sub> <sup>+</sup> ]	ammonium cation

Atoms in aromatic rings are specified by lower case letters, e.g., aliphatic carbon is represented by the capital letter C, aromatic carbon by lower case c. Since attached hydrogens are implied in the absence of brackets, the following atomic symbols are valid SMILES notations.

**2) Hydrogen atoms automatically structure free valences and are omitted (Simple hydrogen connection)**

The second letter of two-character symbols must be entered in lower case. Elements in the "organic subset" B, C, N, O, P, S, F, Cl, Br, and I may be written without brackets if the number of attached hydrogens conforms to the lowest normal valence consistent with explicit bonds. "Lowest normal valences" are B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6), and 1 for the halogens.

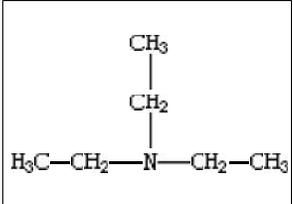
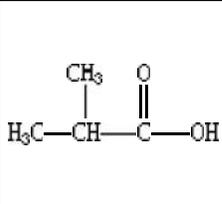
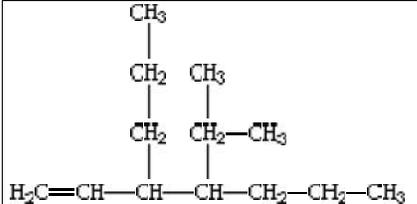
C	Methane	CH <sub>4</sub>
P	Phosphine	PH <sub>3</sub>
N	Ammonia	NH <sub>3</sub>
S	Hydrogen Sulfide	H <sub>2</sub> S
O	Water	H <sub>2</sub> O
Cl	Hydrochloric acid	HCl

**3) Neighboring atoms stand next to each other.**

**4) Double and triple bonds are characterized by "=" and "#" respectively.**

CC	ethane	(CH <sub>3</sub> CH <sub>3</sub> )
C=O	formaldehyde	(CH <sub>2</sub> O)
C=C	ethene	(CH <sub>2</sub> =CH <sub>2</sub> )
O=C=O	carbon dioxide	(CO <sub>2</sub> )
COC	dimethyl ether	(CH <sub>3</sub> OCH <sub>3</sub> )
C#N	hydrogen cyanide	(HCN)
CCO	ethanol	(CH <sub>3</sub> CH <sub>2</sub> OH)
[H][H]	molecular hydrogen	(H <sub>2</sub> )

**5) Branches are represented by parentheses.**

		
CCN(CC)CC	CC(C)C(=O)O	C=CC(CCC)C(C(C)C)CCC
Triethylamine	Isobutyric acid	3-propyl-4-isopropyl-1-heptene

**6) Rings are described by allocating digits to the two "connecting: rings atoms**

Example:

C1CCCCC1      Cyclohexane (C<sub>6</sub>H<sub>12</sub>)

c1ccccc1      Benzene (C<sub>6</sub>H<sub>6</sub>)

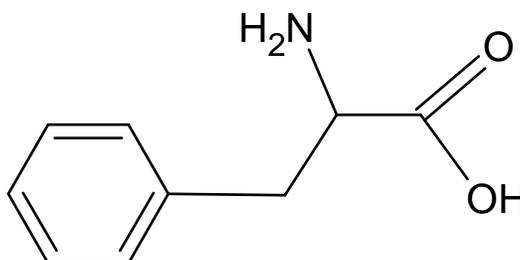
## SMILES Representation of Organic Compounds

C	Methane (CH <sub>4</sub> )
CC	Ethane (CH <sub>3</sub> CH <sub>3</sub> )
C=C	Ethene (CH <sub>2</sub> CH <sub>2</sub> )
C#C	Ethyne (CHCH)
COC	Dimethyl ether (CH <sub>3</sub> OCH <sub>3</sub> )
CCO	Ethanol (CH <sub>3</sub> CH <sub>2</sub> OH)
CC=O	Acetaldehyde (CH <sub>3</sub> -CH=O)
C#N	Hydrogen Cyanide (HCN)
[C-]#N	Cyanide anion

Branches are specified by enclosures in parentheses and can be nested or stacked, as shown in these examples.

CC(C)CO	Isobutyl alcohol (CH <sub>3</sub> -CH(CH <sub>3</sub> )-CH <sub>2</sub> -OH)
CC(CCC(=O)N)CN	5-amino-4-methylpentanamide

Example:



SMILES: NC(Cc1ccccc1)C(=O)O

## SMILES Advantages and Disadvantages

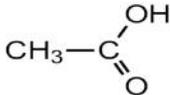
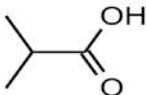
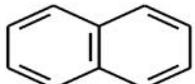
### Advantages:

- 1) Simplest linear code and easy to learn
- 2) Fast data exchange format
- 3) Supports Markush, stereochemistry and reaction coding
- 4) Unambiguous

### Disadvantages:

- 1). Not unique (except Unique SMILES)
- 2). Some problems with aromaticity perception

**Table 2-2.** SMILES syntax.

<i>SMILES code</i>	<i>Chemical structure</i>	<i>Compound name</i>
<i>Atoms:</i> Atoms are represented by their atomic symbols. Ambiguous two-letter symbols (e.g., Nb is not NB) have to be written in square brackets. Otherwise, no further letters are used. Free valences are saturated with hydrogen atoms.		
C	CH <sub>4</sub>	methane
[Fe+ 2] or [Fe++]	Fe <sup>2+</sup>	iron (II) cation
<i>Bonds:</i> Single, double, triple, and aromatic (or conjugated) bonds are indicated by the symbols “ - ”, “ = ”, “ # ” and “ : ”, respectively; single and aromatic bonds should be omitted.		
C=C	H <sub>2</sub> C=CH <sub>2</sub>	ethene
O=CO	HCOOH	formic acid
<i>Disconnected structures in the molecule:</i> Individual parts of the compound are separated by a period. The period indicates that there is no connection between atoms or parts of a molecule. The arrangement of the parts is arbitrary.		
[Na+].[OH-]	NaOH	sodium hydroxide
<i>Branches:</i> Branches are indicated within parentheses.		
CC(=O)O		acetic acid
CC(C)C(=O)O		isobutyric acid
<i>Cyclic structures:</i> Rings are described by breaking the ring between two atoms and then labeling the two atoms with the same number.		
C1CCCCC1		cyclohexane
<i>Aromaticity:</i> Aromatic structures are indicated by writing all the atoms involved in lower-case letters.		
o1cccc1		furan
c12c(ccc1)ccc2 same as c1cc2ccccc2cc1		naphthalene

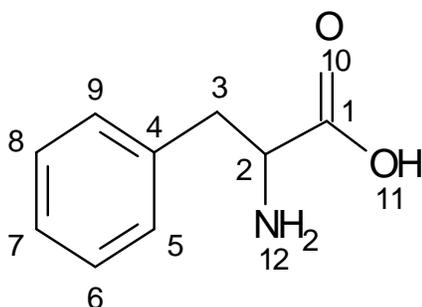
## Application

The compact textual coding requires no graphical input and additionally permits a fast transmission. These are important advantages of using SMILES in chemical applications via the internet and in online services. SMILES are also for the input of structure in the daylight toolkit.

## Limitations:

- SMILES is proprietary and it is not an open project. This has led different chemical software developers to use different SMILES generation algorithms, resulting in different SMILES versions for the same compound.
- SMILES strings obtained from different databases or research groups are not interchangeable unless they used the same software to generate the SMILES strings.
- With an aim to address this interchangeability issue of SMILES, an open-source project has launched to develop an open, standard version of the SMILES language called OpenSMILES.28 .

## WLN, SMILES, ROSDAL Notation of Phenyl alanine



IUPAC: 2-amino-3-phenylpropanoic acid

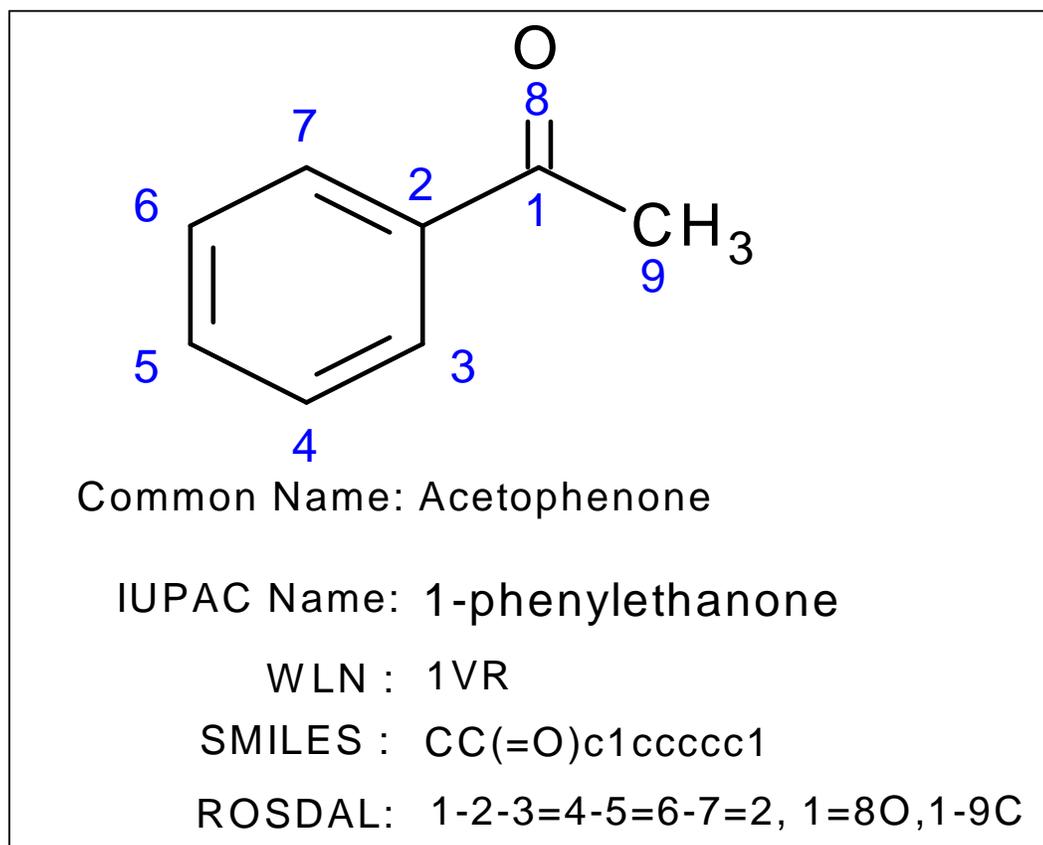
Systematic Name: Phenyl alanine

WLN: VQYZ1R

SMILES: NC(Cc1ccccc1)C(=O)O

ROSDAL: 1-2-3-4-5=6-7=8-9=4, 1=10O,1-11O, 2-12N,

## WLN, SMILES, ROSDAL Notation of Acetophenone



### Structure exchange formats

The chemistry of the numerous software programs is available to handle structure information on molecules. The scope of the programs leads from drawing structure diagrams to expert systems that process the data and produce new information. All of these systems have one task in common: to save data in file. Many organizations and software's suppliers have developed their own connection table format and quite a few have made provisions for their import or export of other files format. The processing of data from data to information and finally to knowledge, usually asks for the interaction and cooperation of several different software systems and databases. In this process, the exchange of chemical structure information plays a pivotal role : the internal file formats of one software systems has to be understood by another, i.e converted into its internal file format. This exchange process is usually handled through an external ASCII, file format.

Format type	Codename of the format
Document formats	mrv, Marvin Documents (MRV) cdx, cdxml, ISIS/Draw sketch file (SKC) skc, ChemDraw sketch file (CDX, CDXML)
Molecule file formats	mol, rgf, sdf, rdf, csmol, csrgf, cssdf, csrdf, cml, smiles, cxsmiles, abbrevgroup, peptide, sybyl, mol2, pdb, xyz, inchi, name
Graphics formats	jpeg, msbmp, png, pov, svg, emf, tiff, eps
Compression and Encoding	gzip, base64

Table: The most important file formats for exchange of chemical structure information

<i>File format</i>	<i>Suffix</i>	<i>Comments</i>	<i>Support</i>
MDL Molfile	*.mol	Molfile; the most widely used connection table format	<a href="http://www.mdli.com">www.mdli.com</a>
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds	<a href="http://www.mdli.com">www.mdli.com</a>
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions	<a href="http://www.mdli.com">www.mdli.com</a>
SMILES	*.smi	SMILES; the most widely used linear code and file format	<a href="http://www.daylight.com">www.daylight.com</a>
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides	<a href="http://www.rcsb.org">www.rcsb.org</a>
CIF	*.cif	Crystallographic Information File format; for 3D structure infor- mation on organic molecules	<a href="http://www.iucr.org/iucr-top/cif/">www.iucr.org/iucr-top/cif/</a>
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format	<a href="http://www.jcamp.org/">www.jcamp.org/</a>
CML	*.cml	Chemical Markup Language; extension of XML with speciali- zation in chemistry	<a href="http://www.xml-cml.org">www.xml-cml.org</a>

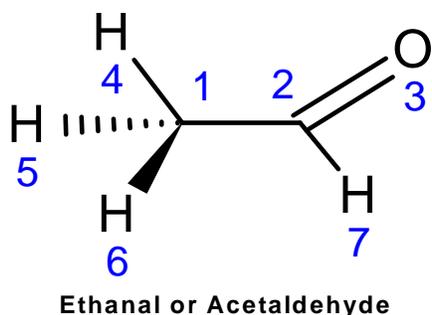
There are many file formats for storing information about molecular structures. Only some of them have widely accepted by the cheminformatics community and are used as standard format for the exchange of information of chemical structures and reactions.

The Molfile and SDFfile formats were developed by Molecular Design Limited (MDL website). These formats became popular for the storage and exchange of information on molecular structures and properties.

Molfile: It describes a single molecular structure which can contain disjointed fragments. The default filename extension for Molfiles is “.mol”.

In order to understand the Molfile format let us look at a sample file and recognize its fundamental structure. For simplicity, some less important details will be omitted in the discussion.

### Molfile Example: Ethanal or Acetaldehyde



The header consists of up to 3 lines:

**Line:1** Comment or molecule name line;

**Line:2** A line detailing the type of application which generated the table, the date and time of its creation.

**04131617572D**

**04<sup>th</sup>Month/13<sup>th</sup> date/16 th year/1757 Time of created/2D structure**

**Line:3** Comment or description line that even if empty must be present.

**Line:4** These lines are followed by the counts line which gives the number of atoms and bonds (plus several additional flags) in the molecule.

### Line: 5 to 11

The atom block, containing the x, y, and z co-ordinates of the atom and the element symbol for the atom. As this study used only topological (rather than topographical) structural information the positional coordinates were not used.

### Line: 12 to 17

Finally, for the bond block, the first two numbers of each line indicate the numbers of the respective atoms from the atom list, and the third number indicates the bond type between these two atoms. (A value of 1 indicates a single bond connection, 2 a double bond and 3 a triple bond.)

Line: 18 to 24

Line: 25 END

```
1. Blank Line or Molecule Name
2. ACD/Labs04131617572D [Creator Name]
3. Blank Line
4. 7 6 0 0 0 0 0 0 0 0 0 8 V2000
5. -0.8077 -0.0155 0.1782 C 0 0 0 0 0 0 0 0 0 0 0 0
6. 0.6883 -0.0351 -0.0110 C 0 0 0 0 0 0 0 0 0 0 0 0
7. 1.2933 1.0026 -0.1570 O 0 0 0 0 0 0 0 0 0 0 0 0
8. -1.1027 0.8443 0.8174 H 0 0 0 0 0 0 0 0 0 0 0 0
9. -1.2933 0.0808 -0.8174 H 0 0 0 0 0 0 0 0 0 0 0 0
10. -1.1471 -0.9553 0.6649 H 0 0 0 0 0 0 0 0 0 0 0 0
11. 1.2343 -1.0026 -0.0232 H 0 0 0 0 0 0 0 0 0 0 0 0
12. 1 2 1 0 0 0 0
13. 1 4 1 0 0 0 0
14. 1 5 1 0 0 0 0
15. 1 6 1 0 0 0 0
16. 2 3 2 0 0 0 0
17. 2 7 1 0 0 0 0
18. M ZC 1 1
19. M ZC 2 2
20. M ZC 3 3
21. M ZC 4 4
22. M ZC 5 5
23. M ZC 6 6
24. M ZC 7 7
25. M END
```

Header block (Line:1 to 3)

Counts Line (Line:4)

Atom Block (Line:5 to 11)

Bond block (Line:12 to 17)

Stereochemistry (Line:18 to 24)

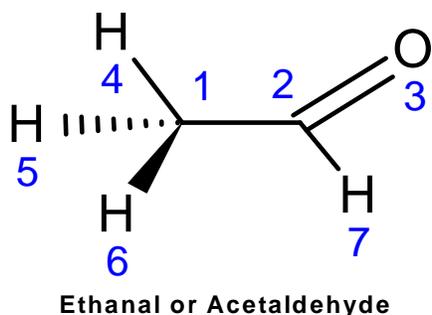
Fig. Schematic Representation of Molfile for Acetaldehyde molecule or Ethanal molecule

### SDfile Format:

SDfile contains the structural information and non structural information. This makes it useful for exchange of data between databases and between two computational softwares. Within an SDfile, each molecule is represented by its Molfile with additional data items describing its non-

structural properties molecular weight, heat of formation, molecular descriptors, biological activity, CAS number etc.)

In order to understand the SDfile format let us look at a Acetaldehyde or Ethanal molecule.



The header consists of up to 3 lines:

**Line:1** Comment or molecule name line;

**Line:2** A line detailing the type of application which generated the table, the date and time of its creation.

**07011509283D**

**07<sup>th</sup> Month/1<sup>st</sup> date/15<sup>th</sup> year/09.28 time/ 3D file**

**Line:3** Comment on copy right

**Line:4** These lines are followed by the counts line which gives the number of atoms and bonds (plus several additional flags) in the molecule.

**Line: 5 to 11**

The atom block, containing the x, y, and z co-ordinates of the atom and the element symbol for the atom. As this study used only topological (rather than topographical) structural information the positional coordinates were not used.

**Line: 12 to 17**

Finally, for the bond block, the first two numbers of each line indicates the numbers of the respective atoms from the atom list, and the third number indicates the bond type between these two atoms. (A value of 1 indicates a single bond connection, 2 a double bond and 3 a triple bond.)

**Line: 18** End for structural data.

**Line 19 to 20:** Header block for non structural data

**Line 21 to 47:** Nonstructural data like, date of creation, CAS number, Method of generation, Dipole moment, Electronic Energy, Rotational constants, Software used for generation and Contributor name.

**Line: 48:** Delimiter.

Line No:

```
1. Acetaldehyde or Ethanal
2. NIST 07011509283D 1 1.00000 -153.83012
3. Copyright by the U.S. Sec. Commerce on behalf of U.S.A. All rights reserved.
4. 7 6 0 0 0 0 0 0 0 0 0999 V2000
5. 0.5984 0.8354 0.7035 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6. 2.0205 1.1102 1.1250 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7. 2.5939 2.1646 0.9686 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8. 0.0019 0.5440 1.5781 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9. 0.1623 1.7187 0.2315 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10. 0.5736 -0.0140 0.0082 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11. 2.5387 0.2513 1.6103 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12. 1 2 1 0 0 0 0
13. 1 4 1 0 0 0 0
14. 1 5 1 0 0 0 0
15. 1 6 1 0 0 0 0
16. 2 3 2 0 0 0 0
17. 2 7 1 0 0 0 0
18. M END
19. > <COPYRIGHT>
20. Collection (C) 2016 copyright by the U.S. Secretary of Commerce on behalf of the United States of America. All rights reserved.
21.
22. > <DATE>
23. 2015-07-01
24.
25. > <CAS.NUMBER>
26. 75-07-0
27.
28. > <METHOD>
29. B3LYP/6-31G*
30.
31. > <DIPOLE.MOMENT>
32. 2.6401 debye
33.
34. > <ELECTRONIC.ENERGY>
35. -153.830121544 hartree
36.
37. > <ROTATIONAL.CONSTANTS>
38. 56.86315 GHz
39. 10.07531 GHz
40. 9.03745 GHz
41.
42. > <SOFTWARE>
43. Gaussian 09, Revision D.01
44.
45. > <CONTRIBUTOR>
46. Avi Newman
47.
48. $$$
```

Line 1 to 3: Header block

Line 4: Counts line

Line 5 to 11: Atom block

Line 12 to 17: Bond block

Line 18: End

Line 21 to 48: Non structural data

Fig. Schematic Representation of SDfile for Acetaldehyde molecule or Ethanal molecule

## Chemical structure drawing soft wares

Chemical structure drawing soft wares aids in teaching key chemistry concepts to high school, undergraduate, and graduate chemistry students and researchers at higher education. In addition, students benefit from exposure in the learning environment to the same tools they will encounter in the workforce

Types of the Structure drawing tools:

**a) ChemDraw:**

Molecule editor developed by the cheminformatics company CambridgeSoft. For Windows and Mac.

**b) ACD/ChemSketch:**

Molecule editor developed by ACD/Labs. Also available as freeware, with tools for 2D structure cleaning, 3D optimization and viewing, InChI generation and conversion, drawing of polymers, organometallics, and Markush structures. For Windows only.

**c) ChemWindow:**

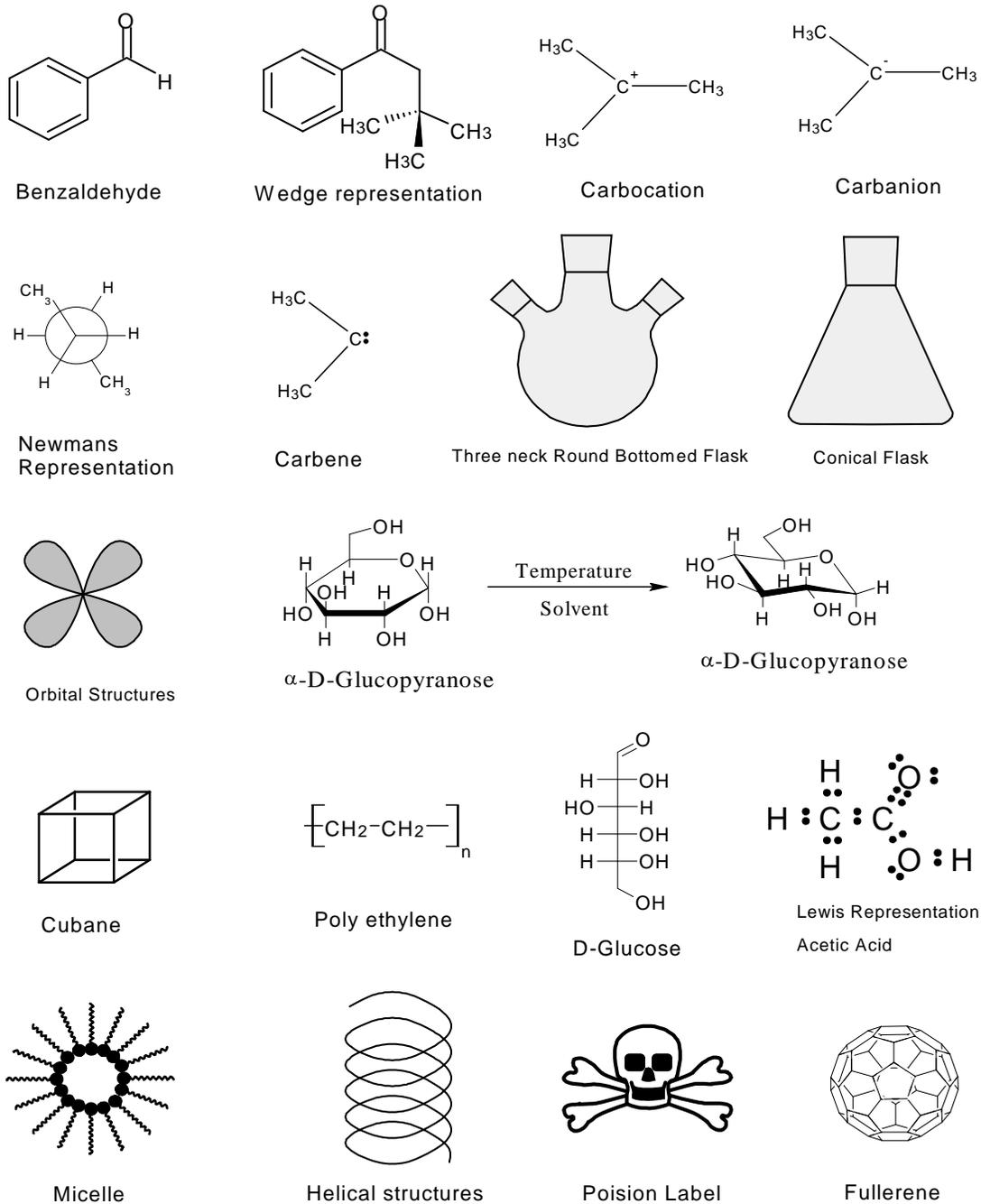
It is the software chemists choose for chemical structure drawing and publishing worldwide. Now with integrated solutions to modify, store, search, and retrieve chemical structures and properties, ChemWindow offers scientists even more solutions.

- i. 2D Chemical Structure Drawing
- ii. 3D Molecular Rendering
- iii. Chemistry Database Building (structures, chemical property information, etc.)
- iv. Scientific Publishing / Reporting Tools

**d) CACTV/JME Molecular Editor:**

The CACTVS molecule editor is a graphical mouse-oriented X11 Unix-based tool for the input of molecular structures. It can be used both in stand-alone mode, optionally with Drag&Drop to and from other stand-alone programs from the CACTVS tool series.

Java applet which allows to draw / edit molecules and reactions (including generation of substructure queries) and to depict molecules directly within an HTML page. Editor can generate Daylight SMILES or MDL Molfile of created structures.



**Fig.** Schematic representation of possible representation of chemical structures by using the molecular editor softwares (Chemsketch, Chemdraw, ChemWindo Softwares)

## UNIT 5: CHEMINFORMATICS

### Part – A Questions

#### 1. Define Cheminformatics.

Chemical informatics is the application of information technology to help chemists investigate new problems and organize, analyze, and understand scientific data in the development of novel compounds, materials, and processes.

#### 2. Define line notations. Mention different types of line notations.

Line notations represent the structure of chemical compounds as a linear sequence of letters and numbers. The IUPAC nomenclature represents such kind of the line notations. However, the IUPAC nomenclature makes an alternative way to represent and communicate a molecular graph is through the use of a linear notation. A linear notation uses alphanumeric characters to encode the molecular structure. Linear notations are more compact than connection tables and so they can be particularly useful for storing and transmitting large numbers of molecules.

Types of linear notations

- 1) Line notations
- 2) Wiswesser Line Notation (WLN)
- 3) Representation of Organic Structures Description Arranged Linearly (ROSDAL) notation
- 4) Simplified Molecular Input Line Entry Specification (SMILES) notation

#### 3. What are the uses of line notations?

#### 4. Define Wiswesser line notations?

#### 5. What are the string characters used for WLN?

#### 6. Create the WLN for (i) Acetone and (ii) Diethyl ether.

**Acetone:**

The two "1"s stand for saturated one-carbon chains, i.e. methyl groups. The "V" stands for a carbon doubly-bonded to oxygen.

**Diethyl Ether:**

Given nothing more than the above example, the encoding of diethyl ether should be completely clear: "O" simply stands for a divalent oxygen atom.

7. What are the advantages and disadvantages of Wiswesser line notation?

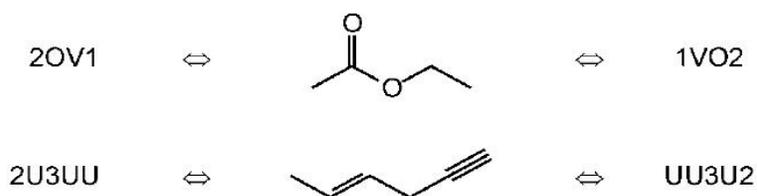
**Advantages:**

- 1) WLN is remarkably compact, especially when compared to SMILES and InChI.
  - WLN for 4-chloroacetophenone : GRDV1
  - InChI for 4-chloroacetophenone: InChI=1/C8H7ClO/c1-6(10)7-2-4-8(9)5-3-7/h2- 5H,1H3
- 2) The functional group recognition is easy for humans, it's orders of magnitude easier for machines

**Disdvantages:**

- 1) Encoding WLN rules into a computer programme is difficult, and the rules for the canonicalization were computationally intractable.

Example:



8. What is SMILES?

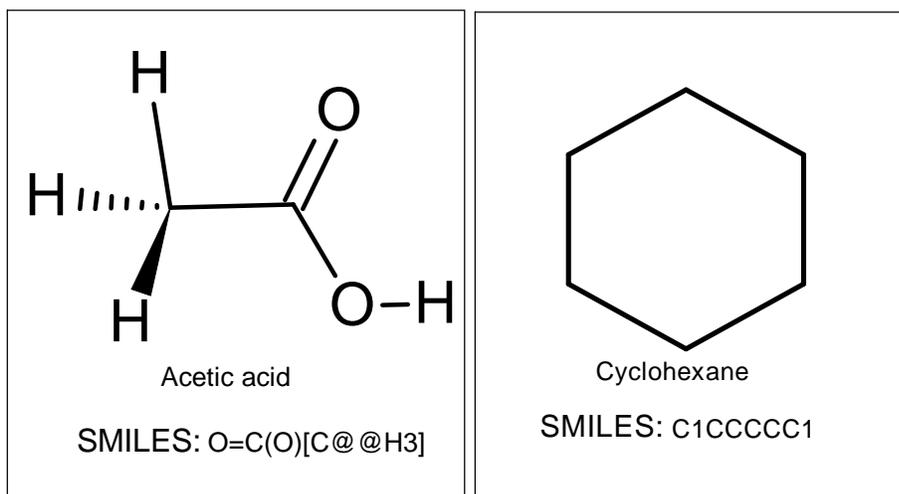
Ans: The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings.

9. Write the common rules for the generation of SMILES string for organic compounds.

Ans: The basic SMILES rules are:

- Atoms are represented by their atomic symbols
- Hydrogen atoms automatically saturate free valences and are omitted (simple hydrogen connection).
- Neighboring atoms stand next to each other
- d) Double and triple bonds are characterized by “=” and “#”, respectively.
- Branches are represented by parentheses.
- Rings are described by allocating digits to the two “connecting” ring atoms.

10. Generate SMILE string for molecules (i) Acetic acid (ii) Cyclohexane.



11. What are the advantages and disadvantages of SMILES coding?

**Advantages:**

- Simplest linear code and easy to learn
- Fast data exchange format

- c) Supports Markush, stereochemistry and reaction coding
- d) Unambiguous

**Disadvantages:**

- a) Not unique (except Unique SMILES)
- b) Some problems with aromaticity perception

12. Write the link for getting detailed information on the SMILES code.

13. What is ROSDAL syntax?

14. Write the ROSDAL string characters for a single bond, a double bond and a triple bond.

**Ans:** Bond types are described as follows:

“-“ for a single bond

“=” for a double bond

“#” for triple bond

“?” for any connection

15. Mention the application of ROSDAL notation.

16. What are the advantages and disadvantages of ROSDAL syntax?

**Advantages:**

- 1) Simple code, easy to learn
- 2) Fast data exchange format
- 3) Includes stereochemistry

**Disadvantages**

- 1) No support for coding reactions
- 2) Not Unique

17. Mention any four standard chemical structure file formats.

<b>File format</b>	<b>Suffix</b>	<b>Comments</b>	<b>Support</b>
MDL Molfile	*.mol	Molfile; the most widely used connection table format	<a href="http://www.mdli.com">www.mdli.com</a>
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds	<a href="http://www.mdli.com">www.mdli.com</a>
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions	<a href="http://www.mdli.com">www.mdli.com</a>
SMILES	*.smi	SMILES; the most widely used linear code and file format	<a href="http://www.daylight.com">www.daylight.com</a>
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides	<a href="http://www.rcsb.org">www.rcsb.org</a>
CIF	*.cif	Crystallographic Information File format; for 3D structure information on organic molecules	<a href="http://www.iucr.org/iucr-top/cif/">www.iucr.org/iucr-top/cif/</a>
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format	<a href="http://www.jcamp.org/">www.jcamp.org/</a>
CML	*.cml	Chemical Markup Language; extension of XML with specialization in chemistry	<a href="http://www.xml-cml.org">www.xml-cml.org</a>

18. Define pdf

file format.

19. Mention the support link with file name extension of MDL Molfile and SDfile.

20. Name any two chemical structure drawing package molecule editors.

A molecule editor is a computer program for creating and modifying representations of chemical structures.

- a) ACD/ChemSketch Software developed by ACD/Labs. A chemically intelligent drawing interface that allows you to draw almost any chemical structures. Freeware version available

- b) ChemDraw software was developed by CambridgeSoft. It is used to draw chemical structures and reactions.
- c) ChemWindo software was developed by BioRad company. It is used to draw chemical structure. It is a free ware for academic research and teaching.

21. Write the support link for Chemsketech and Chemwindow tutorials.

<http://www.acdlabs.com>

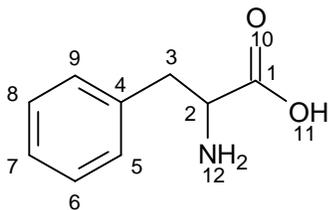
<http://www.cambridgesoft.com/software/overview.aspx>

<http://www.bio-rad.com/en-fr/spectroscopy>

22. What is meant by similarity search?

### **Part – B Questions**

1. Write the IUPAC name, WLN, SMILES and ROSDAL notation for phenylalanline



IUPAC: 2-amino-3-phenylpropanoic acid

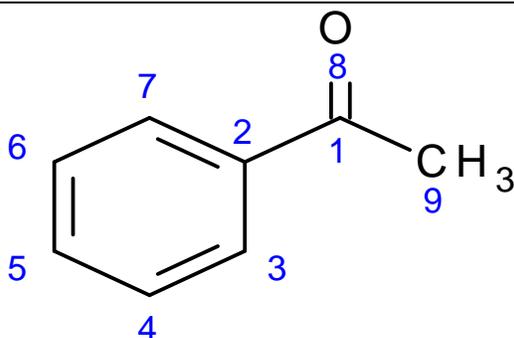
Systematic Name: Phenyl alanine

WLN: VQYZ1R

SMILES: NC(Cc1ccccc1)C(=O)O

ROSDAL: 1-2-3-4-5=6-7=8-9=4, 1=10O,1-11O, 2-12N,

2. Write the IUPAC name, WLN, SMILES and ROSDAL notation for acetophenone.



Common Name: Acetophenone

IUPAC Name: 1-phenylethanone

WLN : 1VR

SMILES : CC(=O)c1ccccc1

ROSDAL: 1-2-3=4-5=6-7=2, 1=8O,1-9C

3. Explain WLN coding for at least five structural units.

WLN for CH<sub>4</sub> (Methane) : 1H

WLN for CH<sub>3</sub>-CH<sub>3</sub> (Ethane): 2H

WLN for CH<sub>3</sub>-CH<sub>2</sub>-CH<sub>3</sub> (Propane) :3H

WLN for C<sub>7</sub>HCl<sub>5</sub>O<sub>2</sub> (Pentachlorobenzoate) : QVR BG CG DG EG FG

**Acetone:**



The two "1"s stand for saturated one-carbon chains, i.e. methyl groups. The "V" stands for a carbon doubly-bonded to oxygen.

**Diethyl Ether:**

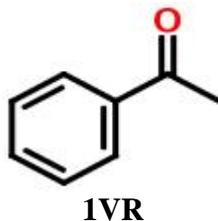


Given nothing more than the above example, the encoding of diethyl ether should be completely clear:

"O" simply stands for a divalent oxygen atom.

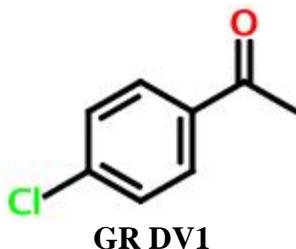
**Acetophenone:**

The benzene ring is one of the most ubiquitous functional groups in organic chemistry. Wiswesser knew this and wanted to make it easy to encode aromatic compounds. His solution is simplicity itself. Consider acetophenone:



The "R" stands for a benzene ring. WLN canonicalization gives it the lowest priority and this is why it appears last.

#### 4-chloroacetophenone:



4. Explain in detail the basic rules of SMILES syntax.

5. Write the sequence for setting up a ROSDAL notation.

ROSDAL (Representation of organic structure description arranged linearly) syntax was developed by S. Welford, J. Barnard and M.F. Lynch in 1985 for the Beilstein Institute. This line notation was intended to transmit structure information between the user and the Beilstein was DIALOG system (Beilstein-Online) during database retrieval queries and structure displays. This exchange of structure information by the ROSDAL ASCII character string is very fast.

#### **ROSDAL syntax is characterized by**

- g) A simple coding of a chemical structure using alphanumeric symbols which can easily be learned by a chemist.
- h) In the Linear structure representation, each atom of the structure is arbitrarily assigned a unique number, except for the hydrogen atoms.
- i) Carbon atoms are shown in the notation only by digits.
- j) The other types of atoms carry, in addition their atomic symbols.

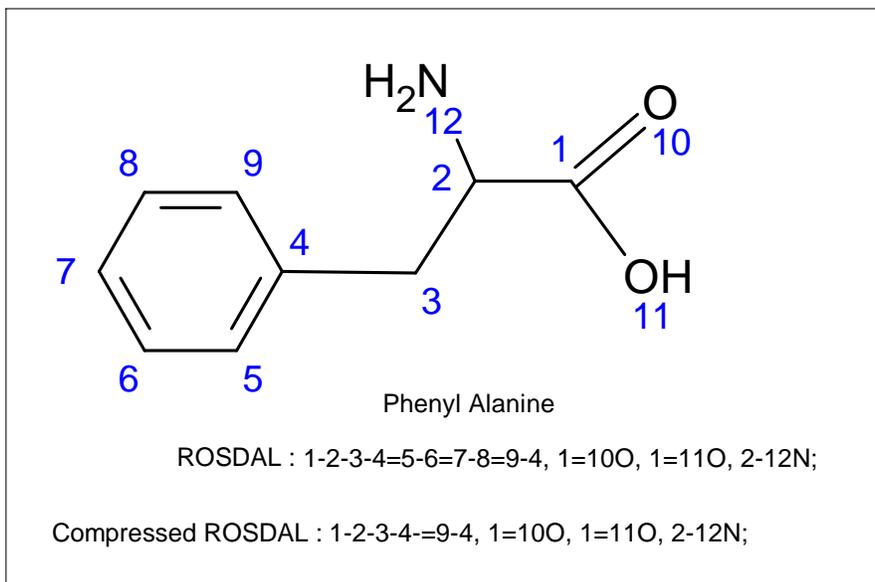
- k) In order to describe the bonds between atoms, bond symbols are inserted between the atoms numbers.
- l) Branches are marked and separated from the other parts of the code by commas. The ROSDAL linear notation is unambiguous but not unique.

**The sequence for setting up a ROSDAL notation is**

1. The structure diagram is drawn and the atoms are arbitrarily numbered (each atom is assigned a unique number).
2. Atomic symbols are usually written directly behind the index of an atom.
3. Usually only the indices of the carbon atoms are written, not the symbols: hydrogen atoms can have, but do not need, an atom number.
4. Bond types are described as follows:
  - “-“ for a single bond
  - “=” for a double bond
  - “#” for triple bond
  - “?” for any connection
5. Simplifications are allowed, such as writing alternating bonds as “-=”.
6. Commas separate branches and substituents.

Examples:

**Phenylalanine:**



**ROSDAL Advantages and disadvantages**

**Advantages:**

- 1) Simple code, easy to learn

- 2) Fast data exchange format
- 3) Includes stereochemistry

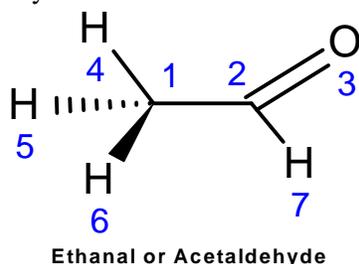
**Disadvantages**

- 1) No support for coding reactions
- 2) Not Unique

**6. Explain the structure of MDL Molfile format.**

In order to understand the Molfile format let us look at a sample files and recognize its fundamental structure. For simplicity, some less important details will be omitted in the discussion.

Molfile Example: Ethanal or Acetaldehyde



```

1. Blank Line or Molecule Name
2. ACD/Labs04131617572D [Creator Name]
3. Blank Line
4. 7 6 0 0 0 0 0 0 0 0 0 8 v2000
5. -0.8077 -0.0155 0.1782 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6. 0.6883 -0.0351 -0.0110 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7. 1.2933 1.0026 -0.1570 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8. -1.1027 0.8443 0.8174 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9. -1.2933 0.0808 -0.8174 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10. -1.1471 -0.9553 0.6649 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11. 1.2343 -1.0026 -0.0232 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12. 1 2 1 0 0 0 0
13. 1 4 1 0 0 0 0
14. 1 5 1 0 0 0 0
15. 1 6 1 0 0 0 0
16. 2 3 2 0 0 0 0
17. 2 7 1 0 0 0 0
18. M ZC 1 1
19. M ZC 2 2
20. M ZC 3 3
21. M ZC 4 4
22. M ZC 5 5
23. M ZC 6 6
24. M ZC 7 7
25. M END

```

Atom Block (Line:5 to 11)

Header block (Line:1 to 3)

Counts Line (Line: 4)

Bond block (Line:12 to 17)

Stereochemistry (Line:18 to 24)

Fig. Schematic Structure of Molfile for Acetaldehyde molecule or Ethanal molecule

The header consists of up to 3 lines:

**Line:1** Comment or molecule name line;

**Line:2** A line detailing the type of application which generated the table, the date and time of its creation.

**Line:3** Comment or description line that even if empty must be present.

**Line:4** These lines are followed by the counts line which gives the number of atoms and bonds (plus several additional flags) in the molecule.

**Line: 5 to 11**

The atom block , containing the x, y, and z co-ordinates of the atom and the element symbol for the atom. As this study used only topological (rather than topographical) structural information the positional coordinates were not used.

**Line: 12 to 17**

Finally, for the bond block , the first two numbers of each line indicates the numbers of the respective atoms from the atom list, and the third number indicates the bond type between these two atoms. (A value of 1 indicates a single bond connection, 2 a double bond and 3 a triple bond.)

**Line: 18 to 24**

**Line: 25 END**

7. Explain the structure of SDfile format.

8. Explain in brief any two commercial chemical structure and drawing software.