KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore – 641 021.

SYLLABUS
DEPARTMENT OF CHEMISTRY

STAFF NAME: Dr. S. RAVI
SUBJECT NAME: CHEMINFORMATICS SUB.CODE:16CHU501A
SEMESTER: V CLASS: III- B.Sc (CHEMISTRY)

Semester-V

18CHU501A CHEMINFORMATICS 3H 3C

Instruction Hours/week: L:3 T:0 P:0 Marks: Internal:40 External: 60 Total:100

Programme objectives

This course enables the student to

1. Understand the introduction to cheminformatics
2. Understand the Representation of molecules and chemical reactions
3. Understand the searching methods for chemical structures
4. Understand the computer assisted structure elucidations.

Programme outcome

The student will understand

1. The introduction to cheminformatics
2. The Representation of molecules and chemical reactions
3. The searching methods for chemical structures
4. The computer assisted structure elucidations.

UNIT I

Introduction to Chemoinformatics: History and evolution of chemoinformatics, Use of chemoinformatics, Prospects of chemoinformatics, Molecular Modelling and Structure elucidation.

UNIT II

Representation of molecules and chemical reactions: Nomenclature, Different types of notations, SMILES coding, Matrix representations, Structure of Molfiles and Sdfiles, Libraries and toolkits, Different electronic effects, Reaction classification.

UNIT III
Searching chemical structures: Full structure search, sub-structure search, basic ideas, similarity search, three dimensional search methods, basics of computation of physical and chemical data and structure descriptors, data visualization.

UNIT IV
Applications: Prediction of Properties of Compounds; Linear Free Energy Relations; Quantitative Structure-Property Relations; Descriptor Analysis; Model Building; Modelling Toxicity; Structure-Spectra correlations; Prediction of NMR, IR and Mass spectra.

UNIT V
Computer Assisted Structure elucidations; Computer Assisted Synthesis Design, Introduction to drug design; Target Identification and Validation; Lead Finding and Optimization; Analysis of HTS data; Virtual Screening; Design of Combinatorial Libraries; Ligand-Based and Structure Based Drug design; Application of Chemoinformatics in Drug Design.

Suggested Readings

Text Books:

1. Andrew R. Leach & Valerie, J. Gillet (2007). *An introduction to Chemoinformatics.* Springer: The Netherlands.
2. Gasteiger, J. & Engel, T. (2003).*Chemoinformatics: A text-book.* Wiley-VCH.

Reference Book

1. Gupta, S. P. (2011).*QSAR & Molecular Modeling.* New Delhi: Anamaya Pub.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore – 641 021.

LECTURE PLAN
DEPARTMENT OF CHEMISTRY

STAFF NAME: Dr. S.RAVI
SUBJECT NAME: CHEMINFORMATICS SUB.CODE:16CHU501A
SEMESTER: V CLASS: III-B.Sc (CHEMISTRY)

S.No.	Lecture Duration Period	Topics to be Covered	Support Material/Page Nos
		UNIT-I	
1	1	Introduction	T1: 1-3
2	1	History and evolution of chemoinformatics	T1: 9-10
3	1	Terms and Definitions	T1: 8-9
4	1	Use of chemoinformatics,	T1:487-490
5	1	Prospects of chemoinformatics	T1: 4-8
6	1	Molecular Modelling and Structure elucidation.	T1: 10-11
7	1	Recapitulation	
	Total No of Hours Planned For Unit 1=07		
		UNIT-II	
1	1	Representation of molecules and chemical reactions: Nomenclature	T1: 15-18; T2:1-6
2	1	Different types of notations,	T1: 9-12
3	1	SMILES coding, Matrix representations,	T1: 12-25
4	1	Structure of Molfiles and Sdfiles,	T1:33-40
5	1	Libraries and toolkits	T1: 123-150
6	1	Different electronic effects, Reaction classification.	T1: 172-173; 176-179
7	1	Recapitulation	
	Total No of Hours Planned For Unit II=07		

		UNIT-III	
1	1	Searching chemical structures:	T1:230-236
2	1	Full structure search, sub-structure search, basic ideas,	T1: 293-300
3	1	Similarity search, three dimensional search methods,	T1:313-315
4	1	Basics of computation of physical and chemical data	T2:99-106
5	1	Structure descriptors,	T1:403-408
6	1	Data visualization.	T1: 408-409
7	1	Recapitulation	
	Total No of Hours Planned For Unit III=07		
		UNIT-IV	
1	1	Prediction of Properties of Compounds;	T1: 176-180
2	1	Linear Free Energy Relations;	T1: 179-182
3	1	Quantitative Structure-Property Relations;	T1: 303-312; 489-490; T2:82-87
4	1	Descriptor Analysis; Model Building;	T1: 490-491
5	1	Modelling Toxicity; Structure-Spectra correlations;	T1:504-508
6	1	Prediction of NMR, IR and Mass spectra;	T1:518-535
7	1	Recapitulation	
	Total No of Hours Planned For Unit IV=07		
		UNIT-V	
1	1	Computer Assisted Structure elucidations;	T1:535-536
2	1	Computer Assisted Synthesis Design, Introduction to drug design;	T1:567-570
3	1	Target Identification and Validation; Lead Finding and Optimization	T1: 600-602
4	1	Analysis of HTS data; Virtual Screening;	T1:603-605; T2: 141-181
5	1	Design of Combinatorial Libraries; Ligand-Based and Structure Based Drug design;	T1:581-585 T2: 183-190
6	1	Application of Chemoinformatics in Drug Design.	T1:598-610
7	1	Recapitulation	

8	1	Discussion of previous year end semester question papers	
9	1	Discussion of previous year end semester question papers	
10	1	Discussion of previous year end semester question papers	
	Total No of Hours Planned for unit V=10		
Total Planned Hours	**38**		

Text Books:

1. Gasteiger, J. & Engel, T. (2003).*Chemoinformatics: A text-book.* Wiley-VCH.

2. Andrew R. Leach & Valerie, J. Gillet (2007). *An introduction to Chemoinformatics.* Springer: The Netherlands.

UNIT-I

SYLLABUS

Introduction to Chemoinformatics: History and evolution of chemoinformatics, Use of chemoinformatics, Prospects of chemoinformatics, Molecular Modelling and Structure elucidation.

UNIT I

Origins of Chemoinformatics

Chemoinformatics, in all but name, has existed for many decades. It could be argued that research has been conducted in this area since the advent of computers in the 1940s. However, the term *chemoinformatics* has only been in existence for the past decade yet it has quickly become a popular term with a number of books published that provide excellent overviews of the field. However, there is no true definition of chemoinformatics, most likely due to its highly interdisciplinary characteristics, and this has been a source of debate in recent years. Here, a few quotes

The mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization.

Frank K. Brown

[Chemoinformatics involves]. . . the computer manipulation of two- or three-dimensional chemical structures and excludes textual information. This distinguishes the term from chemical information, largely a discipline of chemical librarians and does not include the development of computational methods.

Peter Willett,

. . . the application of informatics to solve chemical problems . . . [and] chemoinformatics makes the point that you're using one scientific discipline to understand another scientific discipline.

From these quotations from leading scientists in the field of chemoinformatics, it is clear that there is still some dispute as to the "true" sphere of influence of chemoinformatics. Indeed, even the spelling of chemoinformatics is hotly debated, and the field is also referred to as *cheminformatics*, *chemical informatics*, *chemi-informatics*, and *molecular informatics*; and more recently our group at The Institute of Cancer Research is called *In Silico Medicinal Chemistry*.

The Similar-Structure, Similar-Property Principle

Much of chemoinformatics is essentially based on the fundamental assertion that similar molecules will also tend to exhibit similar properties; this is known as the *similar-structure, similar-property principle*, often simply referred to as the *similarproperty principle*, and described for molecules by Rouvray, thus: ". . . the so-called principle of similitude, which states that systems constructed similarly on different scales will possess similar properties." In large part, this is true; however, it must be emphasized that the similar-property principle is a heuristic and therefore will break down in certain circumstances. Another important caveat is that there are many ways of defining similarity.

The term cheminformatics so referred as Chemoinformatics/Chemiinformatics/Chemical information/ Chemical informatics has been recognised in recent years as a distinct discipline in computational molecular sciences. Cheminformatics is also known as interface science as it combines Physics, Chemistry, Biology, Mathematics, Biochemistry, Statistics and informatics. The primary focus of cheminformatics is to analyse/simulate/modelling/manipulate chemical information which can represented either in 2D structure or in 3D structure. Industry sectors such as, agrochemicals, food and pharmaceutical are distinct areas where cheminformatics plays significant role in the recent history of molecular sciences.

Cheminformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.

According to S.K.Brown, “The use of information technology and management has become a critical part of the drug discovery process. Cheminformatics is the mixing of information resources to transform data into information and information into knowledge which is collectively referred as inductive learning as shown in Fig.1. for the intended purpose of making better decisions faster in the areas of drug lead identification and organization. cheminformatics can be viewed as the application of informatics methods to solve chemical problems.

Cheminformatics has mainly dealt with small molecules, whereas bioinformatics addresses genes, proteins, and other larger chemical compounds (shown in Figure below). Chem and Bioinformatics complements each other for bimolecular process, like structure and function of proteins, the binding of a ligand to its binding site, the conversion of a substrate within its enzyme receptor, and the catalysis of a biochemical reaction by an enzyme.

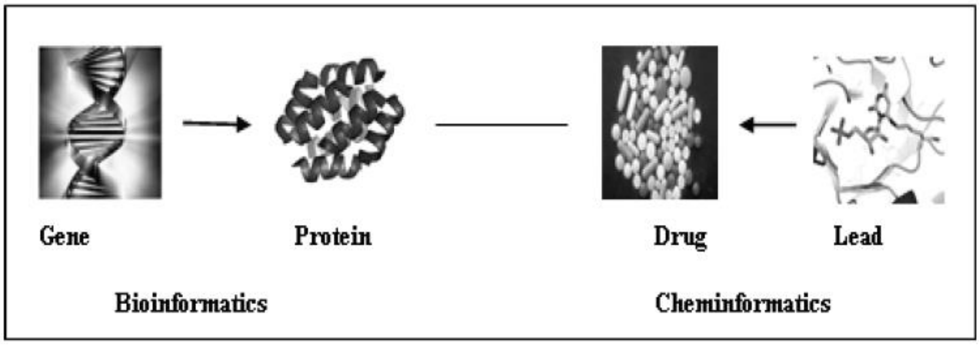


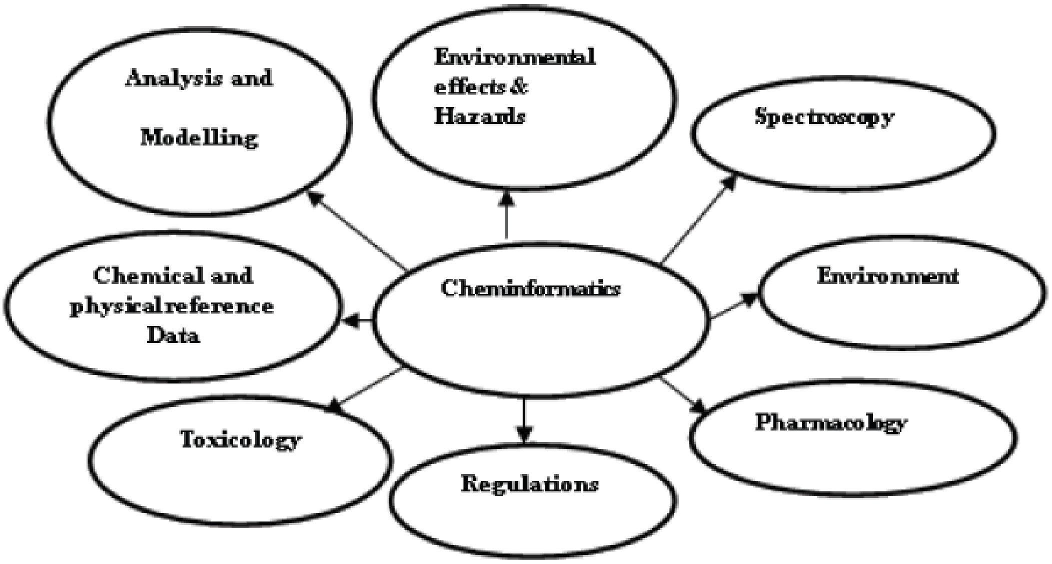
Fig. 2. The Cooperation of Bioinformatics and Cheminformatics

Different tools and methods are available to represent chemical structure, database to store chemical data, to perform searching process, Quality Structure- Activity Relationship(QSAR), Quality Structure Property Relationship(QSPR), to predict physical, chemical and biological properties of a molecule.

Need and importance of cheminformatics

Cheminformatics plays a key role to maintain and access enormous amount of chemical data, produced by chemist (more than 45 million chemical compounds are known and the number may increase in million every year,) by using a proper database. Also, the field of chemistry needs a novel technique for knowledge extraction from data to model complex relationships between the structure of the chemical compound and biological activity or the influence of reaction condition on chemical reactivity. Cheminformatics has wider range of application and Fig. 3. shows influence if cheminformatics in some specific research areas.

The need for cheminformatics



Three major aspects of Cheminformatics are

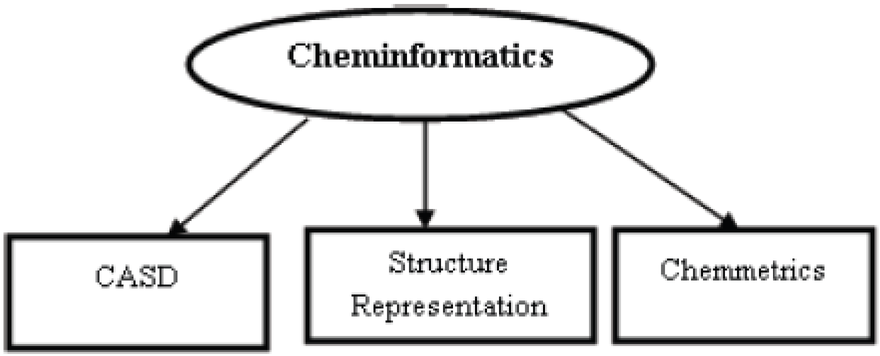
i) Information Acquisition, is a process of generating and collecting data empirically (experimentation) or from theory (molecular simulation)

ii) Information Management deals with storage and retrieval of information and

iii) Information use, which includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences.

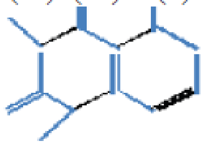
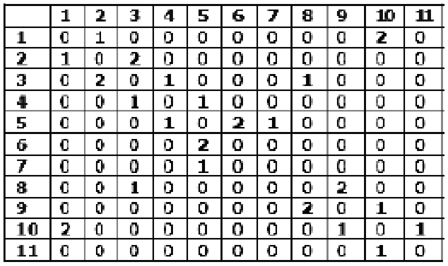
Cheminformatics and its Applications

Cheminformatics is a significant application of information technology to help chemists for investigating new problems, organize, analyse, and understand scientific data in the development of novel compounds, materials and processes. Primary modules of cheminformatics are Computer-Assisted Synthesis Design, Structure representation and chemmetrics



Computer-Assisted Synthesis Design (CASD) is applied mainly where artificial intelligence technique can be applied. This technique is applied in various applications which included pharmaceutical, food industry, textile industry and agro industry. Various forms of machine readable chemical representation play basic property to design chemical database where the chemical information are stored for analysis and manipulation. The chemical structure representations can be linear, 2D or in 3D format. Some of the chemical structure representations are shown in Table 1. SMILES (Simplified Molecular Input Line Entry Specification) is one of the linear chemical notation format which is widely used among chemist [38] for various clinical and analysis purpose. Structure representation deals with Reaction Representation, Structure Descriptors, Molecular Modelling, Structure Searching, and Computer-Assisted Structure Elucidation (CASE).

Table 1: Some of the Chemical Structure Representation

Representation	Name
Caffine	Common Name
trimethylxanthine coffeine, theine, mateine,	Synonyms
$C_8H_{10}N_4O_2$	Empirical formula
3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione	IUPAC Name
58-08-2	CAS Registry Number
T56 BN DN FNVNVJ B1 F1 H1	WLN Notation
CN1C=NC2=C1C(=O)N(C(=O)N2C)C	SMILES
1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	Inchl
	Markush Structure
	Connection Table

	1	2	3	4	5	6	7	8	9	10	11
1	0	1	0	0	0	0	0	0	0	2	0
2	1	0	2	0	0	0	0	0	0	0	0
3	0	2	0	1	0	0	0	1	0	0	0
4	0	0	1	0	1	0	0	0	0	0	0
5	0	0	0	1	0	2	1	0	0	0	0
6	0	0	0	0	2	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	2	0	0
9	0	0	0	0	0	0	0	2	0	1	0
10	2	0	0	0	0	0	0	0	1	0	1
11	0	0	0	0	0	0	0	0	0	1	0

[OH]c1cccc1	Fragment Code
0000100110100111	Fingerprint
5244987098423150	Hash Code

CASD and Structure representation

Structure Searching involves in determination of features like bond orders, rings and aromaticity. It includes searching the whole structure, substructure, structure similarity and diversity. CASE builds on information obtained from various spectroscopic methods like IR, NMR, MS, etc. Structure Descriptor used to identify the physical, chemical and biological properties of chemical compound and relationship between two structures. The descriptors fall into four classes such as,

i) Topological,
ii) Geometrical,
iii) Electronic and

iv) Hybrid or 3D Descriptors

Chemmetrics is used for quantitative analyse of the chemical data by using mathematical and statistical methods. It also deals with property prediction of chemical information

Applications of cheminformatics

The range of applications of cheminformatics is rich indeed; any field of chemistry can profit from its methods. The following lists different areas of chemistry and indicates some typical applications of cheminformatics.

a) Storing data generated through experiments or from molecular simulation Retrieval of chemical Structures from chemical database (Software libraries).

b) Prediction of physical, chemical and biological properties of chemical compounds.

c) Elucidation of the structure of a compound based on spectroscopic data.

d) Structure, Substructure, Similarity and diversity searching from chemical database

e) High Throughput Screening (HTS) is the integration of technologies (laborat ory automation, assay technology, micro plate based instrumentation, etc.) to quickly screen chemical compounds in search of a desired activity.

e) Docking - Interaction between two macromolecules.

f) Drug Discovery.

h) Molecular Science, Materials Science, Food Science (nutraceuticals), Atmospheric chemistry, Polymer chemistry, Textile Industry, Combinatorial organic synthesis (COS)

Tools Used for cheminformatics

The development of software and tools for computer assisted organic synthesis are under vast development. This has resulted in many tools and representations for chemical structures. Some of the tools are listed below.

ISIS-Draw is a chemical structure drawing program for Windows, published by MDL Information Systems. It is the interfacial software to ISIS/Base database.

ChemDraw is a molecule editor developed by the cheminformatics company Cambridge Soft. ChemDraw is, along with Chem3D and ChemFinder, part of the ChemOffice suite of programs and is available for Macintosh and Microsoft Windows.

ChemWindow, is a chemical structure drawing program with several template. The template can be created by the customer can be saved in template folder and opened in preference dialogue box .

ChemSketch, is a chemical structure drawing program with predefined templates are available for drawing and it is more powerful and user friendly tool for structure analysis

ChemReader is a software developer toolkit for translating digital raster images of chemical structures into standard, chemical file formats that can be searched and analyzed with other open source or commercial cheminformatic software.

JME Molecular Editor is a Java applet which allows to draw / edit molecules and reactions (including generation of substructure queries) and to depict molecules directly within an HTML page

LogCHEM, an Inductive Logic Programming (ILP) based tool for discriminative interactive mining of chemical fragments.

PLSR (PLS-Regression), a simple chemmetrics tool, which relates two matrix X and Y through linear multivariate model and has the ability to analyse data with many, noisy, collinear, and even incomplete variables in both X and Y.

Wendi (Web Engine for Nonobvious Drug Information), a web based integrative data mining tool. It attempts to find non-obvious relationships between a query compound and scholarly publications, biological properties, genes and diseases using multiple information sources.

ChemMine tool is an online service for small molecule analysis. It provides an interface between cheminformatics and data mining tools for various analytical analyses in chemical genomics and drug discovery.

CML (Chemical Markup Language) is degined as combination of semantic text and nontextual information of chemical strucutre on the internet. It acts like HTML pages.

MyChemise (My Chemical Structure Editor) is a new 2D structure editor. It is designed as a Java applet that enables the direct creation of structures in the Internet using a web browser. MyChemise saves files in a digital format (.cse) and the import and export of .mol files using the appropriate connection tables is also possible.

PubChem is an open repository for small molecules and their experimental biological activity. It integrates and provides search, retrieval, visualization, analysis, and

Open Babel is a chemical tool box which interconverts chemical structures between different formats, over 110 formats.

AmberTools is used Biomolecular simulation and analysis of polymers, nucleotides, and synthetic organic structures.

Some other tools such as, CAS Draw, DIVA (Diverse Information, Visualization and Analysis), Structure Checker Accord, DS Accord Chemistry Cartridge, MarvinSketch PowerMV, TINKER, APBS, ArgusLab, Babel, ioSolveIT, ChemTK, Chimera, CLIFF, Dragon, gOpenMol, Grace, JOELib, Jmol, IA_LOGP, Lammps, MIPSIM, Mol2Mol, AMSOL, MOLCAS, Molexel, ICM-Pro, ORTEP, Packmol, Polar, XLOGP,PREMIER Biosoft, Q-chem, ALOGPS, Qmol, SageMD, ChemTK Lite, Transient, CLOGP,TURBOMOLE, UNIVIS, VMD, WHATIF, GCluto, COSMOlogic, KOWWIN are also used for similar kind of applications mentioned above.

The what and why of cheminformatics

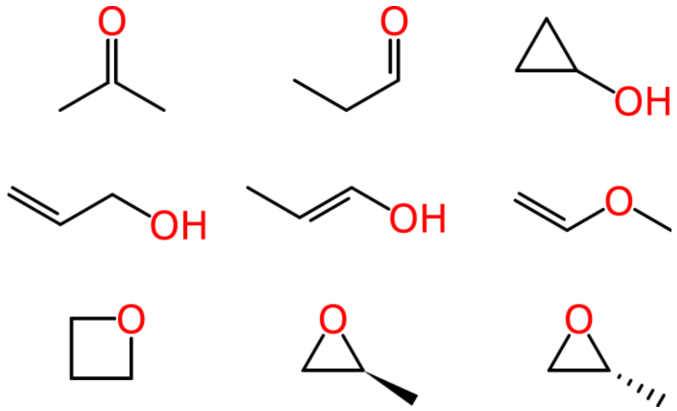
Communicating with chemists vs. communicating with computers

How does communication between chemists differ from communication between a chemist and a machine? If one chemist was to recommend to another that a reaction should be performed using "chloroform" as a solvent for a reaction, this would generally be a successful exercise in communication. For all practical purposes, this word is understood by every chemist, and has no ambiguity.

If this fact were to be communicated to a machine, things start to get a little murky. Humans are quite accustomed to learning common facts, and after sufficiently many years spent studying at university, they tend to become very good at looking up information and inferring missing data. Software algorithms, however, are supremely literal. Because "chloroform" is a so-called *trivial name*, there is no formula for converting it into the actual chemical structure that it represents, and so a machine will not be able to participate in this exchange of information unless it has been explicitly instructed as to what the word means, using a format that it can work with.

A more descriptive way to communicate the composition that is chloroform is by chemical formula, in this case $CHCl_3$. If this were handed over to a computer program, it would be a simple matter for it to understand that the substance being described is a molecule with 5 atoms: 1 carbon, 1 hydrogen and 3 chlorine. Assembling this into a molecule with bonds is a very simple matter, because 4 of the atoms are normally monovalent, and one of them is normally tetravalent. It is quite simple to create a software algorithm that can join the atoms together in the most obvious way, which also happens to be correct.

Beyond such tiny simple molecules, difficulties soon arise. Some of these ambiguities affect human chemists in the same way that they affect machines. Consider the molecular formula of C_3H_6O, which is associated with multiple reasonable structures, including a ketone, an aldehyde, a cyclic alcohol, oxygenated alkenes and cyclic ethers, one of which exists as two enantiomers:

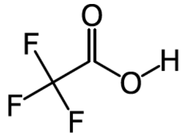


Structural formulas as chemical graphs

As we discussed usually, the most effective way to communicate with another chemist about the structure of a compound is to draw its structural formula.

In order to do cheminformatics, we need to express chemical structure in a way that can be understood by machines as well as humans. It just so happens that structural formulas can be fairly directly mapped to a computer-friendly data structure. Structural formulas can be interpreted as a kind of *graph*: a set of nodes (in our case, atoms), certain pairs of which are linked by edges (in our case, bonds). For example, consider this structural formula for trifluoroacetic acid, CF_3CO_2H:

(a)



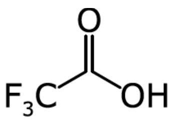
The diagram in has no ambiguity, since all atoms are represented in a graph: there are 8 nodes, each of which is labelled according to the element, except for carbon (which is the default). Each of the 7 bonds is represented by an edge within the graph, and the bond order is represented by showing the number of lines. The way that this formula is drawn on paper is completely compatible with a data structure based on a labelled graph.

Such *molecular graphs* are typically stored in dedicated file formats designed for chemical information. (There are many to choose from; we'll discuss the most popular ones a little later in this module and throughout the rest of this course.) This is good news, because cheminformaticians and computer scientists have come up with all kinds of clever data structures and algorithms for storing and analyzing datasets that can be represented as graphs.

However, just because structural formulas look like a graphs doesn't mean that they always look like a chemically-meaningful graph – or like a graph with the same chemical meaning that the chemist who drew the structure intended. There is a long list of issues that need to be handled carefully in order to reliably encode all of the chemical information contained in a structural formula in a machine-readable manner. (Long enough to ensure that cheminformatics will continue to be a lively field of research for a very long time!)

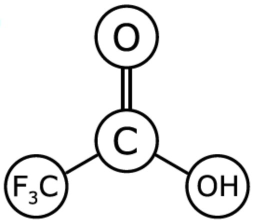
Consider a more common way of drawing trifluoroacetic acid:

(b)

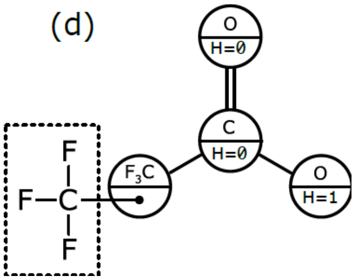


As we discussed chemists have become accustomed to condensing portions of structural formulas to make them easier to draw, read, and compare at a glance. The diagram (b) differs from the previous representation in that the carbon trisubstituted with fluoride has been collapsed into F_3C, a single node (to use the terminology of graphs). So has the hydroxyl part of the acid (OH). A schematic form of the corresponding graph is shown in (c). In this case, the underlying graph has 4 nodes, not 8. The labels of these nodes are [O, C, F_3C, OH]. Note that only two of these are elements from the periodic table. As shown, the structure represented by (c) *is not a molecular graph*, because some of its notes are labeled with something other than a symbol corresponding to an atom of a particular element.

(c)



(d)



In order to be interpreted by a computer as if it were a molecule, the molecular graph needs to be labelled in a slightly different way, which is represented schematically in (d). Here, we have annotated the graph nodes in a more systematic way.

Note that each node now has two labels. Three of the nodes are labeled so that the primary property is an atomic element, and the secondary property is the number of hydrogen atoms attached to it. These three node definitions are [O:H=0, C:H=0, O:H=1].

The fourth node, which represents a more complex group of atoms, is labeled slightly differently. Its primary property is the label F_3C, which is displayed for the benefit of human chemists reading the structural formula. However, its secondary property is labeled with*another graph* that stores the configuration of atoms and bonds that makes up F_3C. This is the underlying chemical information that the human chemist picks up easily when she sees “F_3C” in a structural formula, but that needs to be described explicitly in order for the computer to properly understand and use it.

Graph (d) is the best of both worlds. The structure can easily be displayed in the way that chemists expect to see it, but it can also be easily interpreted by a computer algorithm, because the definition follows rules that allow the full atomic structure to be reassembled behind the scenes.

Most importantly, always remember that just because a chemical structure has been *digitized* and stored on a computer does not mean that the information can be used by cheminformatics. In fact, most of the chemical structures that have been generated by scientists and put on computers are available not in a *graph*-based file format, but in a nonchemical *graphics* format. (There are two main types of such nonchemical image files: *bitmapped* and *vector* graphics. Neither is of much use for cheminformatics purposes.

In order to make use of *any* of the capabilities that cheminformatics offers, the molecules involved must be represented as a molecular graph, rather than such generic print-ready forms. Furthermore, the molecular graph must be sufficiently well defined that an algorithm can use the information to piece together the *complete* structure. Every single atom and bond must be present and accounted for, in some way or another.

Different data structures do so in different ways, each of which has its advantages and disadvantages. For example, some approaches try to keep the representation as similar as possible to the human-friendly version (e.g. Kekulé forms for aromatic rings). Others favor a selection of properties that is more similar to the quantum wave-function of the molecule (e.g. use of resonance/aromatic bonds to denote equivalence).

Molecular Modeling

Molecular modeling is a method that includes a variety of computational schemes that are aimed at simulating molecular structures, their properties and behavior in silico. In particular, this should also include molecular manipulations, that is, visualizing molecules on the screen using different modes, merging molecules, superimposing, and rotating molecules in space and bonds within individual molecules, and so on, as well as molecular predictions, that is, predicting molecular shape by 3D structure generation and modeling or forecasting chemical properties or eventual biological activity or effects. In particular, modeling virtual molecular structures themselves is not a trivial problem and can be achieved on the different level of approximation. In novel approaches we often sample VCS by systematically changing various molecular moieties in the user-directed mode. This can demand generation of thousands or even millions of structures and this operation can be achieved only by using the automated way. Such an operation can be easily programmed in a variety of environments, for example MATLAB, basing on SMILES codes whose syntax is simple enough. In a variety of chemical research, we simplify the real structure of a chemical molecule to its molecular configuration (cf. Section 4.14.2). What we usually mean by molecular configuration is a simplified 3D molecular structure, for example, we are classifying E and Z isomers as two different configuration series, although some other effects such as steric hindrance can further affect individual structures. Actually, in organic chemistry, we often rely on such simplification. However, molecules are 3D objects, which means each atom can be described by its exact space location. We can observe this by applying X-ray diffraction pattern on crystals, which allows us to reveal the 3D structure of the atomic lattice and thus to describe the 3D structure of the molecule. This effect is limited to condensed matter (crystals). Although there are many further approaches that allow chemists to disclose some structural data concerning the 3D atomic pattern, for example, by the application of NMR, current physics and chemistry do not have general technology for the observation of the 3D molecular structure. X-ray crystallography poses problems related to production of crystals, which is not always an easy task, and there is also the question of the relationship between condensed matter atom configuration and configuration in other environments. Even though nowadays we have data for quite a number of structures (Figure 12) including peptides or drug–ligand complexes, it is only a small percentage of the compounds described.

Possible Questions

Part A (Online multiple choice questions)

Part B (Each question carries two marks)

1. Define Cheminformatics
2. Write briefly about the history of cheminformatics
3. Why the process of learning cheminformatics is called an inductive learning
4. What is meant by an image file
5. How structures can be represented by a image file
6. How chemical structures are stored in a computer as molecular graphs.
7. What is meant by a connection table and how the chemical structures are stored in it
8. What is meany by bio-informatics
9. What are the three main aspects of cheminformatics
10. Write notes on Chemdraw.
11. What are the software tools available for cheminformatics
12. Classify structure descriptors.

Part C (Each question carries eight marks)

1. Explain the objectives of cheminformatics
2. Describe the needs to study cheminformatics
3. Explain the different ways of which structures can be represented
4. Explain in detail the different categories of databases on chemical structures
5. Classify and explain the different categories of databases
6. Explain the scope and applications of cheminformatics
7. Explain some of the chemical structures representation of caffine
8. What are the tools used for cheminformatics

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore – 641 021.

UNIT I (Multiple choice Questions)

S.No	Question	A	B	C	D	Answer
1	Structural data refers to	the 1-, 2- or 3-D representations of molecules.	biological activity, pka, log P, or analytical results	information such as experimental notes that are associated with a structure or data point.	any structure or data point may have associated graphical information such as spectra or plots.	the 1-, 2- or 3-D representations of molecules.
2	Numerical data refers to	the 1-, 2- or 3-D representations of molecules.	biological activity, pka, log P, or analytical results	information such as experimental notes that are associated with a structure or data point.	any structure or data point may have associated graphical information such as spectra or plots.	biological activity, pka, log P, or analytical results
3	Annotation/text data refers to	the 1-, 2- or 3-D representations of molecules.	biological activity, pka, log P, or analytical results	information such as experimental notes that are associated with a structure or data point.	any structure or data point may have associated graphical information such as spectra or plots.	information such as experimental notes that are associated with a structure or data point.
4	Graphical data refers to	the 1-, 2- or 3-D representations of molecules.	biological activity, pka, log P, or analytical results	information such as experimental notes that are associated with a structure or data point.	any structure or data point may have associated graphical information such as spectra or plots.	any structure or data point may have associated graphical information such as spectra or plots.
5	the 1-, 2- or 3-D representations of molecules. Is component of	Structural data	Numerical data	Annotation/text data	Graphical data	Structural data
6	biological activity, pka, log P, or analytical results is a	Structural data	Numerical data	Annotation/text data	Graphical data	Numerical data

	component of					
7	information such as experimental notes that are associated with a structure or data point. Is a component of	Structural data	Numerical data	Annotation/text data	Graphical data	Annotation/text data
8	any structure or data point may have associated graphical information such as spectra or plots.	Structural data	Numerical data	Annotation/text data	Graphical data	Graphical data
9	Literature databases consists of	author names, titles, journals or books etc	contains alphanumeric data or compounds	contains information on chemical structure and compounds.	the 1-, 2- or 3-D representations of molecules.	author names, titles, journals or books etc
10	Factual databases consists of	author names, titles, journals or books etc	contains alphanumeric data or compounds	contains information on chemical structure and compounds.	the 1-, 2- or 3-D representations of molecules.	contains alphanumeric data or compounds
11	Structure databases consists of	author names, titles, journals or books etc	contains alphanumeric data or compounds	contains information on chemical structure and compounds.	the 1-, 2- or 3-D representations of molecules.	contains information on chemical structure and compounds.
12	A data base consists of author names, titles, journals or books etc	Literature databases	Structure databases	Factual databases	Molecularity data base	Literature databases
13	A data base contains alphanumeric data or compounds	Literature databases	Structure databases	Factual databases	Molecularity data base	Factual databases
14	A data base contains information on chemical structure and compounds.	Literature databases	Structure databases	Factual databases	Molecularity data base	Structure databases
15	Chemical Abstracts was started in	1907	1975	1832	1961	1907
16	The first information systems and services	Annalen der Pharmacie	Chemical Abstracts	journal for chemical information	Handbook of Computer Handling of Chemical Structure Information	Annalen der Pharmacie
17	The term Chemoinformatics was defined by	J.Gastegeir	F.K. Brown	G. Paris	M. Hann, R	F.K. Brown

18	The first journal for chemical information	Annalen der Pharmacie	Chemical Abstracts	Handbook of Computer Handling of Chemical Structure Information	Handbook of Computer Handling of Chemical Structure Information	Handbook of Computer Handling of Chemical Structure Information
19	Chemdraw & ISIS/Draw represent	Image file	Graph theory	Connection tables	Linear notation	Image file
20	Chemical structures are usually stored in a computer as molecular graphs, it is called	Image file	Graph theory	Connection tables	Linear notation	Graph theory
21	it is a means to communicate the molecular graph to and from the computer	Image file	Graph theory	Connection tables	Linear notation	Connection tables
22	it represents the structure of chemical compounds as a linear sequence of letters and numbers	Image file	Graph theory	Connection tables	Linear notation	Linear notation
23	Linear notation is	it represents the structure of chemical compounds as a linear sequence of letters and numbers	it is a means to communicate the molecular graph to and from the computer	Chemical structures are usually stored in a computer as molecular graphs,	Chemdraw & ISIS/Draw	it represents the structure of chemical compounds as a linear sequence of letters and numbers
24	Connection tables are	it represents the structure of chemical compounds as a linear sequence of letters and numbers	it is a means to communicate the molecular graph to and from the computer	Chemical structures are usually stored in a computer as molecular graphs,	Chemdraw & ISIS/Draw	it is a means to communicate the molecular graph to and from the computer
25	Graph theory is	it represents the structure of chemical compounds as a linear sequence of letters and numbers	it is a means to communicate the molecular graph to and from the computer	Chemical structures are usually stored in a computer as molecular graphs,	Chemdraw & ISIS/Draw	Chemical structures are usually stored in a computer as molecular graphs,

26	Example for Image file	it represents the structure of chemical compounds as a linear sequence of letters and numbers	it is a means to communicate the molecular graph to and from the computer	Chemical structures are usually stored in a computer as molecular graphs,	Chemdraw & ISIS/Draw	Chemdraw & ISIS/Draw
27	Information Acquisition, is a process of	generating and collecting data empirically (experimentation) or from theory	storage and retrieval of information	includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences	Bioinformatics	generating and collecting data empirically (experimentation) or from theory
28	Information Management deals	generating and collecting data empirically (experimentation) or from theory	storage and retrieval of information	includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences	Bioinformatics	storage and retrieval of information
29	Information use, which includes	generating and collecting data empirically (experimentation) or from theory	storage and retrieval of information	includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences	Bioinformatics	includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences
30	generating and collecting data empirically or from theory is called	Information Acquisition	Information Management	Information use	cheminformatics	Information Acquisition
31	storage and retrieval of information is called	Information Acquisition	Information Management	Information use	cheminformatics	Information Management
32	includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences	Information Acquisition	Information Management	Information use	cheminformatics	nformation use
33	Bioinformatics is the study of	Information Acquisition	Information Management	Information use	addresses genes, proteins, and other larger chemical compounds	addresses genes, proteins, and other larger chemical compounds
34	The code represents caffine 5244987098423150	Hash code	Fingerprint	Fragment Code	Empirical formula	Hash code
35	The code represents	Hash code	Fingerprint	Fragment Code	Empirical formula	Fingerprint

	caffine 0000100110100111					
36	The code represents caffine [OH]c1cccc1	Hash code	Fingerprint	Fragment Code	Empirical formula	Fragment Code
37	The code represents caffine C8H10N4O2	Hash code	Fingerprint	Fragment Code	Empirical formula	Empirical formula
38	Hash code for caffine	5244987098423150	0000100110100111	[OH]c1cccc1	C8H10N4O2	5244987098423150
39	Fingerprint of caffine	5244987098423150	0000100110100111	[OH]c1cccc1	C8H10N4O2	0000100110100111
40	Fragment Code of caffine	5244987098423150	0000100110100111	[OH]c1cccc1	C8H10N4O2	[OH]c1cccc1
41	ISIS-Draw is a chemical structure drawing program for Windows	published by MDL Information Systems.	published by CambridgeSoft	Inductive Logic Programming	Web Engine for Nonobvious Drug Information	MDL Information Systems.
42	ChemDraw is a molecule editor	published by MDL Information Systems.	published by CambridgeSoft	Inductive Logic Programming	Web Engine for Nonobvious Drug Information	published by CambridgeSoft
43	ChemReader is a software developer toolkit	published by MDL Information Systems.	published by CambridgeSoft	for translating digital raster images of chemical structures into standard	Web Engine for Nonobvious Drug Information	for translating digital raster images of chemical structures into standard
44	The synonym of caffine is	Theine	morphine	reserpine	codeine	Theine
45	Characterising molecular compound is	cheminformatics	Genomics	Proteomics	Pharmacogenomics	cheminformatics
46	Which of the following terms refers to the molecular modelling computational method that uses equations obeying the laws of classical physics	Quantum mechanics	Molecular mechanics	Molecular calculations	Quantum theory	Molecular mechanics
47	The integration of technolopgies to quickly screen chemical compounds in search of a desired activity	HTS	docking	Drug discovery	QSAR	HTS
48	A molecule editor	Chem draw	Chem window	ISIS draw	Chem reader	Chem draw

	developed by the cheminformatics					
49	A study of interaction of drug molecule and a protein is called	HTS	docking	Drug discovery	QSAR	Docking
50	The 1998 Nobel Chemistry Prize was awarded to Pople and Kohn for their work in	Computational Chemistry and Molecular Modelling	Nanotechnology	Green chemistry	Electrochemistry	Computational Chemistry and Molecular Modelling
51	In a chemical database, structural data refers to	2D representation of a molecule	Log P value	Experimental notes	Spectra or plots	2D representation of a molecule
52	The subject which plays a key role to maintain and access enormous amount of chemical data, produced by chemist is	Statistics	Bioinformatics	Cheminformatics	Vector algebra	Cheminformatics
53	A list of the bonds, stomic numbers, bonds, hybridization state and bond order constitute	Image file	Graph theory	Connection table	Linear notation	Connection table
54	The line notation in which hydrogen is not included	WLN	ROSDAL	SMILES	SLN	SLN
55	The line notation in which hydrogen is removed	WLN	ROSDAL	SMILES	SLN	SMILES
56	The line notation in which a symbol was allotted for conjugated bonds	WLN	ROSDAL	SMILES	SLN	SLN
57	In the representation of structures the branches are represented inside parenthesis	Line notations	Graph theory	Matrix representation	Full structure representation	Line notations

UNIT II

Representation of molecules: Nomenclature, Different types of notations, SMILES coding, Matrix representations, Structure of Molfiles and Sdfiles, Libraries and toolkits, Different electronic effects, Reaction classification.

Nomenclature of Inorganic compounds

1. Electropositive elements are listed first.
2. The stoichiometry of the elements is indicated at the lower right hand side by index numbers.
3. The charges of ions are placed at the top right-hand side next to the element symbol (eg S^{2-}).
4. In ions of complexes, the central atom is specified before the ligands are listed in alphabetical order, the complex ion is set in square brackets.

Nomenclature of organic compounds

1. The elements of an organic compound are listed in empirical formulas and the stoichiometry is indicated by index numbers.
2. Carbon and hydrogen atoms were positioned in the first and second places respectively and the hetero atoms following them in alphabetical order (eg. $C_9H_{11}NO_2$).
3. However since different compounds have the same empirical formulae, formulas were developed that indicate the presence of certain structural units and functional groups.

Eg. the empirical formulae of phenylalanine may be split into a more extended form that shows the presence of a phenyl ring, as well as an amino acid and a carboxylic acid group.

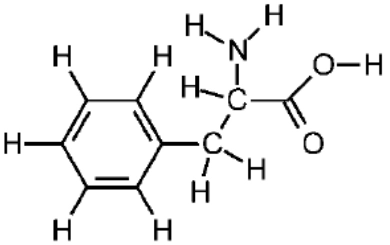
Empirical formula	Structure diagram	Condensed formula
$C_9H_{11}NO_2$		$C_6H_5CH_2CH(NH_2)CO_2H$

Fig: Different representations of phenylalanine.

Systematic IUPAC nomenclature of compounds

1. Trival names are short and simple to memorise.

2. IUPAC name can be quite long and cumbersome. The aim is to describe particular parts of the structure (fragments) in a systematic manner, with special expressions from a vocabulary of terms.

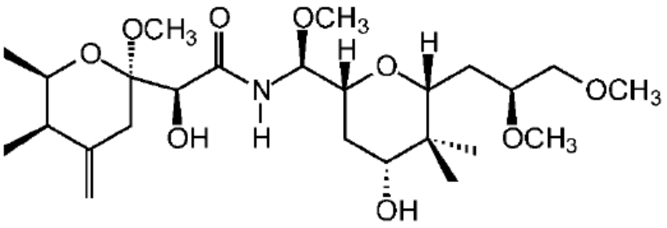
3. So the systematic nomenclature is used in Chemical Abstracts Service as index for chemical structures.

4. However this does not directly allow the extraction of additional information about the molecule, such as bond orders or molecular weight.

5. There are two basic rules for the nomenclature of organic compounds.
    a. No. of carbon atoms in the longest continuous aliphatic chain of carbon atoms has to be indicated. Branching of the skeleton, and the presence of rings, have to be specified by prefixes.
    b. Functional groups are to be specified by a prefix and /or suffix to indicate the family to which the compounds belongs.
    c. Substituents are listed in the name in alphabetical order.

Eg. The IUPAC name of phenylalanine is 2-amino-3-phenylprppanoic acid. It indicates a carbon chain of length three (propan-) of the acid (propanoic acid) with an additional two structural units, the phenyl and the amino group at different positions on the carbon chain: phenyl at carbon atom number 3 and amino at 2, with the counting beginning with the carbon atom of the acid group (COOH).

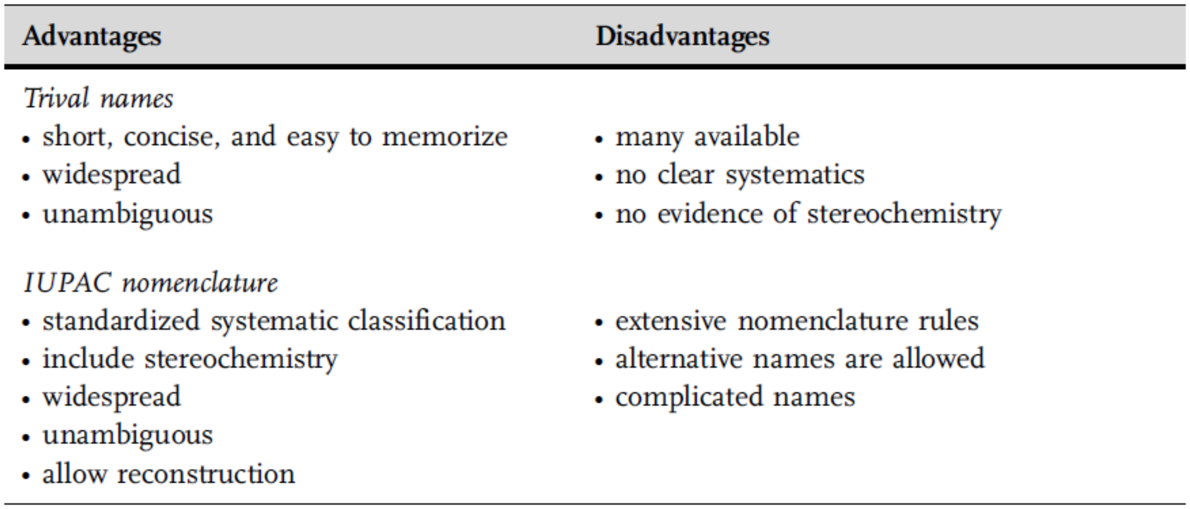
Neither a trival name not the IUPAC name, which both represent the structure as an alphanumerical (text) string, is ideal for computer processing. The reason is that various valid compound names can describe one chemical structure. As a consequence, the name/structure correlation is unambiguous but not unique. Nowadays programmes can translate names of structures and structures to names, to make published structures accessible in electronic journals



- D-***manno*-Nonitol**, 2,6-anhydro-3,5,7-trideoxy-1-*C*-{[hydroxy-(tetrahydro-2-methoxy-5,6-dimethyl-4-methylene-2*H*-pyran-2-yl)acetyl]amino}-5,5-dimethyl-1,8,9-tri-*O*-methyl-,{2*R*-[2α,2[*S**(*S**)], 5β,6β]}-
- **2*H*-Pyran-2-acetamid**, *N*-[[6-(2,3-dimethoxypropyl)tetrahydro-4-hydroxy-5,5-dimethyl-2*H*-pyran-2-yl]methoxymethyl]tetrahydro-α-hydroxy-2-methoxy-5,6-dimethyl-4-methylene]-

Fig. Various logical compound names can describe one logical structure

Evaluation of chemical nomenclature systems for representing a chemical structure.



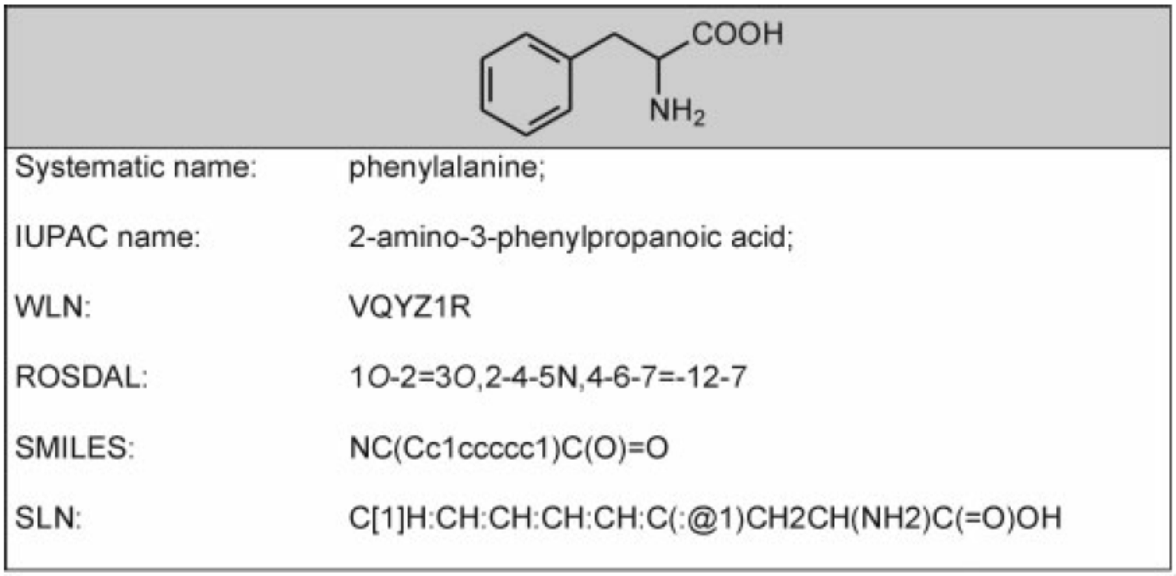
Advantages	Disadvantages
Trival names	
• short, concise, and easy to memorize • widespread • unambiguous	• many available • no clear systematics • no evidence of stereochemistry
IUPAC nomenclature	
• standardized systematic classification • include stereochemistry • widespread • unambiguous • allow reconstruction	• extensive nomenclature rules • alternative names are allowed • complicated names

Notations

Line notation

It represents the structure of chemical compounds as a linear sequence of letters and numbers. IUPAC nomenclature represents a kind of line notation. However it makes difficult to obtain additional information on the structure of a compound directly from its name.

A chemist trained in this line notation, could enter the code of large molecules faster than with a structure-editing program.



Systematic name:	phenylalanine;
IUPAC name:	2-amino-3-phenylpropanoic acid;
WLN:	VQYZ1R
ROSDAL:	1O-2=3O,2-4-5N,4-6-7=-12-7
SMILES:	NC(Cc1ccccc1)C(O)=O
SLN:	C[1]H:CH:CH:CH:CH:C(:@1)CH2CH(NH2)C(=O)OH

Fig: Different line notations for the structure diagram of phenylalanine

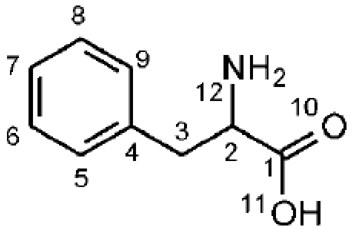
Wiswesser Line Notation (WLN)

It was introduced in 1946. A line notation represents a chemical structure by an alphanumeric sequence, which significantly simplifies the processing by the computer. In many cases the WLN uses the symbols for the chemical elements. Additionally, functional groups, ring systems, positions of ring substituents and position of condensed rings are assigned to individual letters or combination of symbols. It facilitates the search for particular functional groups and for fragments in a molecule. Thus, the machine retrieval of WLN characterizes the parts of a molecule. It uses 40 symbols from the following character sets.

1. Capital letters: 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 position.

ROSDAL (Representation of Organic Structures Description Arranged Linearly)

1. This line notation was intended to transmit structural information between the user and the Beilstein (DIALOG) system during database retrieval queries and structure displays. This exchange of structure information by the ROSDAL ASC II character string is very fast.
2. It is characterised by a simple coding of a chemical structure using alphanumeric symbols which can easily be learned by a chemist.
3. In the linear structure representation each atom of the structure is arbitrarily assigned a unique number, except for the hydrogen atoms.
4. Carbon atoms are shown in the notation only by digits.
5. The other types of atoms carry, in addition, their atomic symbol.
6. In order to describe the bonds between atoms, bond symbols are inserted between the ato numbers.
7. Branches are marked and separated from the other parts of the code by commas.



a) 1-2-3-4=5-6=7-8=9-4,1=10O,1-11O,2-12N;

b) 1-2-3-4-=9-4,1-11O,1=10O,2-12N;

Fig: A possible ROSDAL code for phenylalanine in a) a complete and b) a compressed notation.

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 a triple bond

   “$” for any connection
5. Simplifications are allowed, such as writing alternating bonds as “-=”.
6. Commas separate branches and substituents.

SMILES Coding (Simplified Molecular Input Line Entry System)

a. Chemical structure information is highly compressed and simplified in this notation
b. The flexible, easy to learn language describes chemical structures as a line notation.
c. The SMILES language has found widespread distribution as a universal chemical nomenclature for the representation and exchange of chemical structure information, independently of software or hardware architecture.

The basic rules of SMILES are

1. Atoms are represented by their atomic symbols
2. Hydrogen atoms automatically saturate free valences and are omitted (simple hydrogen connection)
3. Neighbouring atoms stand next to each other
4. Double and triple bonds are characterised by “=” and “#” respectively
5. Branches are represented by parenthesis
6. Rings are described by allocating digits to the two connecting ring atoms.

The compact textual coding requires no graphical input and additionally permits a fast transmission.

SMILES code	Chemical structure	Compound name

Atoms: 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_4	methane
[Fe+2] or [Fe++]	Fe^{2+}	iron (II) cation

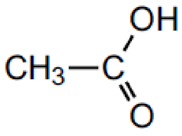
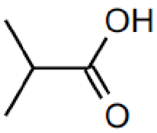
Bonds: Single, double and triple and aromatic (or conjugated) bonds are indicated by the symbols “-“, “=”, “#”, and ‘:” respectively; Single and aromatic bonds should be omitted.

C=C	$H_2C{=}CH_2$	ethene
O=CO	HCOOH	formic acid

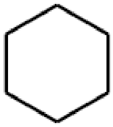
Disconnected structures in the molecule: Individual parts of the compound are separated by a period. The period indicates that there is no connection between atoms or parts of a molecule. The arrangement of the parts is arbitrary.

[Na+].[OH-]	NaOH	sodium hydroxide

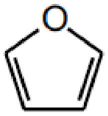
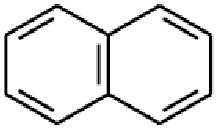
Branches: Branches are indicated within parenthesis.

CC(=O)O		acetic acid
CC(C)C(=O)O		isobutyric acid

Cyclic structures: Rings are described by breaking the ring between two atoms and then labeling the two atoms with the same number.

C1CCCCC1		cyclohexane

Aromaticity: Aromatic structures are indicated by writing all the atoms involved in lower-case letters

o1cccc1		furan
c12c(cccc1)cccc2 same as c1cc2ccccc2cc1		naphthalene

Sybyl Line Notation (SLN)

It is used to represent molecular structures including common organic molecules, macromolecules, polymers and combinatorial libraries. Its main distinction from SMILES is that all hydrogen atoms must be specified, because no assumptions are made regarding standard valences.

The compact textual coding requires no graphical input and additionally permits a fast transmission.

SLN code	Chemical structure	Compound name

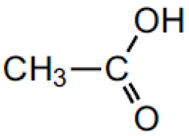
Atoms: Atoms are represented by their atomic symbols. The first letter is upper case, and in two-letter symbols the second letter is lower case. Hydrogen atoms must be specified.

CH4	CH_4	methane
NH2	$-NH_2$	amine

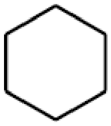
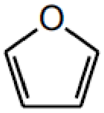
Bonds: Single bonds are omitted; double and triple and aromatic (or conjugated) bonds are indicated by the symbols "=", "#", and ':' respectively; In contrase to SMILES aromaticity is not an atomic property but a property of bonds. A period indicates the start of a new part of the structure.

HC(=O)OH	HCOOH	formic acid
Na.OH	NaOH	sodium hydroxide

Branches: Branches are indicated by parenthesis

CH3C(=O)OH		acetic acid

Cyclic structure: Ring closures are described by a bond to a previously defined atom which is specified by a unique ID number. The ID is a positive integer placed in square brackets behind the atom. An "@" indicates a ring closure.

C[15]H2CH2CH2CH2CH2CH2@15		cyclohexane
O[6]:CH:CH:CH:CH:@6		furan

Matrix Representations

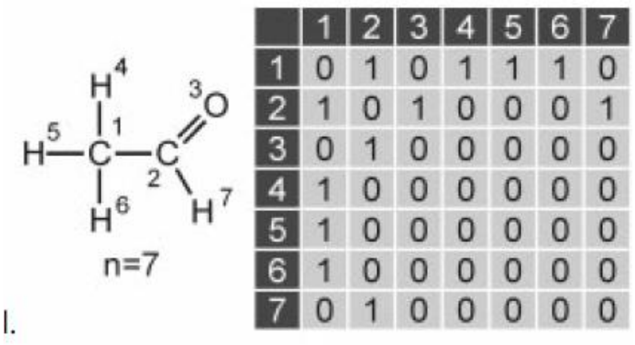
The matrix of a structure with 'n' atoms consists of an array of n x n entries. A molecule with its different atoms and bond types can be represented in matrix form in different ways depending on what kind of entries are chosen for the atoms and bonds. Thus a variety of matrices has been proposed; adjacency, distance, incidence, bond, and bond-electron matrices.

Hydrogen atoms are sometimes not shown, because their numbers and positions can be calculated from organic structures on the basis of the valence rules of the other atoms.

Each atom is described twice- in a column and in a row. Matrices in which all elements are shown twice are called redundant. A non-redundant matrix contains each element only once (eg. only the top right or bottom left triangle of the matrix).

Adjacency Matrix

The adjacency matrix of a molecule consisting of n atoms in a square (n x) matrix with the entries giving all the connectivities of the atoms. The intersection of a row and a column obtains a value of 1 if the corresponding atoms are connected. If there is no bond between the atoms being considered, the position in the matrix obtains the value 0. Thus, this matrix representation is a Boolean matrix with bits (0 or 1).

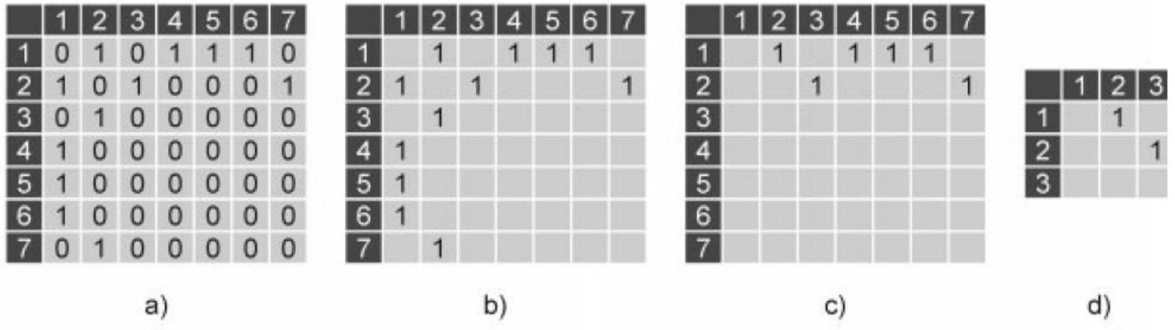


	1	2	3	4	5	6	7
1	0	1	0	1	1	1	0
2	1	0	1	0	0	0	1
3	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0

Fig : Adjancy (7 x 7) matrix of ethanal

As can be seen the diagonal elements of the matrix are always zero and it is symmetric around the diagonal elements. Thus it is a reductant matrix and can be reduced to half of its entries. For clarity, all zero entries are omitted in fig b-d.

With such a matrix representation, the storage space is dependent only on the number of nodes (atoms) and independent of the number of bonds. All the essential information in an adjancy matrix can also be found in the much smaller non-reductant matrix. But the adjancy matrix is unsuitable for reconstructing the constitution of a molecule, because it does not provide any information about the bond orders.



a)

	1	2	3	4	5	6	7
1	0	1	0	1	1	1	0
2	1	0	1	0	0	0	1
3	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0

b)

	1	2	3	4	5	6	7
1		1		1	1	1	
2	1		1				1
3		1					
4	1						
5	1						
6	1						
7		1					

c)

	1	2	3	4	5	6	7
1		1		1	1	1	
2			1				1
3							
4							
5							
6							
7							

d)

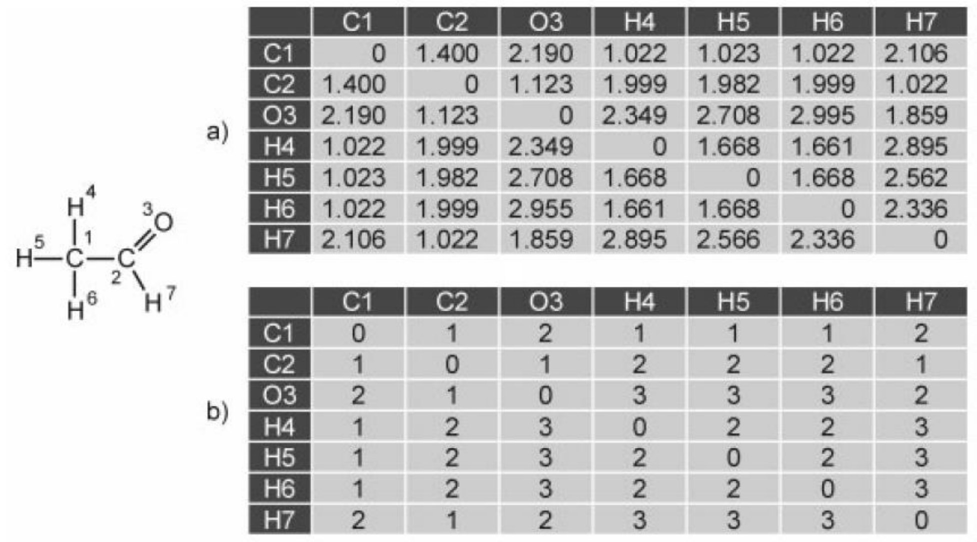
	1	2	3
1		1	
2			1
3			

Figure 2-14. a) The redundant adjacency matrix of ethanal (see Figure 2-13) can be simplified step by step by b) omitting the zero values, c) reducing it to the top right triangle, and, finally, d) omitting the hydrogen atoms.

Distance matrix

The elements of a distance matrix contain values which specify the shortest distance between the atoms involved. Distances can be expressed either as geometric distances (in A) or as topological distances (in number of bonds)

.


H
4
3
O
H
5
1
C
C
2
H
6
H
7

a)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1.400	2.190	1.022	1.023	1.022	2.106
C2	1.400	0	1.123	1.999	1.982	1.999	1.022
O3	2.190	1.123	0	2.349	2.708	2.995	1.859
H4	1.022	1.999	2.349	0	1.668	1.661	2.895
H5	1.023	1.982	2.708	1.668	0	1.668	2.562
H6	1.022	1.999	2.955	1.661	1.668	0	2.336
H7	2.106	1.022	1.859	2.895	2.566	2.336	0

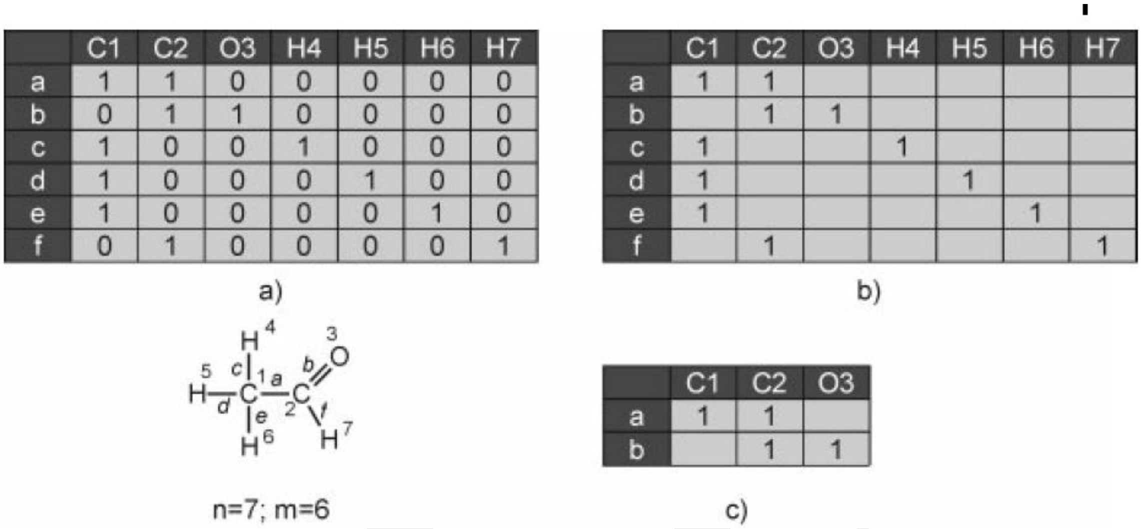
b)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1	2	1	1	1	2
C2	1	0	1	2	2	2	1
O3	2	1	0	3	3	3	2
H4	1	2	3	0	2	2	3
H5	1	2	3	2	0	2	3
H6	1	2	3	2	2	0	3
H7	2	1	2	3	3	3	0

Figure 2-15. Distance matrices of ethanal with a) geometric distances in Å and b) topological distances. The matrix elements of b) result from counting the number of bonds along the shortest walk between the chosen atoms.

Incidence Matrix

The incidence matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m). An entry obtains the value of 1 if the corresponding edge ends in this particular node.



	C1	C2	O3	H4	H5	H6	H7
a	1	1	0	0	0	0	0
b	0	1	1	0	0	0	0
c	1	0	0	1	0	0	0
d	1	0	0	0	1	0	0
e	1	0	0	0	0	1	0
f	0	1	0	0	0	0	1

a)

	C1	C2	O3	H4	H5	H6	H7
a	1	1					
b		1	1				
c	1			1			
d	1				1		
e	1					1	
f		1					1

b)

H 4
3 O
5 H
c
1 a
b
C
C
d
e
2
f
H 6
H 7

n=7; m=6

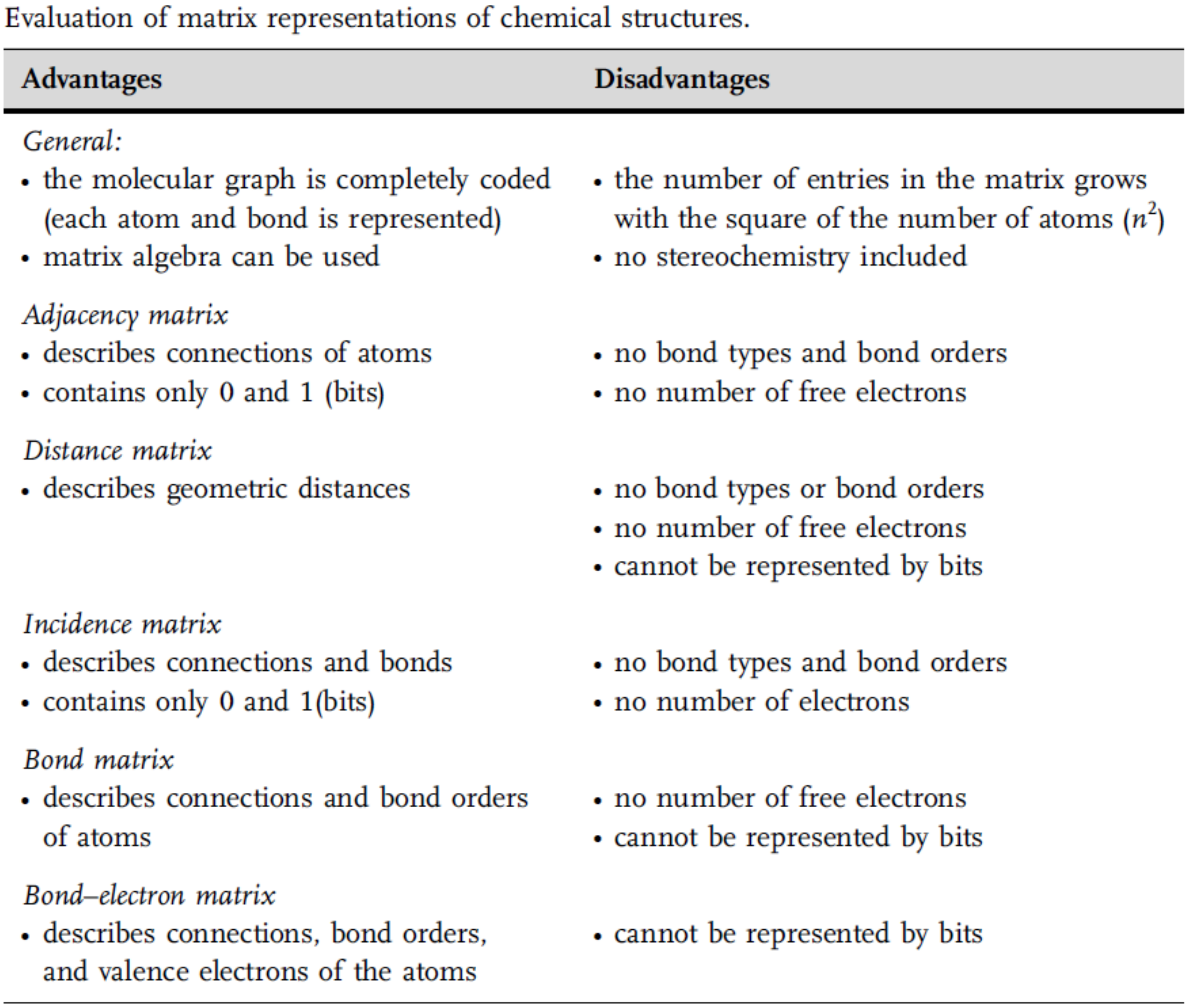
	C1	C2	O3
a	1	1	
b		1	1

c)

Figure 2-16. a) The redundant incidence matrix of ethanal can be compressed by b) omitting the zero values and c) omitting the hydrogen atoms. In the non-square matrix, the atoms are listed in columns and the bonds in rows.

Bond Matrix

The bond matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms. Elements of the matrix obtain the value of 2 if there is a double bond between the atoms. Eg. between atoms 2 and 3. Otherwise the value can be 0,1 or 3 for other bonding combinations. This representation is redundant as well.

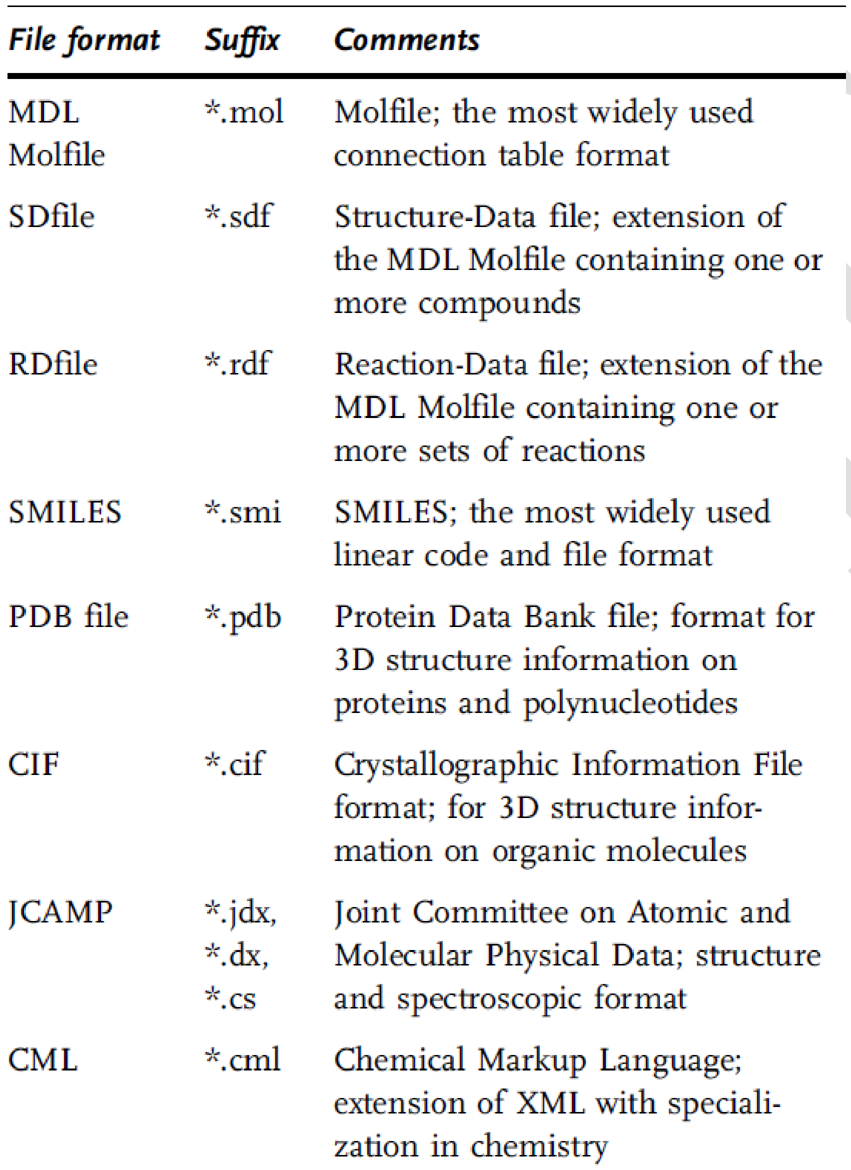


Evaluation of matrix representations of chemical structures.

Advantages	Disadvantages
General:	
• the molecular graph is completely coded (each atom and bond is represented) • matrix algebra can be used	• the number of entries in the matrix grows with the square of the number of atoms (n^2) • no stereochemistry included
Adjacency matrix	
• describes connections of atoms • contains only 0 and 1 (bits)	• no bond types and bond orders • no number of free electrons
Distance matrix	
• describes geometric distances	• no bond types or bond orders • no number of free electrons • cannot be represented by bits
Incidence matrix	
• describes connections and bonds • contains only 0 and 1(bits)	• no bond types and bond orders • no number of electrons
Bond matrix	
• describes connections and bond orders of atoms	• no number of free electrons • cannot be represented by bits
Bond–electron matrix	
• describes connections, bond orders, and valence electrons of the atoms	• cannot be represented by bits

Structure of MOL files

In chemistry, numerous software programs are available to handle structure information on molecules. The one task in common is to save data in a file. Many organizations have developed their own connection table format and a quite a fe have made provisions for the import or export of other file formats. 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 format of one software system has to be understood by another i.e. converted into its internal file format. MDL Molfile format developed at Molecular Design Limited became a standard file format. Several extensions have been made to the MDL Molfile format, leading to the SDfile, RGfile, Rxnfile or RDfile, with each one having special additional information on one or several molecules.

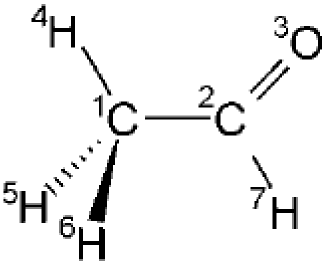


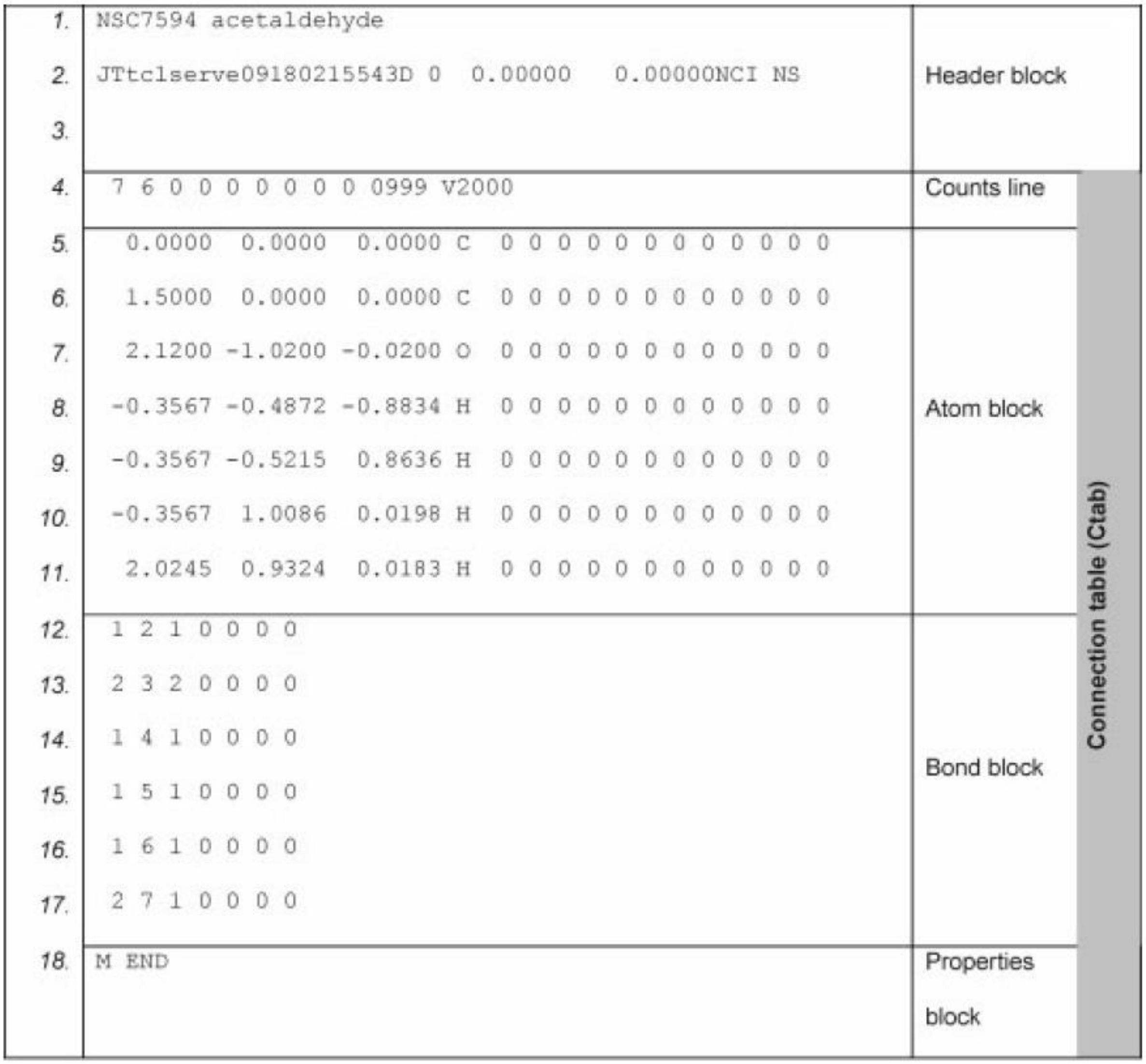
File format	*Suffix*	*Comments*
MDL Molfile	*.mol	Molfile; the most widely used connection table format
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions
SMILES	*.smi	SMILES; the most widely used linear code and file format
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides
CIF	*.cif	Crystallographic Information File format; for 3D structure information on organic molecules
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format
CML	*.cml	Chemical Markup Language; extension of XML with specialization in chemistry

A Molfile describes a single molecular structure which can contain disjointed fragments. An SDfile (SD stands for structure-data) contains structure and data (properties) for any number of molecules, which makes it especially convenient for handling large sets of molecules. All the MDL file formats are referred together as CTfiles.

The structure of the MOLfile

The structure of ethanol molecule and the corresponding MOLfile is given below





1.	NSC7594 acetaldehyde	Header block	
2.	JTtclserve09180215543D 0 0.00000 0.00000NCI NS		
3.			
4.	7 6 0 0 0 0 0 0 0 0999 V2000	Counts line	Connection table (Ctab)
5.	0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0	Atom block	
6.	1.5000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0		
7.	2.1200 -1.0200 -0.0200 O 0 0 0 0 0 0 0 0 0 0 0 0		
8.	-0.3567 -0.4872 -0.8834 H 0 0 0 0 0 0 0 0 0 0 0 0		
9.	-0.3567 -0.5215 0.8636 H 0 0 0 0 0 0 0 0 0 0 0 0		
10.	-0.3567 1.0086 0.0198 H 0 0 0 0 0 0 0 0 0 0 0 0		
11.	2.0245 0.9324 0.0183 H 0 0 0 0 0 0 0 0 0 0 0 0		
12.	1 2 1 0 0 0 0	Bond block	
13.	2 3 2 0 0 0 0		
14.	1 4 1 0 0 0 0		
15.	1 5 1 0 0 0 0		
16.	1 6 1 0 0 0 0		
17.	2 7 1 0 0 0 0		
18.	M END	Properties block	

Each Molfile consists of two parts;

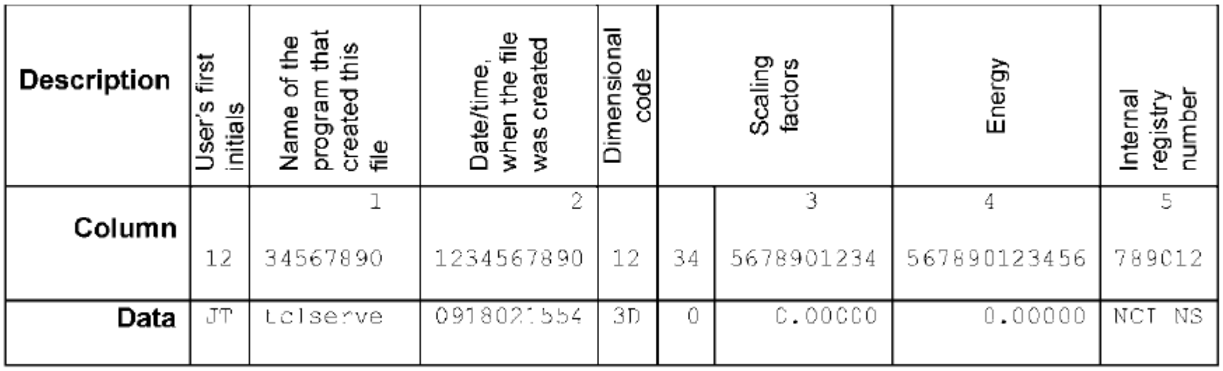
Header Block : (lines 1-3)

Connection Table (Ctab) : (lines 4-18)

The first line of the header block contains the molecule name-No specific format (If no name is available the line is blank)

The second line, however, has a strict format and contains general information about the users name, the programme used to generate this file, and the date and time when the file was created.

a. The date and time information is formed of concatenated two digit values representing the month (09), day (18), year (02), hour (15) and minute (54) respectively.
b. It specifies also whether 2D or 3D atomic coordinates
c. The second line is specified as below.

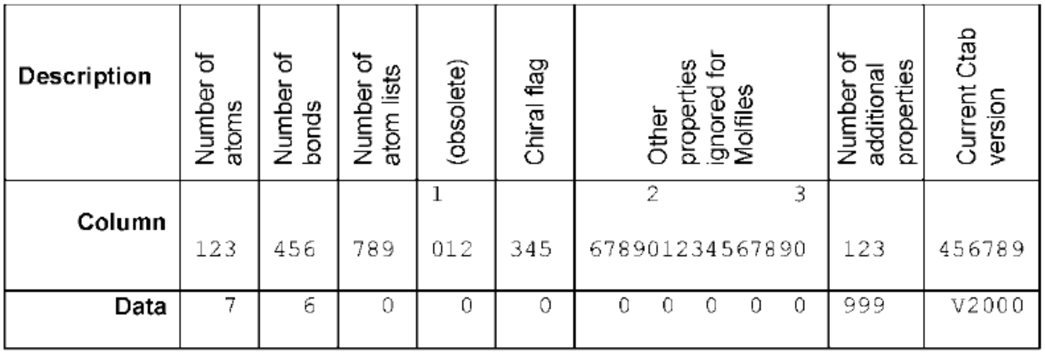


Description	User's first initials	Name of the program that created this file	Date/time, when the file was created	Dimensional code	Scaling factors		Energy	Internal registry number
Column	12	1 34567890	2 1234567890	12	34	3 5678901234	4 567890123456	5 789012
Data	JT	tclserve	0918021554	3D	0	0.00000	0.00000	NCT NS

The third line of the Header Block may be empty or may contain comments.

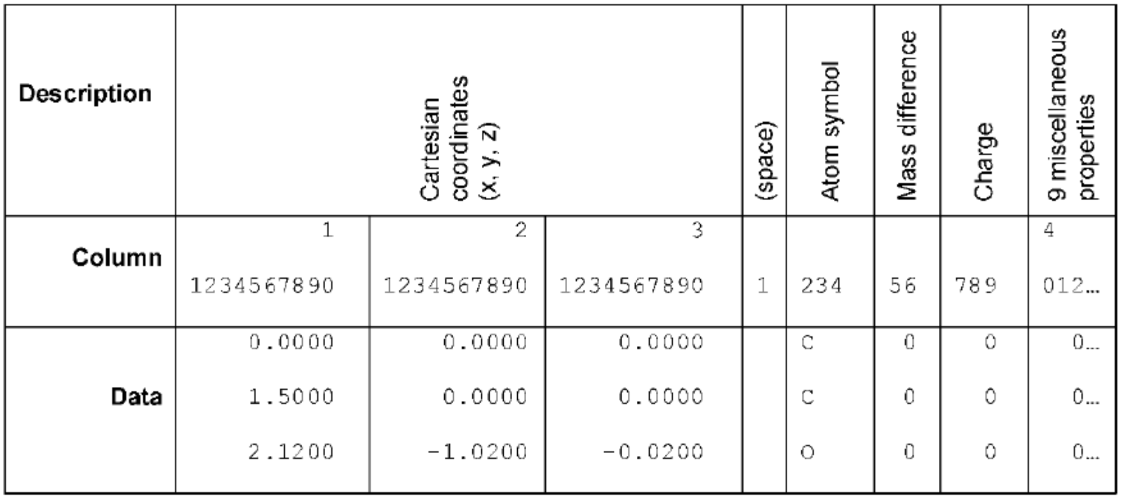
Lines 4-18 form the connection table (Ctab), containing the description of the collection of atoms constituting the given compound, which can be wholly or partially connected by bonds. Such a collection can represent molecules, molecular fragments, substructures, substituent groups, and so on.

a. The first line is called the counts line, specifies how many atoms constitute the molecule represented by this file, how many bonds are within themolecule, whether this molecule is chiral (1 if it is chiral). The last but one entry is always set to 999. The last entry specifies the version of the Ctab (V2000 or V3000).



Description	Number of atoms	Number of bonds	Number of atom lists	(obsolete)	Chiral flag	Other properties ignored for Molfiles	Number of additional properties	Current Ctab version
Column	123	456	789	1 012	345	2 3 678901234567890	123	456789
Data	7	6	0	0	0	0 0 0 0 0	999	V2000

All of the seven atoms declared in the counts line above are described next in an ‘atom block’. Each atom is represented by a single row, which specifies its Cartesian coordinates (2D/3D), atomic symbol, difference from mass in the periodic table, charge and nine other properties, which are usually set to their default values (0s) in Molfiles. 3D atomic coordinates can be recognized in the third column of the atom block. If it contains 0.0, then it is 2D and if it is a different value, it is 3D.

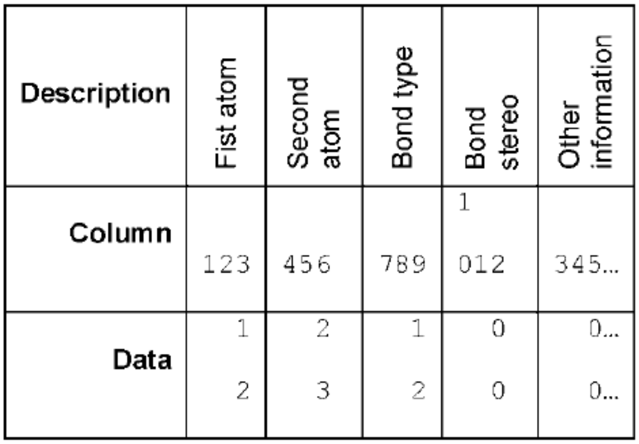


Description	Cartesian coordinates (x, y, z)			(space)	Atom symbol	Mass difference	Charge	9 miscellaneous properties
Column	1 1234567890	2 1234567890	3 1234567890	1	234	56	789	4 012…
Data	0.0000	0.0000	0.0000		C	0	0	0…
	1.5000	0.0000	0.0000		C	0	0	0…
	2.1200	-1.0200	-0.0200		O	0	0	0…

Once the atoms are defined, the bonds between them are specified in a bond block. Each line of this block specifies which two atoms are bonded, the multiplicity of the bond (the bond type entry) and the stereo configuration of the bond(there are also three additional fields that are unused in Molfiles and usually set to 0). The indices of the atoms reflect the order of their appearance in the atom block. In the example analysed “1” relates to the first carbon atom, “2” to second one , “3” to oxygen atom, etc.

The two first lines of the bond block describe the single bond between the two carbon atoms C_1-C_2 and the double bond C_2=O_3, respectively.

The last part of the file is a properties block, which can contain miscellaneous properties.

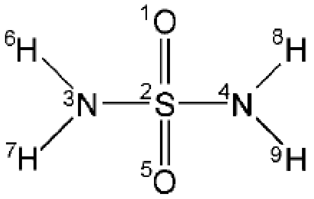


Description	Fist atom	Second atom	Bond type	Bond stereo	Other information
Column	123	456	789	1 012	345...
Data	1 2	2 3	1 2	0 0	0... 0...

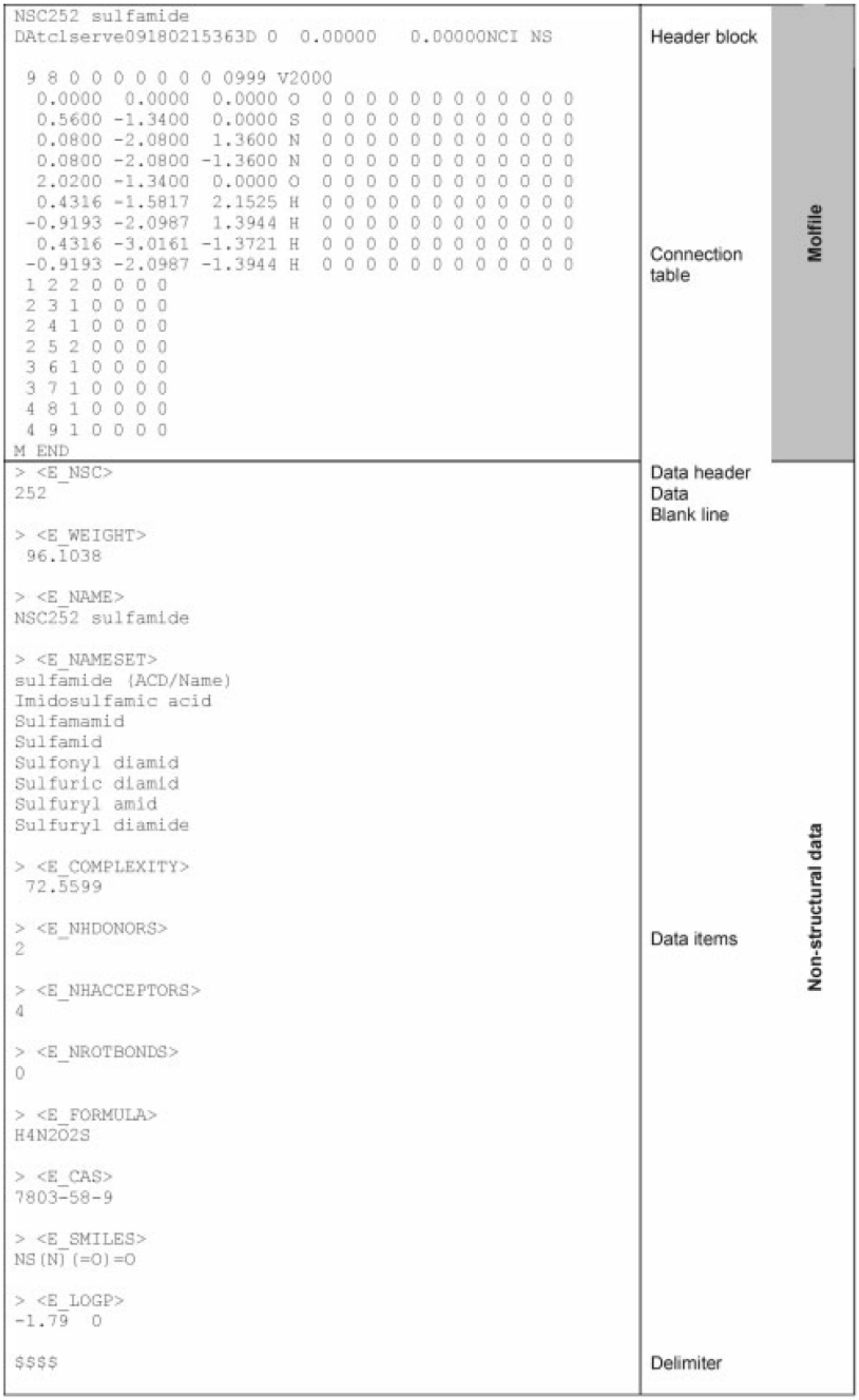
Structure of the bond block

Structure of a SDfile

In the SDfile, each molecule is represented by its Molfile with additional data items describing its non-structural properties (Mol.weight, heat of formation, molecular descriptors, biological activity, etc.). The information on a molecule is terminated by a delimiter line (containing only “$$$$”). Each data item starts with a data header line, which reflects a molecular property name. Next one or more rows contain the actual data; they are terminated by an empty line.



SD file for sulfamide



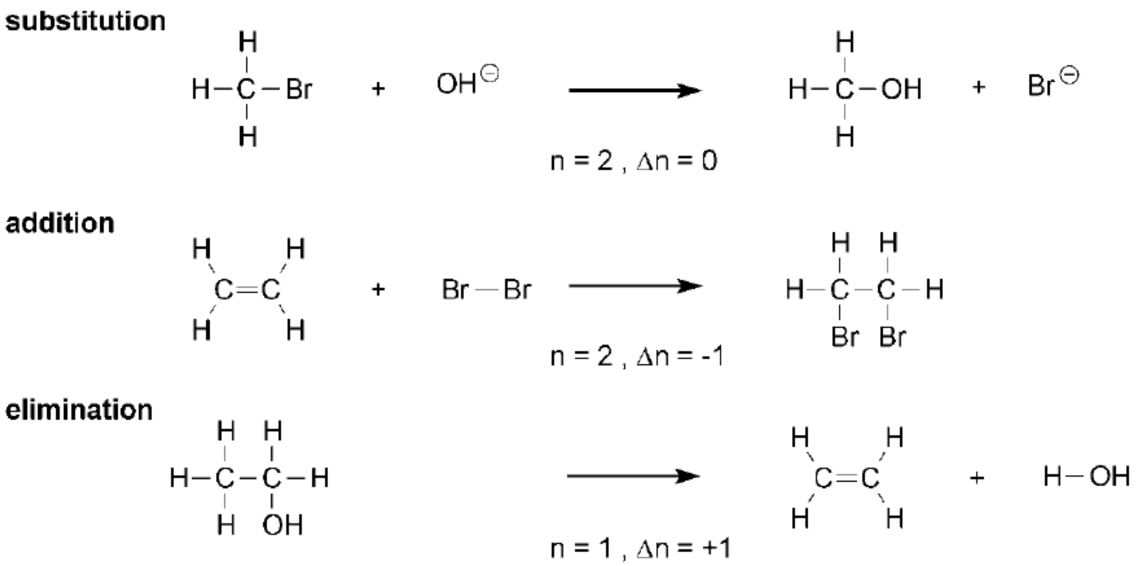
```
NSC252 sulfamide
DAtclserve09180215363D 0   0.00000    0.00000NCI NS

 9 8 0 0 0 0 0 0 0 0999 V2000
  0.0000  0.0000  0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
  0.5600 -1.3400  0.0000 S  0 0 0 0 0 0 0 0 0 0 0 0
  0.0800 -2.0800  1.3600 N  0 0 0 0 0 0 0 0 0 0 0 0
  0.0800 -2.0800 -1.3600 N  0 0 0 0 0 0 0 0 0 0 0 0
  2.0200 -1.3400  0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
  0.4316 -1.5817  2.1525 H  0 0 0 0 0 0 0 0 0 0 0 0
 -0.9193 -2.0987  1.3944 H  0 0 0 0 0 0 0 0 0 0 0 0
  0.4316 -3.0161 -1.3721 H  0 0 0 0 0 0 0 0 0 0 0 0
 -0.9193 -2.0987 -1.3944 H  0 0 0 0 0 0 0 0 0 0 0 0
 1 2 2 0 0 0 0
 2 3 1 0 0 0 0
 2 4 1 0 0 0 0
 2 5 2 0 0 0 0
 3 6 1 0 0 0 0
 3 7 1 0 0 0 0
 4 8 1 0 0 0 0
 4 9 1 0 0 0 0
M END
> <E_NSC>
252

> <E_WEIGHT>
 96.1038

> <E_NAME>
NSC252 sulfamide

> <E_NAMESET>
sulfamide (ACD/Name)
Imidosulfamic acid
Sulfamamid
Sulfamid
Sulfonyl diamid
Sulfuric diamid
Sulfuryl amid
Sulfuryl diamide

> <E_COMPLEXITY>
 72.5599

> <E_NHDONORS>
2

> <E_NHACCEPTORS>
4

> <E_NROTBONDS>
0

> <E_FORMULA>
H4N2O2S

> <E_CAS>
7803-58-9

> <E_SMILES>
NS(N)(=O)=O

> <E_LOGP>
-1.79  0

$$$$
```

Libraries and Toolkits

There are miscellaneous libraries for molecular structure manipulation which support both reading and generating Mol- and SD file formats. OEchem from OpenEye is a commercial library for C++ programmers, which additionally contains links to Python. For the Java programming language there are two intensively developed free libraries, namely JOELib (Java-based re-implementation and extension of the OELib library) and CDK (the Chemical Development Kit) and at least one commercial library- the JChem Library from ChemAxon Ltd.

Another interesting toolkit is SDF toolkit- a set of Perl scripts for manipulating SDFiles. It provides tools for filtering SDFiles, merging them and removing duplicates, adding data from CSV (comma-separated) files to an SD file, and so on.

Reaction Classification



On a more rational basis, reactions can be classified according to the overall change in molecularity, the change Δn in the number of molecules (n) participating in a reaction. A more detailed classification of chemical reactions will give specifications on mechanism of a reaction: electrophilic aromatic substitution, nucleophilic aliphatic substitution, etc. Details on this mechanism can be included to various degrees; thus nucleophilic aliphatic substitutions can further be classified in to SN1 and SN2 reactions. However as reaction conditions such as a change in solvent can shift a mechanism from one type to another, such details are of interest in the discussion of reaction mechanism but less so in reaction classification.

Carbon and hydrogen atoms have similar electron-attracting power, so hydrocarbons have a uniform electron density distribution and no polarity. Heteroatom such as oxygen, nitrogen or halogen atoms has a higher electron attracting power (higher electronegativity) and introduces polarity into organic compounds. A simple picture expressing this notion is drawn by showing these atoms bearing a partial negative charge; consequently, the atoms to which they are bonded carry a partial positive charge. Reagents with sites of high electron density (nucleophilic agents) seek atoms with low electron density and conversely with low electron density (electrophilic agents bind to atoms with high electron density. Quantum mechanical methods have been developed to assign quantitative values to the partial charges of the atoms in a molecule. This opens the door to defining chemical reactivity on a more quantitative basis.

Inductive effect

The polarizing influence of an electronegative atom decreases with the number of intervening σ – bonds. It is generally accepted that the inductive effect is attenuated by a factor of 2-3 by each intervening bond. The inductive effect is not only operative in the ground state of a molecule but also exerts its influence when bonds are broken heterolytically. Then electronegative atoms adjacent to the reacting bond will stabilize incipient negative charges and vice-versa. It has been shown that residual electronegativity as calculated by the PEOE method.

I

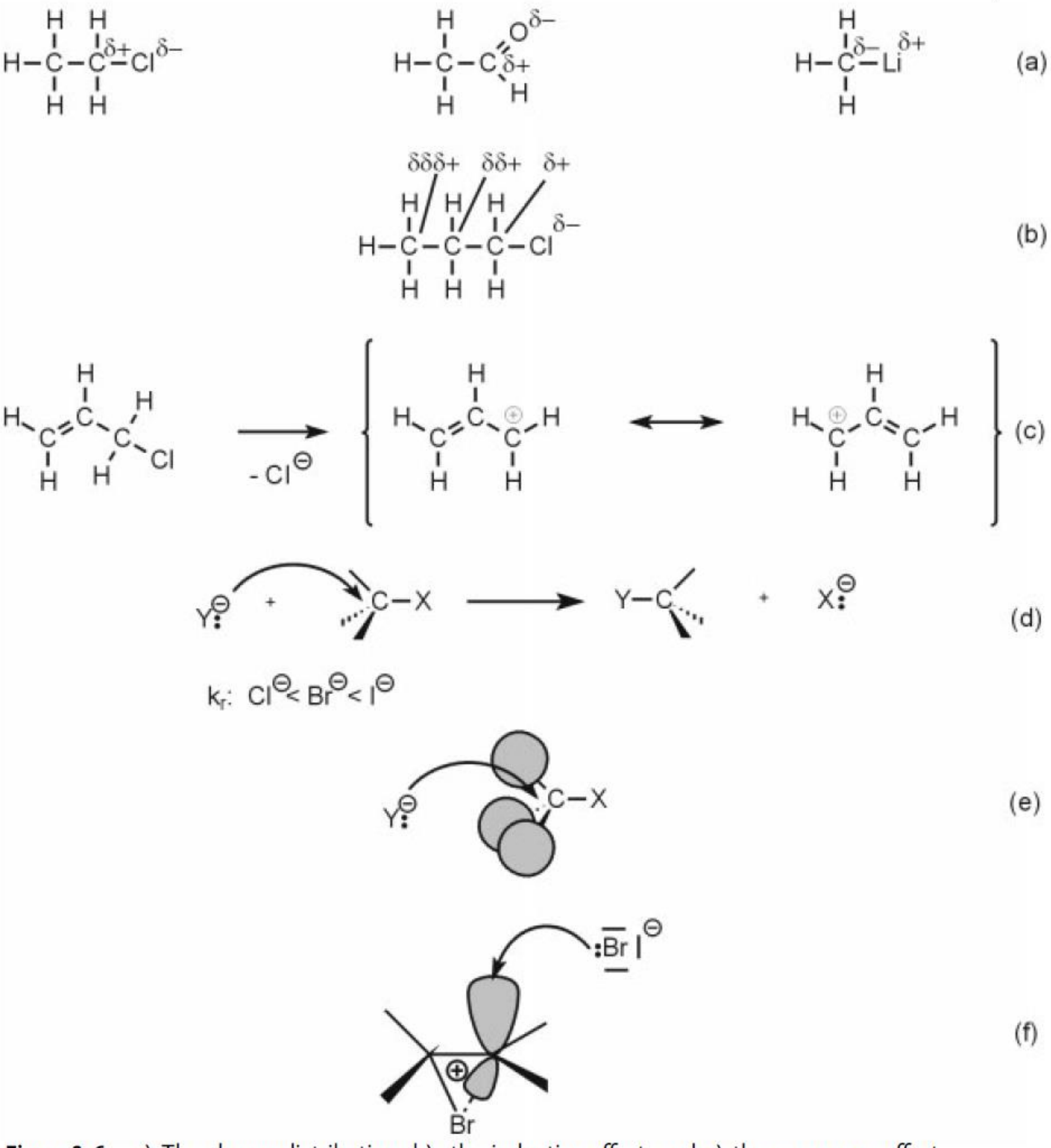


Figure 3-6. a) The charge distribution, b) the inductive effect, and c) the resonance effect, d) the polarizability effect, e) the steric effect, and f) the stereoelectronic effect.

Resonance effect:

The charges that are generated on heterolysis, on polar breaking of a bond, can also be stabilized by delocalization, as observed in conjugated π-systems. This is called the resonance or mesomeric effect and it usually has an even greater influence on chemical reactivity than the inductive effect. The RAMSES data structure is particularly suited for determining conjugated π-systems.

Polaraizability effect

An electric field induces a dipole in a molecule, the magnitude of this dipole is proportional to the polarizability of the molecule, which is measured by the mean molecular polarizability due to motion-averaging. In chemical reactions, charges within a molecule may be generated by bond breaking or bond formation. Atoms that have a high polarizability can stabilize such charges, but they do so more the closer they are to these charges. Furthermore groups that have more polarizable electrons can better provide such electrons to the reacting complex.

The representation of a chemical reaction should include the connection table of all participating species (starting materials, reagents, solvents, catalysts, products), as well as reaction conditions and observations.

However, reactions are only insufficiently represented by the structure of their starting materials and products.

It is essential to indicate also the reaction center and the bonds broken and made in a reaction- in essence, to specify how electrons are shifted during a reaction. In this sense, a representation of chemical reactions should consider some essential features of a reaction mechanisms.

Reaction classification serves to combine several reaction instances into one reaction type. In this way, a vast number of observed chemical reactions is reduced to a manageable number of reaction types.

There are two approaches

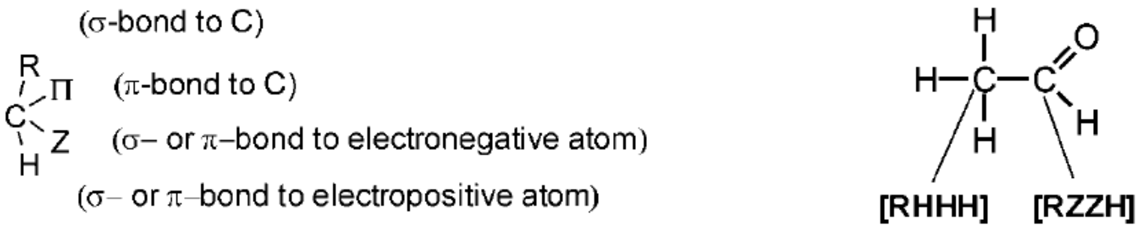
1. Model-driven approaches (classify reactions according to preconceived model, a conceptual framework)

2. Data driven approaches

Hendrickson's scheme

It concentrated mainly on C-C bond forming reactions because the construction of the carbon atom skeleton is the major task in the synthesis of complex organic compounds. Each carbon atom is classified according to which kind of atoms are bonded to it and what kind of bonds (σ or π) are involved. The number of bonds to R, π, Z and H atoms is given by the numbers σ, π, z and h respectively. For any uncharged carbon atom the following equation must hold

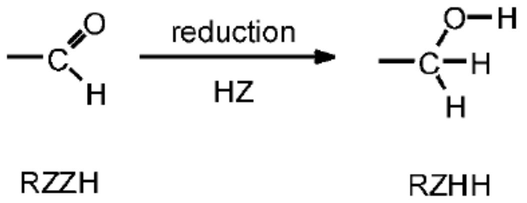
$\sigma + \pi + z + h = 4$


(σ-bond to C)
R
Π
C
Z
H
(π-bond to C)
(σ– or π–bond to electronegative atom)
(σ– or π–bond to electropositive atom)
H
H—C—C
O
H
H
[RHHH]
[RZZH]

Hendrickson's classification of atom types

Unit reactions at each carbon atom are then composed of unit exchanges of one bond type against another. There are 16 such exchanges possible at one carbon atom, each denoted by two letters, the first one for the bond made and the second for the bond broken.

Skeletal changes are characterized by changes in R, with constructions having positive values (+R) and fragmentation negative (-R); functionality changes have +/- π, +/- z, +/- H.



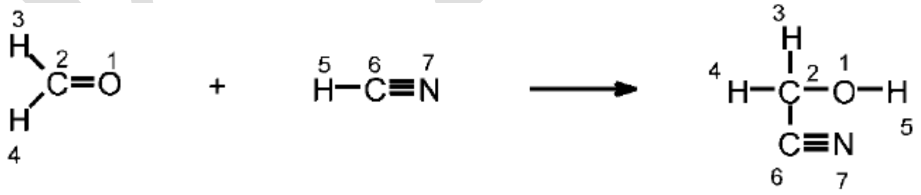
Example of a unit exchange in a reaction.

Table: Possible unit exchanges at any skeletal carbon atom.

		Δx	$\Delta \pi$	$\Delta \sigma$
Substitution	HH, ZZ, RR, ΠΠ	0	0	0
Oxidation	ZH	+2	0	0
Reduction	HZ	–2	0	0
Elimination	ΠH	+1	–1	0
	ΠZ	–1	+1	0
Addition	HΠ	–1	–1	0
	ZΠ	+1	–1	0
Construction	RH	–1	0	+1
	RZ	–1	0	+1
	RΠ	0	–1	+1
Fragmentation	HR	–1	0	–1
	ZR	+1	0	–1
	ΠR	0	+1	–1

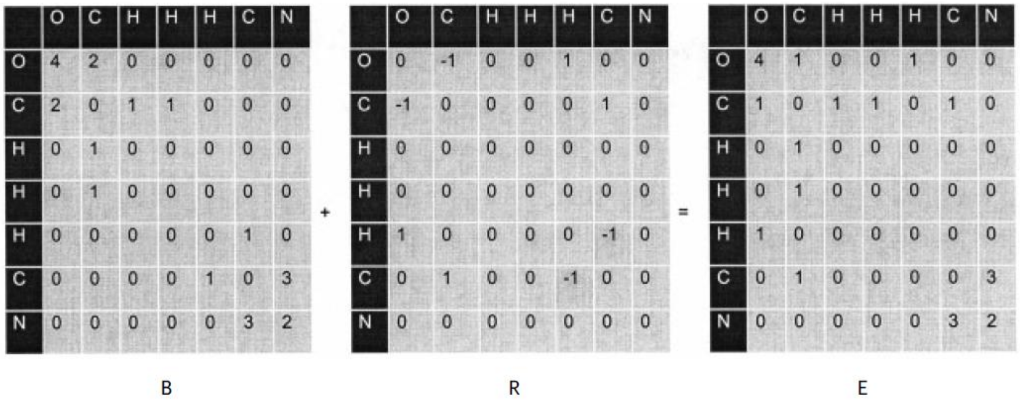
Ugi's scheme

A scheme based on treating reactions by means of matrices- reaction (R-) matrices. The representation of chemical structures by bond and electron (BE-) matrices for single molecule, ensembles of them such as the starting materials (eg. formaldehyde, hydrogen cyanide) and the reaction product (cyanohydrin of formaldehyde) may be constructed.



Having the BE-matrices of the beginning, **B** and the end **E** of a reaction one can calculate

E – **B** = **R** . As can easily be seen, the entries r_{ij} in the R-matrix indicate the bonds broken and made in the course of this reaction.



	O	C	H	H	H	C	N
O	4	2	0	0	0	0	0
C	2	0	1	1	0	0	0
H	0	1	0	0	0	0	0
H	0	1	0	0	0	0	0
H	0	0	0	0	0	1	0
C	0	0	0	0	1	0	3
N	0	0	0	0	0	3	2

B

+

	O	C	H	H	H	C	N
O	0	-1	0	0	1	0	0
C	-1	0	0	0	0	1	0
H	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0
H	1	0	0	0	0	-1	0
C	0	1	0	0	-1	0	0
N	0	0	0	0	0	0	0

R

=

	O	C	H	H	H	C	N
O	4	1	0	0	1	0	0
C	1	0	1	1	0	1	0
H	0	1	0	0	0	0	0
H	0	1	0	0	0	0	0
H	1	0	0	0	0	0	0
C	0	1	0	0	0	0	3
N	0	0	0	0	0	3	2

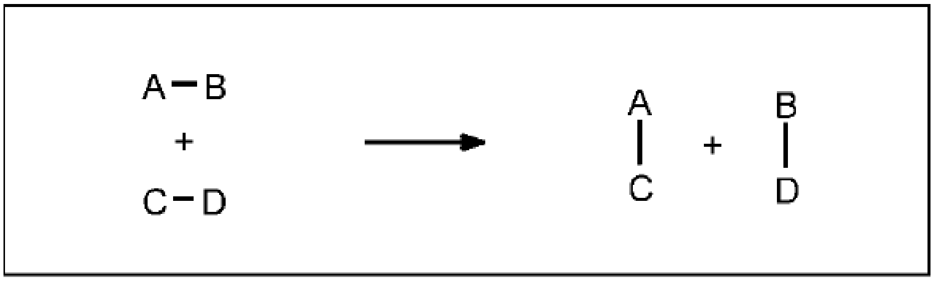
E

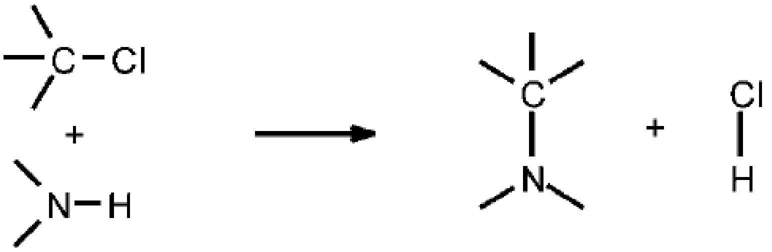
The merit of this mathematical model is due to

1. The representation of chemical species should take account of all valence electrons
2. Reactions should be represented by the shifting of bonds and electrons in the reaction center.

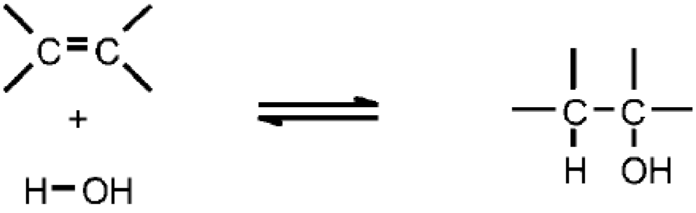
An R-matrix expresses the bond and electron rearrangement in a reaction. It reflects a reaction scheme, the breaking and making of two bonds, that is at the foundation of the majority of all organic reactions.

Eg.

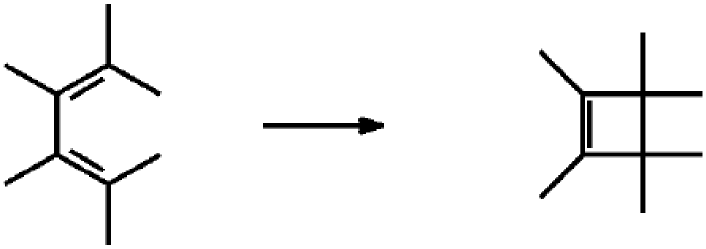




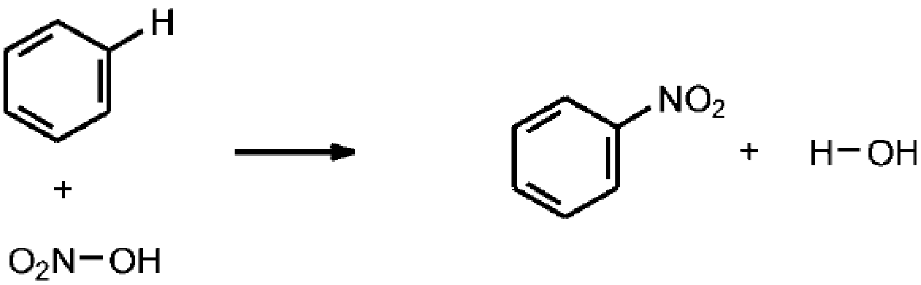
(a)



(b)



(c)

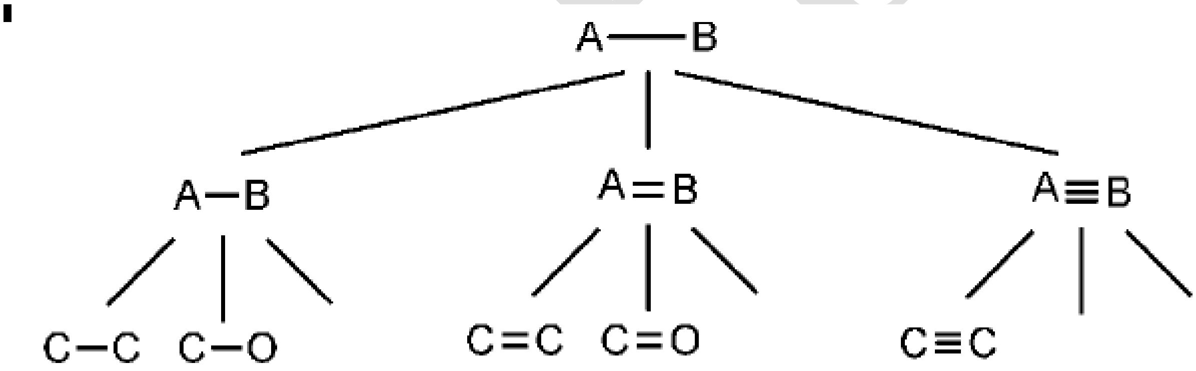


(d)

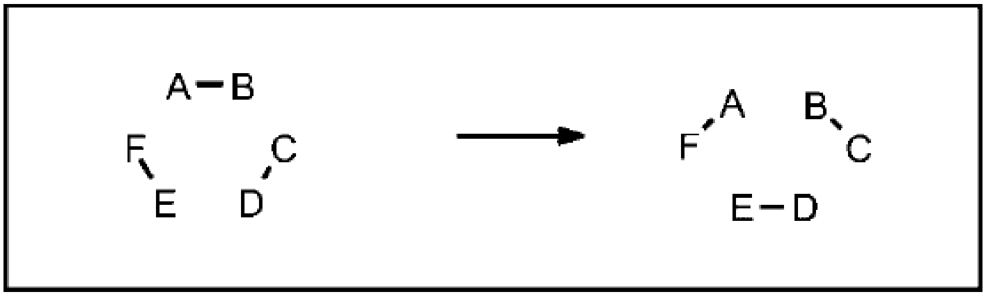
The above reaction scheme consists of

a. Aliphatic substitutions
b. Additions to multiple bonds
c. Electrocyclic reactions
d. Electrophi;lic aromatic substitutions

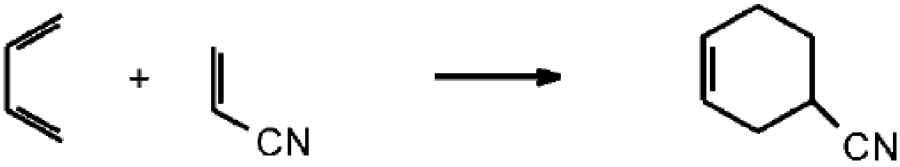
Concentration on the types of bonds broken or made in a reaction provides a basis fopr reaction classification. First let us consider this only for one bond. On the first level of hierarchy, a bond can be distinguished by whether it is a single double or triple bond. Then on the next level, a further distinction can be made on the basis of the atoms that comprise the bond.



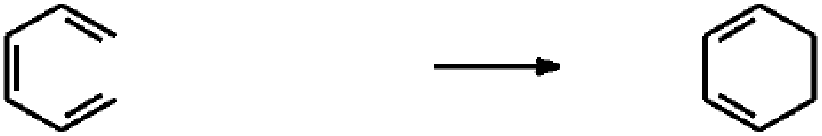
Different levels of specification for a bond participating in a reaction.



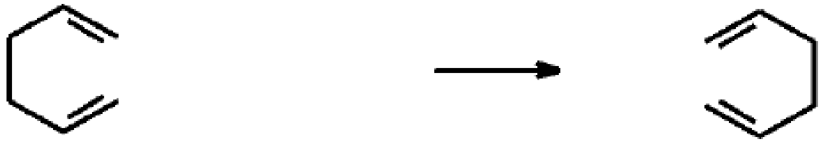
Diels Alder reaction



electrocyclic ring closure



Cope rearrangement



Favorskii rearrangement

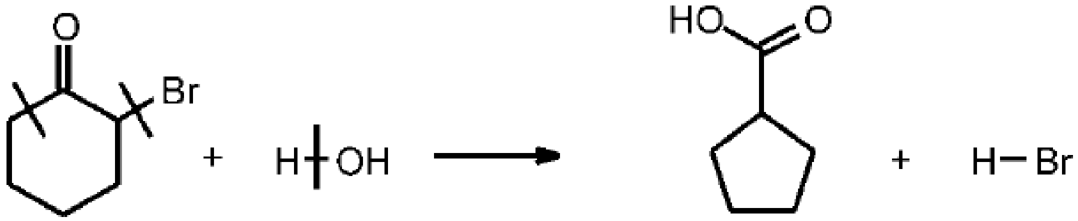


Fig: The reaction scheme breaking three and making three bonds, and some of the reaction type that fall into this scheme.

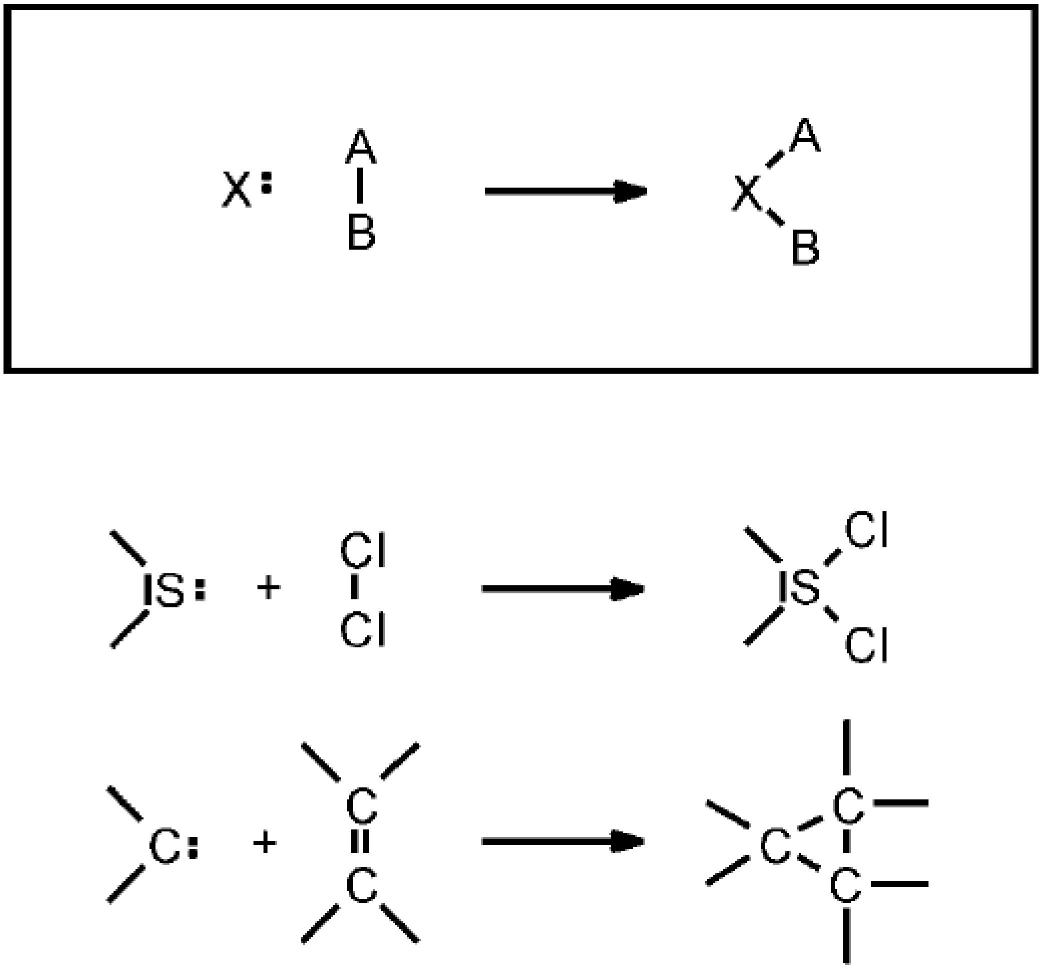


Fig: A reaction scheme that changes the number of bonds at one atom and some specific examples.

InfoChem’s Reaction Classification

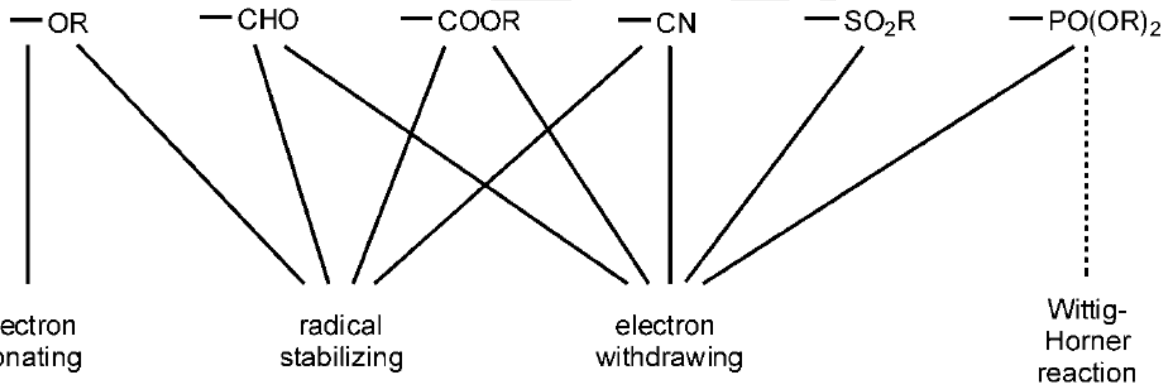
MDL information systems algorithm considers, for each atom of the reaction center, the atom type, valence state, total number of bonded hydrogen atoms, number of π-electrons, aromaticity and formal charge. These pieces of information are merged into a hash code.

Data-driven approaches

The data-driven methods try to derive a classification from the data presented.

a. Horace

Based on the functionalities attached to the reaction center, the method of conceptual clustering derived the features a reaction needed to posses for it to be assigned to a certain reaction type. A functional group can have different effects, depending on the mechanism of a reaction type and depending on the electron on the reaction center.



Possible Questions

Part A (Online multiple choice questions)

Part B (Each question carries two marks)

1. Write the different ways of expressing the molecular formula of phenylalanine
2. What are the advantages and disadvantages of trival names of compounds
3. What are the advantages and disadvantages of IUPAC nomenclature in naming of the compounds
4. What is meant by a line notation
5. What are the different line notations for the structure diagram of phenylalanine
6. What is meant by matrix representation
7. What are the different types of matrix representation of a structure.
8. Explain adjacency matrix of a molecule
9. What is meant by a distance matrix of a molecule
10. What is meant by incidence matrix of a molecule
11. What is meant by bond matrix
12. Explain the libraries and toolkits available for structure representations
13. What are the different approaches of reaction classification in cheminformatics.

Part C (Each question carries eight marks)

1. Explain the rules to be followed in the nomenclature of inorganic and organic compounds
2. What is meant by a line notation. Explain the different types of line notation.
3. Explain the Wiswesser Line Notation (WLN) in detail
4. Explain ROSDAL line notation in detail
5. Explain SMILES line notation in detail
6. Differentiate SMILES and Sybyl Line Notation (SLN) line notations
7. Explain in detail Sybyl Line Notation (SLN).
8. What is meant by a matrix representation. Explain in detail the different types of matrix representations with suitable examples
9. What are the advantages and the disadvantages of the different types of matrix representation of chemical structures.
10. Differentiate adjancy matrix and distance matrix with suitable examples
11. Differentiate incidence matrix and adjancy matrix with suitable examples
12. Explain the structure if a MOLfile in detail.
13. Explain in detail the structure of a SDfile
14. Explain the different types of polar effects in a reaction
15. How reactions are classified in cheminformatics
16. Explain in detail the Hendrickson's scheme of reaction classification
17. Explain in detail the Ugis scheme of reaction classification

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore – 641 021.

UNIT II (Multiple choice Questions)

S.No	Question	A	B	C	D	Answer
1	In the Nomenclature of Inorganic compounds The first named element	Oxygen	Carbon	Electropositive element	Electronegative element	Electropositive element
2	In the nomenclature of inorganic compounds, the stoichiometry of the elements is indicated at	the lower right hand side by index numbers.	the lower left hand side by index numbers.	As a superscript	As a fractional number	The lower right hand side by index numbers.
3	In the nomenclature of inorganic compounds the charges of ions are placed at the top right-hand side next to the element symbol	the top right-hand side next to the element symbol	the lower right hand side by index numbers.	the lower left hand side by index numbers.	As a superscript	the top right-hand side next to the element symbol
4	In ions of complexes, the central atom is specified	After the ligands and are listed in alphabetical order,	before the ligands and are listed in alphabetical order	before the ligands are listed in alphabetical order,	By a neutral atom	before the ligands and are listed in alphabetical order
5	In the Nomenclature of organic compounds	Carbon and hydrogen atoms	Hydrogen and Carbon atoms	Hetroatoms are named first	Heteroatoms and carbon atoms were positioned in	Carbon and hydrogen atoms were positioned

		were positioned in the first and second places respectively	were positioned in the first and second places respectively		the first and second places respectively	in the first and second places respectively
6	IUPAC name of phenylalanine is	3-amino-3-phenylpropanoic acid	3-amino-2-phenylpropanoic acid	2-amino-3-phenylpropanoic acid	phenylpropanoic acid	2-amino-3-phenylpropanoic acid
7	Carbon atoms are shown in the notation only by digits in	WLN notation	SMLES notation	ROSDAL notation	SLN notation	ROSDAL notation
8						
9	Other than carbon atoms carry digits, in addition, their atomic symbol.	WLN notation	SMILES notation	ROSDAL notation	SLN notation	ROSDAL notation
10	Atoms are represented by their atomic symbols	WLN notation	SMILES notation	ROSDAL notation	SLN notation	SMILES notation
11	Rings are described by allocating digits to the two connecting ring atoms.	WLN notation	SMILES notation	ROSDAL notation	SLN notation	SMILES notation
12	The WLN code for phenyl alanine is	VQYZ1R	1O-2=3O, 2-4-5N, 4-6-7=-12-7	NC(Cc1ccccc1)C(O)=O	C[1]H:CH:CH:CH:CH:C(:@1) CH2CH(NH2)C(=O)OH	VQYZ1R
13	SMILES notation for phenylalanine	VQYZ1R	1O-2=3O, 2-4-5N, 4-6-7=-12-7	NC(Cc1ccccc1)C(O)=O	C[1]H:CH:CH:CH:CH:C(:@1) CH2CH(NH2)C(=O)OH	NC(Cc1ccccc1)C(O)=O
14	ROSDAL notation for phenyalanine	VQYZ1R	1O-2=3O, 2-4-5N, 4-6-7=-12-7	NC(Cc1ccccc1)C(O)=O	C[1]H:CH:CH:CH:CH:C(:@1)	1O-2=3O, 2-4-5N, 4-6-7=-12-7

					CH2CH(NH2)C(=O)OH	
15	SLN notation for phenylalanine	VQYZ1R	1O-2=3O, 2-4-5N, 4-6-7=-12-7	NC(Cc1ccccc1)C(O)=O	C[1]H:CH:CH:CH:CH:C(:@1) CH2CH(NH2)C(=O)OH	C[1]H:CH:CH:CH:CH:C(:@1) CH2CH(NH2)C(=O)OH
16	VQYZ1R belongs to the phenylalanine according to	WLN notation	SMILES notation	ROSDAL notation	SLN notation	WLN notation
17	1O-2=3O, 2-4-5N, 4-6-7=-12-7 belongs to the phenylalanine according to	WLN notation	SMILES notation	ROSDAL notation	SLN notation	ROSDAL notation SMILES notation
18	NC(Cc1ccccc1)C(O)=O belongs to the phenylalanine according to	WLN notation	SMILES notation	ROSDAL notation	SLN notation	SMILES notation
19	C[1]H:CH:CH:CH:CH:C(:@1) CH2CH(NH2)C(=O)OH belongs to the phenylalanine according to	WLN notation	SMILES notation	ROSDAL notation	SLN notation	SLN notation
20	The matrix of a molecule consisting of n atoms in a square (n x) matrix with the entries giving all the connectivities of the atoms.	adjacency matrix	distance matrix	incidence matrix	bond matrix	adjacency matrix
21	The elements of a matrix contain values	adjacency matrix	distance matrix	incidence matrix	bond matrix	distance matrix

	which specify the shortest distance between the atoms involved.					
22	The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m).	adjacency matrix	distance matrix	incidence matrix	bond matrix	incidence matrix
23	The matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms	adjacency matrix	distance matrix	incidence matrix	bond matrix	bond matrix
24	adjacency matrix is	The matrix of a molecule consisting of n atoms in a square (n x) matrix with the entries giving all the connectivities of the atoms.	The elements of a matrix contain values which specify the shortest distance between the atoms involved.	The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m).	The matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms	The matrix of a molecule consisting of n atoms in a square (n x) matrix with the entries giving all the connectivities of the atoms.
25	distance matrix is	The matrix of a molecule consisting of n atoms in a square (n x) matrix with the	The elements of a matrix contain values which specify the shortest distance	The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond	The matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms	The elements of a matrix contain values which specify the shortest distance between the atoms involved.

		entries giving all the connectivities of the atoms.	between the atoms involved.	to the rows (m).		
26	incidence matrix	The matrix of a molecule consisting of n atoms in a square (n x) matrix with the entries giving all the connectivities of the atoms.	The elements of a matrix contain values which specify the shortest distance between the atoms involved.	The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m).	The matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms	The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m).
27	bond matrix is	The matrix of a molecule consisting of n atoms in a square (n x) matrix with the entries giving all the connectivities of the atoms.	The elements of a matrix contain values which specify the shortest distance between the atoms involved.	The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m).	The matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms	The matrix is related to the adjacency matrix but gives information also on the bond order of the connected atoms
28	Matrices in which all elements are shown twice are called	Redundant matrix	Non-redudant matrix	bond matrix	incidence matrix	Redundant matrix
29	A matrix contains each element only once	Redundant matrix	Non-redudant matrix	bond matrix	incidence matrix	Non-redudant matrix
30	Redundant matrix is	Matrices in which all elements are	A matrix contains each element only	Describes connections and bonds	Describes geometric distances	Matrices in which all elements are shown twice

		shown twice	once			
31	Non-redudant matrix	Matrices in which all elements are shown twice	A matrix contains each element only once	Describes connections and bonds	Describes geometric distances	A matrix contains each element only once
32	A Molfile describes	a single molecular structure which can contain disjointed fragments	contains structure and data (properties) for any number of molecules	A matrix contains each element only once	Matrices in which all elements are shown twice	a single molecular structure which can contain disjointed fragments
33	SDfile describes	a single molecular structure which can contain disjointed fragments	contains structure and data (properties) for any number of molecules	A matrix contains each element only once	Matrices in which all elements are shown twice	contains structure and data (properties) for any number of molecules
34	The first line of the Header Block of the Molfile contains	molecule name	contains general information about the users and whether 2D or 3D atomic coordinates	empty or may contain comments	connection table (Ctab)	molecule name
35	The Second line of the Header Block of the Molfile contains	molecule name	contains general information about the users and whether 2D or 3D atomic coordinates	empty or may contain comments	connection table (Ctab)	contains general information about the users and whether 2D or 3D atomic coordinates
36	The Third line of the Header Block of the Molfile contains	molecule name	contains general information about the users	empty or may contain comments	connection table (Ctab)	empty or may contain comments

			and whether 2D or 3D atomic coordinates			
37	The fourth line of the Header Block of the Molfile contains	molecule name	contains general information about the users and whether 2D or 3D atomic coordinates	empty or may contain comments	connection table (Ctab)	connection table (Ctab)
38	Each molecule is represented by its Molfile with additional data items describing its non-structural properties is called	SDfile	Molfile	RD file	PDB file	SDfile
39	PDB file consists of	3D structure information of protein	a single molecular structure which can contain disjointed fragments	contains structure and data (properties) for any number of molecules	A matrix contains each element only once	3D structure information of protein
40	CIF file consists of	Crystallographic information of a molecule	3D structure information of protein	a single molecular structure which can contain disjointed fragments	contains structure and data (properties) for any number of molecules	Crystallographic information of a molecule
41	In the MOLfile, it specifies how many atoms constitute the molecule represented by this file	counts line	Atom block	Bond block	Header block	counts line
42	The polarizing influence of an electronegative	Inductive effect	Resonance effect	Polarazability effect	Electromeric effect	

	atom					
43	The charges can also be stabilized by delocalization, as observed in conjugated π-systems.	Inductive effect	Resonance effect	Polarazability effect	Electromeric effect	Resonance effect
44	An electric field induces a dipole in a molecule	Inductive effect	Resonance effect	Polarazability effect	Electromeric effect	Polarazability effect
45	Inductive effect	The polarizing influence of an electronegative atom	The charges can also be stabilized by delocalization, as observed in conjugated π-systems	An electric field induces a dipole in a molecule	The effect present in polymers	The polarizing influence of an electronegative atom
46	Resonance effect is	The polarizing influence of an electronegative atom	The charges can also be stabilized by delocalization, as observed in conjugated π-systems	An electric field induces a dipole in a molecule	The effect present in polymers	The charges can also be stabilized by delocalization, as observed in conjugated π-systems
47	Polarazability effect is	The polarizing influence of an electronegative atom	The charges can also be stabilized by delocalization, as observed in conjugated π-systems	An electric field induces a dipole in a molecule	The effect present in polymers	An electric field induces a dipole in a molecule
48	Hendrickson's scheme	Model-driven approaches	Data driven	Structure driven approach	Target driven approach	Model-driven approaches

	for reaction classification is a		approache			
49	Ugi's scheme for reaction classification is a	Model-driven approaches	Data driven approache	Structure driven approach	Target driven approach	Model-driven approaches
50	InfoChem's Reaction Classification is a	Model-driven approaches	Data driven approache	Structure driven approach	Target driven approach	Model-driven approaches
51	Hendrickson's scheme	It concentrated mainly on C-C bond forming reactions	based on treating reactions by means of matrices- reaction (R-) matrices	considers, for each atom of the reaction center, the atom type, valence state,	Based on the functionalities attached to the reaction center	It concentrated mainly on C-C bond forming reactions
52	Ugi's scheme for reaction classification	It concentrated mainly on C-C bond forming reactions	based on treating reactions by means of matrices- reaction (R-) matrices	considers, for each atom of the reaction center, the atom type, valence state,	Based on the functionalities attached to the reaction center	based on treating reactions by means of matrices- reaction (R-) matrices
53	InfoChem's Reaction Classification	It concentrated mainly on C-C bond forming reactions	based on treating reactions by means of matrices- reaction (R-)	considers, for each atom of the reaction center, the atom type, valence state,	Based on the functionalities attached to the reaction center	considers, for each atom of the reaction center, the atom type, valence state,

			matrices			
54	It concentrated mainly on C-C bond forming reactions	Hendrickson’s scheme of reaction classification	Ugi’s scheme for reaction classification	InfoChem’s Reaction Classification	The data-driven approach of Reaction classification is	Hendrickson’s scheme of reaction classification
55	It is based on treating reactions by means of matrices- reaction (R-) matrices	Hendrickson’s scheme of reaction classification	Ugi’s scheme for reaction classification	InfoChem’s Reaction Classification	The data-driven approach of Reaction classification is	Ugi’s scheme for reaction classification
56	It considers, for each atom of the reaction center, the atom type, valence state	Hendrickson’s scheme of reaction classification	Ugi’s scheme for reaction classification	InfoChem’s Reaction Classification	The data-driven approach of Reaction classification is	InfoChem’s Reaction Classification
57	It is based on the functionalities attached to the reaction center	Hendrickson’s scheme of reaction classification	Ugi’s scheme for reaction classification	InfoChem’s Reaction Classification	The data-driven approach of Reaction classification is	The data-driven approach of Reaction classification is
58	The data-driven approach of Reaction classification is	It concentrated mainly on C-C bond forming reactions	based on treating reactions by means of matrices- reaction (R-) matrices	considers, for each atom of the reaction center, the atom type, valence state	Based on the functionalities attached to the reaction center	Based on the functionalities attached to the reaction center
59	An R-matrix	expresses the bond and electron rearrangement in a reaction	considers, for each atom of the reaction center, the atom type, valence state	Based on the functionalities attached to the reaction center	concentrated mainly on C-C bond forming reactions	expresses the bond and electron rearrangement in a reaction
60	Branches are indicated in parenthesis	WLN notation	SMILES notation	ROSDAL notation	SLN notation	SLN notation

Searching chemical structures: Full structure search, sub-structure search, basic ideas, similarity search, three dimensional search methods, basics of computation of physical and chemical data and structure descriptors, data visualization.

Searching chemical structures

Large chemical databases, combinatorial libraries and data warehouses have become indispensable tools in modern chemical research. Accordingly information must be stored in these databases and searched in an appropriate manner.

Methods have been developed for enabling the computer to perceive both complete chemical structures and fragments of them, as well as their mutual similarity.

Full structure search

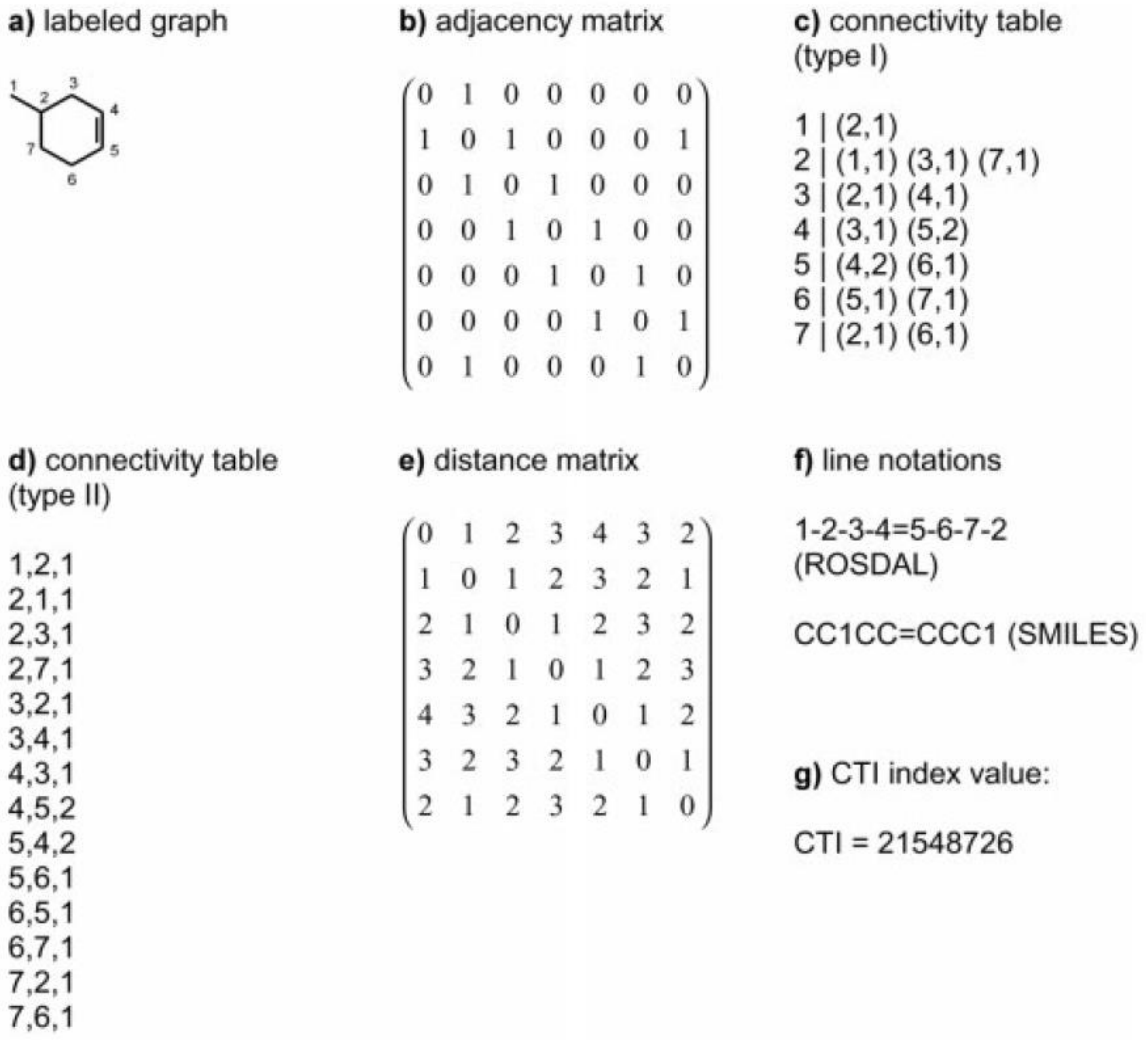
Various approaches have been devised for full structure search. They comprise the use of molecular formulas, molecular weights, trade and /or trival names, various line notations, registry numbers, constitutional diagrams (2D representations), atom coordinates (2D or 3D representations), topological indices, hash codes and others.

Empirical molecular formulas and molecular weights usually identify a whole class of compounds (chemical isomers) rather than a single structure. Furthermore, millions of structures might correspond to each molecular formula, i.e. they are highly degenerate. Hence, they are usually used as supplementary descriptors.

Trade names can be stored and searched as character strings. Their use is the simplest and more suitable way of storing chemical information. However being not subject to strict rules, their formation does not reflect accurately molecular composition.

Plenty of line notations are used and in particular SMILES notation is in widespread use among chemists. Consisting of character strings, these representations are compact and easy to use. Their creation is subject to different rules: a search procedure using them consists of a simple comparision of two strings.

I



a) labeled graph

b) adjacency matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

c) connectivity table (type I)

1 | (2,1)
2 | (1,1) (3,1) (7,1)
3 | (2,1) (4,1)
4 | (3,1) (5,2)
5 | (4,2) (6,1)
6 | (5,1) (7,1)
7 | (2,1) (6,1)

d) connectivity table (type II)

1,2,1
2,1,1
2,3,1
2,7,1
3,2,1
3,4,1
4,3,1
4,5,2
5,4,2
5,6,1
6,5,1
6,7,1
7,2,1
7,6,1

e) distance matrix

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 3 & 2 \\ 1 & 0 & 1 & 2 & 3 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 & 3 & 2 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 3 & 2 & 1 & 0 & 1 \\ 2 & 1 & 2 & 3 & 2 & 1 & 0 \end{pmatrix}$$

f) line notations

1-2-3-4=5-6-7-2
(ROSDAL)

CC1CC=CCC1 (SMILES)

g) CTI index value:

CTI = 21548726

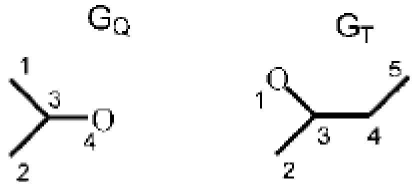
Fig: Different forms of representations of a chemical graph.

Substructure Search

A substructure search algorithm is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures such as identification of equivalent atoms, determination of maximal common substructure, ring detection, calculation of topological indices, etc.

The search for structural fragments (substructures) is very important in medicinal chemistry, QSAR, spectroscopy and many other fields in the process of pharmacophore, chromophore or other –phore perceptions.

Substructure searching is the process of identifying parts of a given structure that are equivalent to a specified query substructure. In graph-theoretical terms substructure searching is the task of checking whether the query graph (GQ) is isomorphic with a subgraph of another target graph (GT). Sometimes the target graph is called a reference graph.



There are several basic strategies for the improvement of the performance of substructure search algorithms.

1. Optimisation of the hardware and software technologies used
2. Usage of various methods to improve the perception of the substructure isomorphic with the query graph or with the rejection of inappropriate target structure candidates as early as possible.
3. Pre-processing of the most time-consuming operations that are independent of the query structure and storing them as an integral part of the data base which can be used at search time.

Backtracking Algorithm

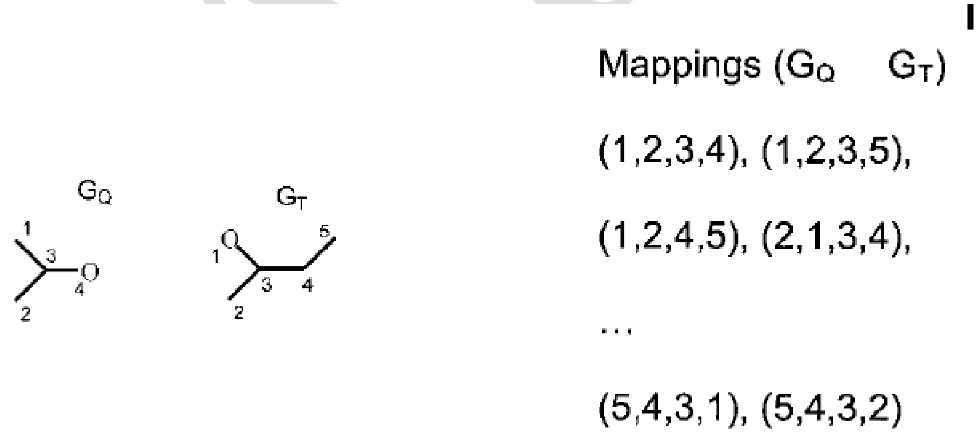
The basic approach to a fast search of an isomorphism among all mappings is so called backtracking algorithm. It searches an isomorphism GQ---GT in such a way that all checked mappings can be organized hierarchically as a tree. Very often this approach is named atom-by-atom searching, since at every step a new single atom is mapped. The backtracking algorithm is the core part of every software system that performs substructure searching.

The optimization of the backtracking algorithm

The optimization of the backtracking algorithm usually consists of an application of several heuristics which reduce the number of candidates atoms for mapping from GQ ---GT. These heuristics are based on local properties of the atoms such as atom types, number of bonds, bond orders and ring membership. According to these properties the atoms in GQ and GT are separated into different classes. This step is known as partitioning.

Table 6.1. Application of partitioning approach for substructure search optimization. According to their local properties, the atoms of the graphs in Figure 6-2 are separated into several classes.

Class description	*Atoms from G_Q*	*Atoms from G_T*
C-atom with one single bond (class I)	1, 2	2, 5
C-atom with two single bonds (class II)		4
C-atom with three single bonds (class III)	3	3
O-atom with one single bond (class IV)	4	1



Although the optimized backtracking algorithm offers considerable improvement it still remains a heavy task to search a structural database with more than 50,000 compounds on conventional computers. The third strategy for optimization of substructure searching is done by a process termed “screening”. Screening systems normally use a predefined set of structural fragments called keys. For each key a preliminary substructure search is performed across the whole structural database. For each database compound a string of bits is contructed. Each bit of this string denotes the presence or absence of a key in the corresponding database compound. The kth bit in the bit-string is set to 1 if the kth key fragment is a substructure of the current data base compound otherwise the kth bit is set to 0. In the same way during the substructure searching a bit-string of the query structure is constructed. This step is usually very fast since a few hundred isomorphism checks are performed for a set of quite simple structural fragments. Further, the query bit-string is compared with each of the bit-strings from the database. The target compounds are screened as follows: each key which is present in the query structure must be present in the target structure. If at least one key that is present in the query graph is not present in the target graph, then this compound is pruned from a further processing. In this way a great many structures which are not likely to survive the isomorphism check are pruned early in the screening stage, thus escaping the much more time-consuming backtracking algorithm. The selection keys are of prime importance in designing a screening system. The basic principle is that one has to use middle frequency keys. The most frequently occurring keys e.g., the $–CH_2$- fragments are not useful since they do not discriminate effectively between the different target structures.

Similarity searching

Similarity searching is an alternative and complement to exact searching. Similarity searching retrieves objects that are similar to a query, sorted in order of their decreasing similarity. High ranked objects are likely to have similar properties to the query and thus be of interest for property prediction. Pattern matching and signature analysis are names given to similarity searching that originate from other application areas.

Similarity measures

In order to compare two chemical objects eg. two molecules we need a measure. Plenty of similarity measures have been proposed. Generally speaking these measures can be divided into two cases: one of qualitative characteristics and the other of quantitative characteristics.

We can give the definition of a similiarity measure as follows: Consider two objects A and B, a is the number of features (characteristics) present in A and absent in B, b is the number of features absent in A and present in B, c is the number of features common to both objects and d is the number of features absent from both objects. Thus c and d measure the present and the absent matches, respectively, i.e. similarity, while a and b measure the corresponding mismatches ie. Dissimilarity. The total number f features is n = a + b + c + d .

The total number of bits set on A is a + c, and the total number of bits set on B is b + c. These totals form the basis of an alternative notation that uses a instead of a + c and b instead of b + c. this notation, however lumps together similarity and dissimilarity "components" – a disadvantage when interpreting a similarity measure.

Similarity search process

The most common objects of interest to a chemist are molecules. Some sources of drug like compounds are the MDL Drug Data Report (MDDR) a licensed database compiled from the patent literature containing about 115 000 compounds, as well as the database of the National Cancer Institute (NCI), containing about 250 000 compounds. The biological data base is formed of three files, which contain data from different types of measurements – TGI, LC50 and GI50.

An example of a fragment based search space is

"… a large set of diverse fragments together with generic definitions of how the fragments can be combined to molecules … The space contains about 17 000 fragments which can be connected to each other via 12 different link types.

Reactions an be considered as composite systems containing reactant and product molecules as well as reaction sites. The similarity of chemical structures is defined by generalized reaction types and by gross structural features. The similarity of reactions can be defined by physicochemical parameters of the atoms and bonds at the reaction site.

Descriptor selection and encoding

The atom pair ap, and topological torsion tt descriptor may be selected. Two other atomic properties are – atomic log P contribution and partial atomic charges.

Similarity Measure Selection

In general, different similarity measures yield different rankings, except when they are monotonic. Improved results are obtained by using fusion methods to combine the rankings resulting from different coefficients. Empirically, the Dice coefficient has worked better than cosine similarity in retrieving actives and is the standard choice for use with the ap and tt descriptors.

To evaluate the performance of the descriptors one needs a database of compounds which the biological activities are known. Queries are selected that are typical of drug-like molecule and from therapeutic categories that

1. Contain a large enough number of actives (eg.> 50) for reasonable statistics.
2. Have several chemical classes present in them
3. Are fairly specific

Three Dimensional Structure Search Methods

Chemists know that 2D representation of molecular moieties gives a very rough picture of their real-world structure. While for some practical applications this representation is sufficient for most modern investigations in all areas of molecular design, 3D structure representation and 3D structure search are highly mandatory.

The first task was the creation of large 3D chemical structure database. A subsequent step was the development of fast 3D search approaches (follows the existing 2D methods such as atomly atom mapping, maximal common substructure, 2D keys, fragments, etc., by substituting them with 3D counterparts). 3D similarity search methods are quite well developed.

3D substructure search in usually known as pharmacophore searching in QSAR. Generally speaking there are two major approaches to it: topological and chemical function queries. In all the 3D search methods the conformational flexibility creates considerable difficulties.

Structure descriptor

It is a mathematical representation of a molecule resulting from a procedure transforming the structural information encoded within a symbolic representation of a molecule.

Molecules can be represented by structure descriptors in a hierarchial manner with respect to

a. Descriptor data type
b. Molecular representation of the compound

The information content of a structure descriptor on two major factors

a. Molecular representation of the compound
b. The algorithm which is used for the calculation of the descriptor.

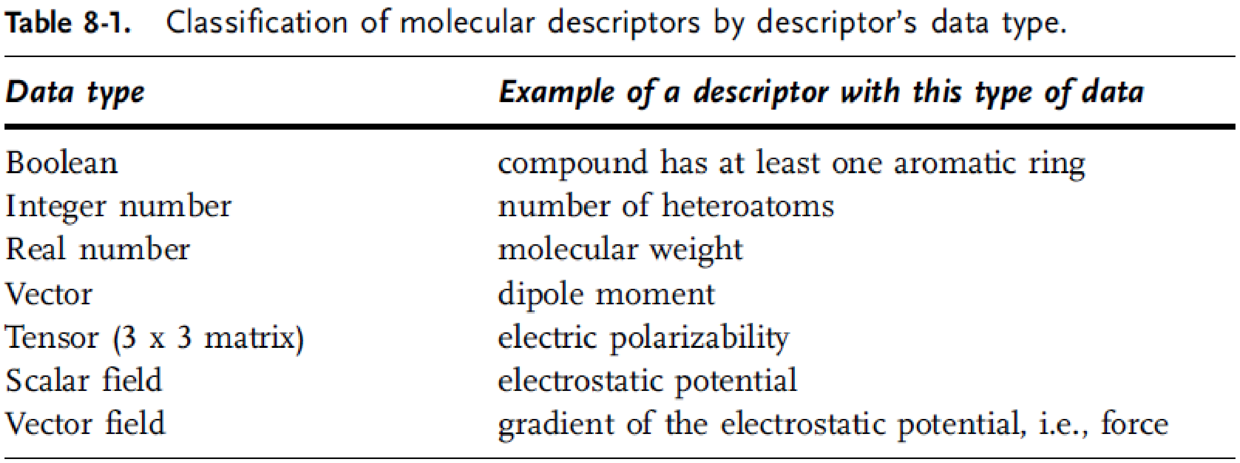


Table 8-1. Classification of molecular descriptors by descriptor's data type.

Data type	*Example of a descriptor with this type of data*
Boolean	compound has at least one aromatic ring
Integer number	number of heteroatoms
Real number	molecular weight
Vector	dipole moment
Tensor (3 x 3 matrix)	electric polarizability
Scalar field	electrostatic potential
Vector field	gradient of the electrostatic potential, i.e., force

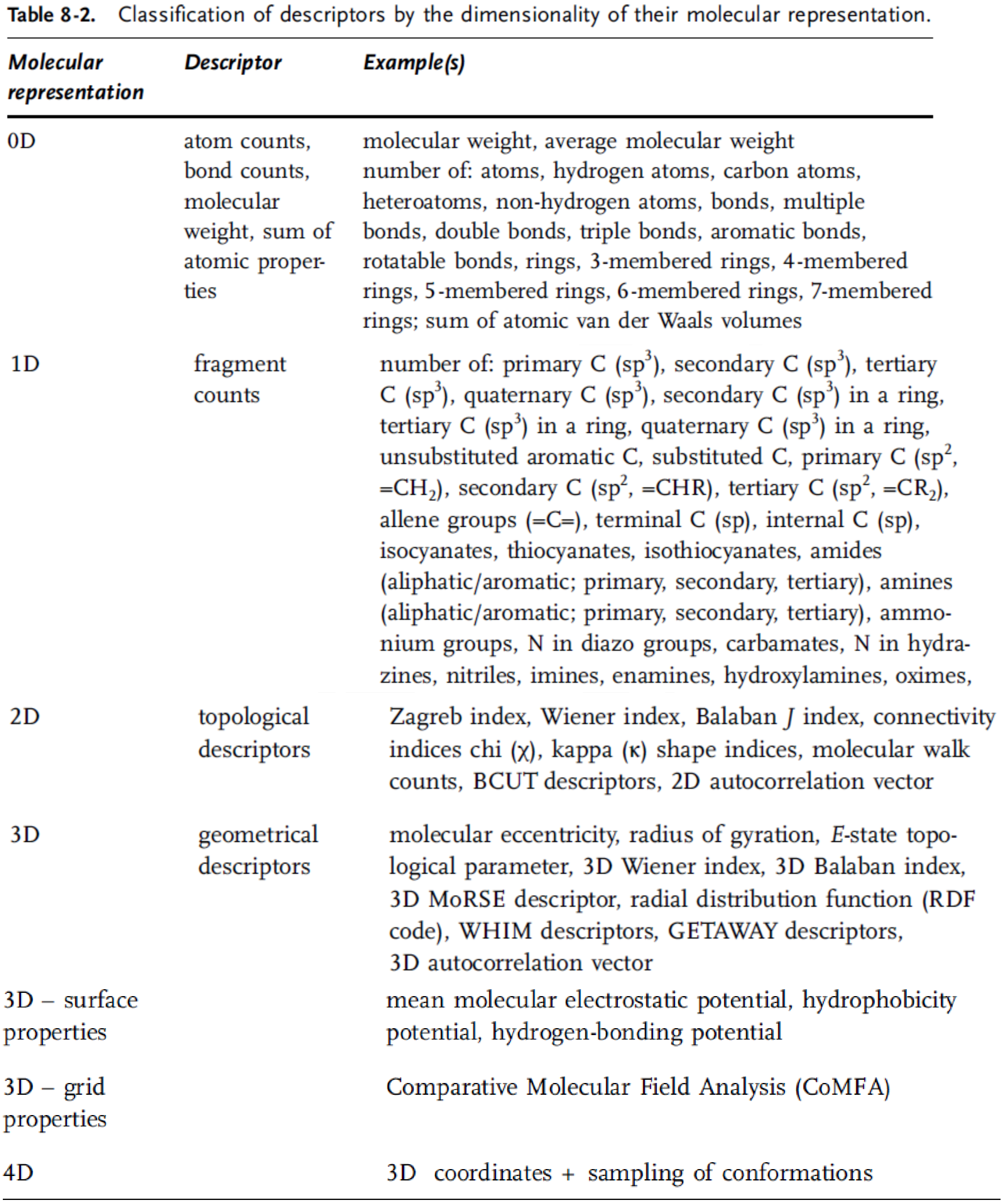


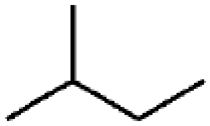
Table 8-2. Classification of descriptors by the dimensionality of their molecular representation.

Molecular representation	*Descriptor*	*Example(s)*
0D	atom counts, bond counts, molecular weight, sum of atomic properties	molecular weight, average molecular weight number of: atoms, hydrogen atoms, carbon atoms, heteroatoms, non-hydrogen atoms, bonds, multiple bonds, double bonds, triple bonds, aromatic bonds, rotatable bonds, rings, 3-membered rings, 4-membered rings, 5-membered rings, 6-membered rings, 7-membered rings; sum of atomic van der Waals volumes
1D	fragment counts	number of: primary C (sp^3), secondary C (sp^3), tertiary C (sp^3), quaternary C (sp^3), secondary C (sp^3) in a ring, tertiary C (sp^3) in a ring, quaternary C (sp^3) in a ring, unsubstituted aromatic C, substituted C, primary C (sp^2, $=CH_2$), secondary C (sp^2, =CHR), tertiary C (sp^2, $=CR_2$), allene groups (=C=), terminal C (sp), internal C (sp), isocyanates, thiocyanates, isothiocyanates, amides (aliphatic/aromatic; primary, secondary, tertiary), amines (aliphatic/aromatic; primary, secondary, tertiary), ammonium groups, N in diazo groups, carbamates, N in hydrazines, nitriles, imines, enamines, hydroxylamines, oximes,
2D	topological descriptors	Zagreb index, Wiener index, Balaban *J* index, connectivity indices chi (χ), kappa (κ) shape indices, molecular walk counts, BCUT descriptors, 2D autocorrelation vector
3D	geometrical descriptors	molecular eccentricity, radius of gyration, *E*-state topological parameter, 3D Wiener index, 3D Balaban index, 3D MoRSE descriptor, radial distribution function (RDF code), WHIM descriptors, GETAWAY descriptors, 3D autocorrelation vector
3D – surface properties		mean molecular electrostatic potential, hydrophobicity potential, hydrogen-bonding potential
3D – grid properties		Comparative Molecular Field Analysis (CoMFA)
4D		3D coordinates + sampling of conformations

Topological descriptors

A huge variety of descriptors are frequently applied in modeling physical, chemical or biological properties of organic compounds. Topological descriptors represent the constitution of these compounds and can be computed from their molecular graph. Each atom is represented by a vertex in the graph. Accordingly, the bonds are described by the edges.

Eg. 2-methyl butane
It consists of five nodes, four edges and the adjacency relationships implied in the structure.



3D descriptors

Physical, chemical and biological properties are related to the 3D structure of the molecule. In essence the experimental sources of 3D structure information are X-ray crystallography, electron diffraction or NMR spectroscopy. For compounds without experimental data on their 3D structure automatic methods for the conversion of the connectivity information into a 3D model are required. Two of the widely used programs for the generation of 3D structures are CONCORD and CORINA. These programmes generate one low-energy conformation for each molecule.

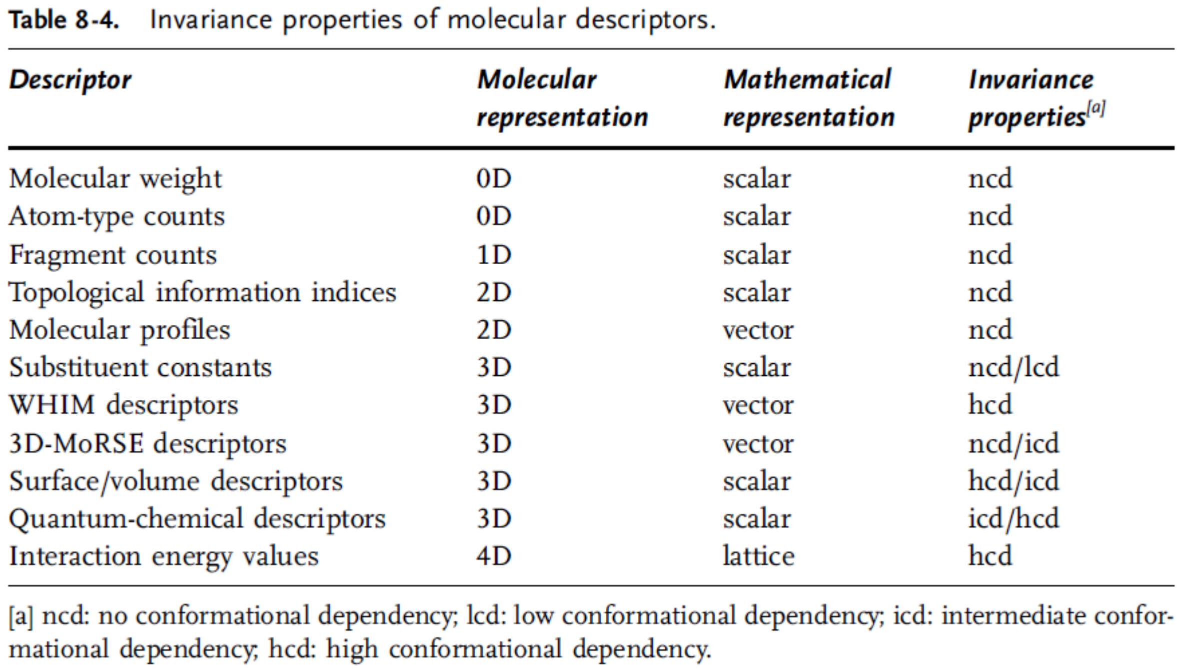


Table 8-4. Invariance properties of molecular descriptors.

Descriptor	*Molecular representation*	*Mathematical representation*	*Invariance properties*[a]
Molecular weight	0D	scalar	ncd
Atom-type counts	0D	scalar	ncd
Fragment counts	1D	scalar	ncd
Topological information indices	2D	scalar	ncd
Molecular profiles	2D	vector	ncd
Substituent constants	3D	scalar	ncd/lcd
WHIM descriptors	3D	vector	hcd
3D-MoRSE descriptors	3D	vector	ncd/icd
Surface/volume descriptors	3D	scalar	hcd/icd
Quantum-chemical descriptors	3D	scalar	icd/hcd
Interaction energy values	4D	lattice	hcd

[a] ncd: no conformational dependency; lcd: low conformational dependency; icd: intermediate conformational dependency, hcd: high conformational dependency.

Possible Questions

Part A (Online multiple choice questions)

Part B (Each question carries two marks)

1. What are the different ways available for searching of a chemical structure.
2. What is meant by a full structure search.
3. State the various approaches devised for full structure search
4. What is meant by a substructure search
5. State the various approaches devised for substructure structure search
6. State any two basic strategies for the improvement of the performance of substructure search
7. What is meant by backtracking algorithm
8. State the similar structure similar-property principle
9. What are the essentials required to evaluate the performance of a similarity search descriptor
10. Write notes on topological descriptors
11. Write notes on 3D descriptors

Part C (Each question carries eight marks)

1. Explain in detail the different methods available for Searching of a chemical structure
2. Explain in detail about the Full structure search in cheminformatics
3. Describe the substructure search in cheminformatics
4. State and explain the backtracking algorithm for chemical structure search.
5. Explain the optimization process of backtracking algorithm for chemical structure search
6. Explain in detail about similarity search
7. What are the different steps to be followed in similarity search approach.
8. Explain the three dimensional structure search methods.
9. Explain the invariance properties of molecular descriptors.

UNIT III (Multiple choice Questions)

S.No	Question	A	B	C	D	Answer
1	Full structure search using empirical formula approach	identify a whole class of compounds (chemical isomers) rather than a single structure	does not reflect accurately molecular composition	retrieves objects that are similar to a query	3D structure representation	identify a whole class of compounds (chemical isomers) rather than a single structure
2	Atom-by-atom searching is pertained to	backtracking algorithm	Substructure Search	Full structure search	Similiarity search	backtracking algorithm
3	The structures are highly degenerate during the Full structure search is made using	empirical formula approach	molecular weights	trival names	various line notations,	empirical formula approach
4	Which is used as a supplementary descriptor in Full structure search	empirical formula approach	molecular weights	trival names	various line notations,	empirical formula approach
5	The simplest and more suitable way of storing chemical information is by using the character strings	empirical formula approach	molecular weights	trival names	Trade names	Trade names
6	When this is used as a character string it does not	empirical formula	molecular weights	trival names	Trade names	Trade names

	reflect accurately themolecular composition	approach				
7						
8	molecular formulas as a character string are used as a used as supplementary descriptors because	The resulted structures are highly degenerate	does not reflect accurately molecular composition	retrieves objects that are similar to a query	3D structure representation	The resulted structures are highly degenerate
9	registry numbers can be used as a charactering string for	backtracking algorithm	Substructure Search	Full structure search	Similiarity search	Full structure search
10	constitutional diagrams (2D representations) are used as a charactering string for	backtracking algorithm	Substructure Search	Full structure search	Similiarity search	Full structure search
11	atom coordinates (2D or 3D representations) are used as a charactering string for	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Full structure search
12	topological indices are used as a charactering string for	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Full structure search
13	Hash codes are used as a charactering string for	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Full structure search
14	Can be stored and searched as character strings	empirical formula approach	molecular weights	trival names	Trade names	
15	Line notations is in widespread use among chemists.	SMILES	ROSDAL	SLN	WLN	SMILES

16	In full structure search Consisting of character strings, these representations are compact and easy to use	SMILES	ROSDAL	SLN	WLN	SMILES
17	The structure search used for identification of equivalent atoms	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
18	The structure search used for determination of maximal common substructure	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
19	The structure search used for ring detection	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
20	The structure search used for calculation of topological indices	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
21	----------is very important in medicinal chemistry	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
22	Which is important in QSAR	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
23	Which is important in spectroscopy and many other fields in the process of pharmacophore perception	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search

24	Which is important in chromophore perception	Substructure Search	Full structure search	Similiarity search	backtracking algorithm	Substructure Search
25	The process of identifying parts of a given structure that are equivalent to a specified query substructure	Full structure search	Similiarity search	backtracking algorithm	Substructure Search	Substructure Search
26	The task of checking whether the query graph (GQ) is isomorphic with a subgraph of another target graph (GT). It is pertained to	Full structure search	Similiarity search	backtracking algorithm	Substructure Search	Substructure Search
27	The task of checking whether the query graph (GQ) is isomorphic with a subgraph of another reference graph. It is pertained to	Full structure search	Similiarity search	backtracking algorithm	Substructure Search	Substructure Search
28	The basic strategy for the improvement of the performance of substructure search algorithms.	Optimisation of the hardware and software technologies used	By using WLN notation	By using MOLfiles	By using SDfiles	Optimisation of the hardware and software technologies used
29	The basic approach to a fast search of an isomorphism among all mappings is so called	Full structure search	Similiarity search	backtracking algorithm	Substructure Search	backtracking algorithm

30	It searches an isomorphism GQ---GT in such a way that all checked mappings can be organized hierarchically as a tree	Full structure search	Similiarity search	backtracking algorithm	Substructure Search	backtracking algorithm
31	backtracking algorithm is otherwise called as	Full structure search	Similiarity search	Atom-by-atom searching	Substructure Search	backtracking algorithm
32	It is the the core part of every software system that performs substructure searching.	Full structure search	Similiarity search	Atom-by-atom searching	Substructure Search	backtracking algorithm
33	The Backtracking Algorithm is	The basic approach to a fast search of an isomorphism among all mappings	is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures	comprise the use of molecular formulas, molecular weights, trade and /or trival names, various line notations	is an alternative and complement to exact searching	The basic approach to a fast search of an isomorphism among all mappings
34	Substructure search is	The basic approach to a fast search of an isomorphism among all mappings	is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures	comprise the use of molecular formulas, molecular weights, trade and /or trival names, various line notations	is an alternative and complement to exact searching	is usually the first step in the implementation of other important topological procedures for the analysis of chemical

						structures
35	Full structure search	The basic approach to a fast search of an isomorphism among all mappings	is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures	comprise the use of molecular formulas, molecular weights, trade and /or trival names, various line notations	is an alternative and complement to exact searching	comprise the use of molecular formulas, molecular weights, trade and /or trival names, various line notations
36	Similiarity search is	The basic approach to a fast search of an isomorphism among all mappings	is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures	comprise the use of molecular formulas, molecular weights, trade and /or trival names, various line notations	is an alternative and complement to exact searching	is an alternative and complement to exact searching
37	Usually consists of an application of several heuristics which reduce the number of candidates atoms for mapping from GQ ---GT.	optimization of the backtracking algorithm	Optimisation of full structure search	Optimization of substructure search	Optimization os similiarity search	optimization of the backtracking algorithm
38	The process in which the atoms in GQ and GT are separated into different classes	partitioning	Union	Intersection	adding	partitioning
39	The third strategy for optimization of	Screening	reduce the number of	partitioning	coagulation	Screening

	substructure searching is done by a process		candidates atoms for mapping from GQ ---GT.			
40	Screening systems normally use a predefined set of structural fragments called	Keys	Fragments	Functional groups	chromophores	Keys
41	The basic principle screening is that one has to use	Middle frequency keys	High frequency keys	Low frequency keys	Ultra frequency keys	Middle frequency keys
42	Pattern matching is also called	Full structure search	Similiarity search	Atom-by-atom searching	Substructure Search	Similiarity search
43	signature analysis is also called.	Full structure search	Similiarity search	Atom-by-atom searching	Substructure Search	Similiarity search
44	Which is not related	Pattern Matching	Screening	Signature analysis	atom-by-atom searching	atom-by-atom searching
45	The similarity of reactions can be defined by	physicochemical parameters of the atoms	Molecular weight	Molecular formula	Pattern Matching	physicochemical parameters of the atoms
46	The similarity of reactions can be defined by	bonds at the reaction site	Molecular weight	Molecular formula	Pattern Matching	bonds at the reaction site
47	The similarity of chemical structures	defined by generalized reaction types	atom-by-atom searching	Defined by functional groups	Defined by chromophores	defined by generalized reaction types
48	The similarity of chemical structures are defined by	by gross structural	atom-by-atom searching	Defined by functional	Defined by chromophores	by gross

		features		groups		structural features
49	In general, different similarity measures yield different rankings, except when they are	Monotonic	Isotonic	isomeric	isotopic	Monotonic
50	Improved results are obtained by using fusion methods to combine the rankings resulting from different coefficients, it is pertained to	Similiarity measure selection	atom-by-atom searching	Substructure search	The Backtracking Algorithm	Similiarity measure selection
51	To evaluate the performance of the descriptors one needs a database of compounds which the	biological activities are known	Functional groups are known	Molecular formula are known	Crystal structures are known	biological activities are known
52	For similarity measure selection Queries are selected from therapeutic categories that	Contain a large enough number of actives for reasonable statistics	Contain a small number of actives for reasonable statistics	Only one chemical class present in them	Are fairly diversified	Contain a large enough number of actives for reasonable statistics
53	For similarity measure selection Queries are selected from therapeutic categories that	Have several chemical classes present in them	Contain a small number of actives for reasonable statistics	Only one chemical class present in them	Are fairly diversified	Have several chemical classes present in them
54	For similarity measure selection Queries are selected from therapeutic categories that	Are fairly specific	Contain a small number of actives for reasonable statistics	Only one chemical class present in them	Are fairly diversified	Are fairly specific

55	The therapeutic categories that Contain a large enough number of actives for reasonable statistics can be taken for	Similiarity measure selection	Substructure search	Full structure search	Fragment search	Similiarity measure selection
56	The therapeutic categories that have several chemical classes present in them can be taken for	Similiarity measure selection	Substructure search	Full structure search	Fragment search	Similiarity measure selection
57	The therapeutic categories that are fairly specific can be taken for	Similiarity measure selection	Substructure search	Full structure search	Fragment search	Similiarity measure selection
58	One among structure search is highly mandatory for most modern investigations in all areas of molecular design	3D structure search	2D structure search	Full structure search	Sub structure search	3D structure search
59	3D substructure search in usually known as	pharmacophore searching in QSAR	Full structure search	Sub structure search	Chromophore and functional group search	pharmacophore searching in QSAR
60	Represent the constitution of these compounds and can be computed from their molecular graph	Topological descriptor	3D descriptor	Molecular profiles	Fragment counts	Topological descriptor

UNIT IV

Applications: Prediction of Properties of Compounds; Linear Free Energy Relations; Quantitative Structure-Property Relations; Descriptor Analysis; Model Building; Modelling Toxicity; Structure-Spectra correlations; Prediction of NMR, IR and Mass

Prediction of Properties of Compounds

The fundamental physical properties of a compound such as its polarity or lipophilicity determine its behavior in chemical, biochemical or environmental processes and are therefore required for understanding and modeling in the fields such as drug design, reaction prediction and biodegradation. Although the amount of experimental data is growing rapidly, the number of newly synthesized or designed compounds is increasing even more quickly, especially through high-throughput methods such as parallel synthesis and combinatorial chemistry. With techniques such as virtual screening, compounds are not synthesized at all but their activity against potential drug receptors should nevertheless be modeled. Thus, the need for reliable methods for the prediction of physicochemical properties is evident.

The basic approach to the problem of estimating properties can be written in a very simple form that states that a molecular property P can be expressed as a function of the molecular structure C

$P = f (C)$

The function f (C) may have a very simple form, as in the case for the calculation of the molecular weight from the relative atomic masses. However f (C) may be complicated in most of the cases. For instance, the partitioning between two phases is a temperature-dependant constant of a substance with respect to the solvent system. It P can be written as a function of molecular structure C, solvent S and Temperature T.

$P = f (C, S, T)$

Two approaches to quantify f (C) i.e. to establish a quantitative relationship between the structural features of a compound and its properties

a. Quantitative structure-property relationship (QSPR)
b. Linear free energy relationships (LFER)

Linear free energy relationships (LFER)

LFER methods are widely used for the properties like partition coefficients, binding constants, or reaction rate constants. This is based on the pioneering work of Hammett for the prediction of chemical reactivity. The basic assumption is that the influence of a structural feature on the free energy change of a chemical process is constant for a congeneric series of compounds. A property φ that is linearly dependant on a free energy change can then be calculated by the property of the basic element of this series, the so-called parent element, and the constant φx for the structural feature X.

$$\Delta G = -2.3RT \log \Phi$$

$$\Delta\Delta G = \Delta G_{R\text{–}X} - \Delta G_{R\text{–}H} = -2.3RT \log \Phi_{R\text{–}X} + 2.3RT \log \Phi_{R\text{–}H}$$

$$\log \Phi_{R\text{–}X} - \log \Phi_{R\text{–}H} = -\frac{1}{2.3RT}\Delta\Delta G = k\Delta\Delta G = \phi_X$$

$$\log \Phi_{R\text{–}X} = \log \Phi_{R\text{–}H} + \phi_X$$

This basic LFER approach has later been extended to the more general concept of fragmentation. Molecules are dissected into substructures and each substructure is seen to contribute a constant increment to the free energy based property. The promise of strict linearity does not hold true in most cases, so corrections have to be applied in the majority of methods based on a fragmentation approach. Correction terms are often related to long range interactions such as resonance or steric effects.

Quantitative Structure-Property Relationships (QSPR)

The general procedure in a QSPR approach consists of three steps

a. Structure representation
b. Descriptor analysis
c. Model building

Structure representation

Descriptors have to be found representing the structural features which are related to the target property. This is the most important step in QSPR, and the development of powerful descriptors is of central interest in this field. Descriptors can range from simple atom or functional group counts to quantum chemical descriptors. They can be derived on the basis of the connectivity (2D descriptors), the 3D structure, or the molecular surface (3D descriptors) of a structure. Which kind of descriptors should or can be used is primarily dependent on the size of the data set to be studied and the required accuracy.

Descriptor Analysis

In general, the set of calculated descriptors should not be used directly for the model building process, mainly because of three problems

1. Different elements of the descriptor set may intercorrelate ie. Different descriptors basically encode the same structural aspect.
2. Descriptors may encode features that do not contribute to the property at all
3. The overall size of a descriptor set may become unmanageable large.

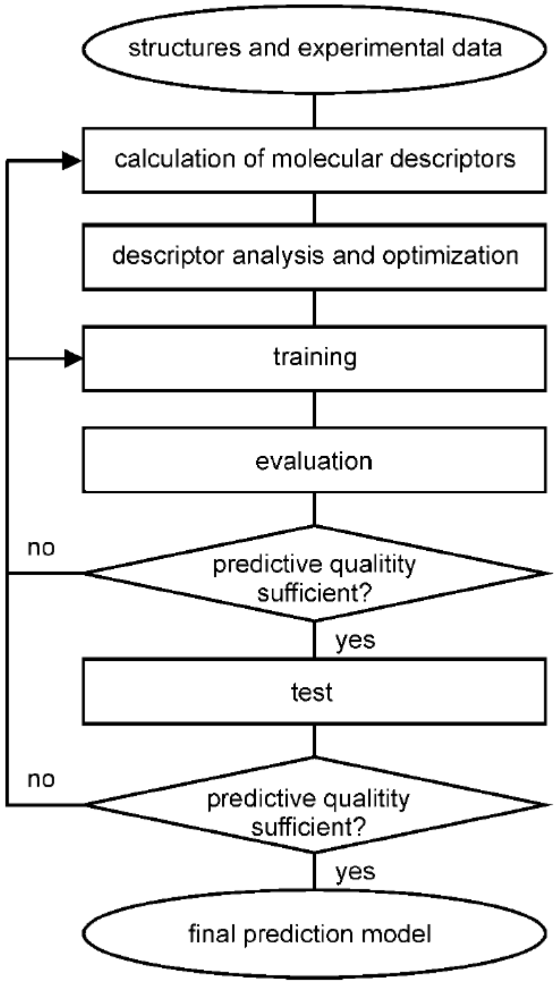
Each case requires a pre-processing of the descriptor set such that the essential information is extracted into a reduced descriptor set with higher information density related to the target property. Mainly, two statistical measures are used to judge the quality of the descriptors: the variance and the correlation coefficient among the descriptors. The former is a measure of the variation of a descriptor across a data set. A low variance indicates little information content of a descriptor. The latter is a measure of internal redundancy. Completely independent descriptors have a correlation coefficient of 0.0 and are said to be orthogonal. This ideal case is of course hardly ever found, and the correlation of two descriptors should normally be not greater than 0.6, but reports of acceptable correlation coefficients between descriptors have ranged from less than 0.4 to 0.9 in the literature. The descriptor set can then be reduced by eliminating candidates that show such bad characteristics.

Model Building

The model building step deals with the development of mathematical models to relate the optimized set of descriptors with the target property. Two statistical measures indicate the quality of a model, the regression coefficient, r, or its square r^2, and the standard, σ.

Model building consists of three steps; training, evaluation and testing. In the ideal case the whole training data set is divided into three portions, the training, the evaluation set, and the test set. A wide variety of statistical or neural network methods can be used to derive QSPR and QSAR models. The most frequently used methods are Multiple Linear Regression Analysis (MLRA) and feed forward neural networks with back propagation of errors. Once a model has been derived with the training set, the evaluation set can be used to test the predictive power of the resulting models, ie, to predict the target property for compounds yet unknown to the model and to optimize model parameters thereby. In many cases the data set is too small to allow its splitting. Therefore cross-validation techniques are applied for the evaluation step. In k-fold cross validation the training set is split in k subsets. Then k-1 subsets are used as a training set and one subset as test set. This procedure is repeated k times. As we have now a prediction for each compound, we can calculate cross-validated errors of the predictions. Values of k usually range from 5 to 10. If k equals N, the number of cases in the training set, the procedure is called leave-one-out cross-validation.

Finally a model has to be tested using an independent data set with compounds yet completely unknown to the model: the test set. The complete process of building a prediction model is depicted below.


structures and experimental data
calculation of molecular descriptors
descriptor analysis and optimization
training
evaluation
no
predictive qualitity sufficient?
yes
test
no
predictive qualitity sufficient?
yes
final prediction model

Prediction of the toxicity of compounds

Toxicity may be one of the most difficult properties to model, especially for high throughput screening in the drug discovery process. The difficulties arise from the following facts: The effects of toxicants are species specific, organ specific and time-dependent. This has the consequence that the concentration at which adverse effects occur can vary over several orders of magnitude depending on the species and the type of test.

Toxic effects are measured through a wide variety of tests. Roughly one can distinguish two types

1. In vivo: with organisms such as rodents, fish, water fleas, earthworms, algae

2. In-vitro tests with single cells, organelles like mitochondria or even just enzymes that are affected by toxicant. The classical in vivo test for acute toxicity of a chemical is the LC50 value. This is the concentration at which 50% of the test species are killed by the toxic effects of a compound in a given time period. Until now this has also been one of the most common values to be predicted with QSAR equations.

After a chemical is released and distributed in the environment it might enter our model fish through its gills or to some extent also through food. Then it is distributed in the body (mainly between the aqueous phase, storage lipids and membrane lipids). At the same time metabolism and excretion processes take place. Metabolism leads in most cases to less toxic compounds, but in some cases the contrary can happen: the product of metabolism is more toxic than the mother compound. Thus this possibility needs to be kept in mind too, if compounds are tested in vivo. The fraction of the compound reaching the target then causes the adverse effect. Therefore, if toxicity data are modeled one should always have a clear picture of whether one is modeling toxicokinetic or toxicodynamic effects, or both effects together.

The most widely used descriptor for the hydrophobicity term in toxicology is the distribution coefficient between octanol and water, log P_{ow}. The bulk solvent octanol is of course a rather crude model to approximate uptake and transport in a cell membrane and has shortcomings especially for charged compounds. Some scientists therefore work increasingly with membrane/water distribute coefficients in order to obtain a more realistic picture of how compounds distribute in cell membranes. However for screening large databases it is still necessary to use the easily calculated log P value.

Structure-Spectra correlations; Prediction of NMR, IR and Mass spectra

NMR spectra have been predicted using quantum chemistry calculations, data-base searches, additive methods, regressions and neural networks.

Several methods have been developed for establishing correlations between IR vibrational bands and substructure fragments. Counter propagation neural networks were used to make predictions of the full spectra from RDF codes of the molecules.

Correlations between structure and mass spectra were established on the basis of multivariate analysis of the spectra, database searching, or the development of knowledge-based systems, some including explicit management of chemical reactions.

The investigation of molecular structures and of their properties is one of the most fascinating topics in chemistry. Chemistry has a language of its own for molecular structures which has been developed from the first alchemy experiments to modern times. With the improvement of computational methods for chemical information processing, several descriptors for the handling of molecular information have been developed and used in a wide range of applications.

^{13}C-NMR spectra

NMR spectroscopy is probably the singly most powerful technique for the confirmation of structural identity and for structure elucidation of unknown compounds. Additionally, the relatively low measurement times and the facility for automation contribute to its usefulness and industrial interest.

Thus in the area of combinatorial chemistry many compounds are produced in short time ranges, and their structures have to be confirmed by analytical methods. A high degree of automation is required, which has fueled the development of software that can predict NMR spectra starting from the chemical structure, and that calculates measures of similarity between simulated and experimental data. These tools are obviously also of great importance to chemists working with just a few compounds at a time, using NMR spectroscopy for structure confirmation.

Furthermore, the prediction of ^{1}H and ^{13}C NMR spectra is of great importance in systems for automatic structure elucidation. In many such systems, all isomers with a given molecular formulae are automatically produced by a structure generation, and are then ranked according to the similarity of the spectrum predicted for each isomer to the experimental spectrum.

^{1}H NMR spectra are basically characterized by the chemical shift and coupling constants of signals. The chemical shift for a particular atom is influenced by the 3D arrangement and bond types of the chemical environment of the atom and by its hybridization. The multiplicity of a signal depends on the coupling partners and on the bond types between atom.

However the situation is much less complex for ^{13}C NMR spectra. Whereas measurement conditions may have a high impact on ^{1}H NMR spectra, chemical shift values in ^{13}C NMR spectra are less sensitive to changes in solvent and experimental conditions. In fact, there is usually a good correlation between the 2D arrangement of atoms and their ^{13}C NMR spectra chemical shifts. In addition ^{13}C NMR spectra are usually measured with the protons decoupled and then show few coupling effects except fluorine and phosphorous atoms in the vicinity of the carbon atom. These are the reasons why ^{13}C NMR spectra can be represented quite well as pairs of chemical shift values and intensities.

Several empirical approaches for ^{13}C NMR spectra prediction are based on the availability of large ^{13}C NMR spectra databases. By using special methods for encoding substructures that correspond to particular parts of the ^{13}C NMR spectrum, the correlation of substructures and partial spectra can be modeled. Substructures can be encoded by using the additive model greatly developed. The authors represented skeleton structures and substituent's by individual codes and calculation rules. A more general additive model was introduced later.

However one of the most successful approaches to systematically encoding substructures for ^{13}C NMR spectrum prediction was introduced called HOSE code(Hierarchial Organisation of Spherical Environments) to describe structures. The chemical shift value of a carbon atom is basically influenced by the chemical environment of the atom. The HOSE code describes the environment of an atom in several virtual spheres. It uses spherical layers around the atom to define the chemical environment. The first layer is defined by all the atoms that are one bond away from the central atom, the second layer includes the atoms within the two-bond distance and so on. This idea is described as an atom center fragment (ACF) concept.

The spectral signals are assigned to the HOSE codes that represent the corresponding carbon atom. This approach has been used to create algorithms that allow the automatic creation of "substructure-sub-spectrum" databases that are now used in systems for predicting chemical structures directly from ^{13}C NMR.

The basic HOSE code ignores stereochemical information such as cis-trans isomer interaction that can contribute significantly to the chemical shift values. Few researchers adopted the HOSE code method by extending it with descriptors for three, four and five bond interactions and with information about axial/equatorial substitution patterns. He used the adopted method software for identification of organic compounds and automated assignment of ^{13}C NMR spectra. In the prediction process, the software searches for matches between the HOSE codes of the model and the database of chemical shifts. The software compares all the sub-spectra of library fragments with the experimental ^{13}C NMR spectrum of the analysed compound. Besides the chemical shifts and multiplicity of the signals, the signal area is needed for automatic assignment. Substructures that have corresponding library sub-spectra coinciding with the query sub-spectra within specified deviation limits are selected and ranked according to their size. The chemical structure is generated by superimposing the atoms common to different fragments.

A number of other software packages are available to predict ^{13}C NMR spectra. The use of large ^{13}C NMR spectral databases is the most common approach; it utilizes assigned chemical structures. In an advanced approach, parameters such as solvent information can be used to refine the accuracy of the prediction. A typical application works with tables of experimental chemical shifts from experimental ^{13}C NMR spectrum. Each chemical shift value is assigned to a specific structural fragment. The query structure is dissected into fragments that are compared with the fragments in the data base. For each coincidence, the experimental chemical shift from the database is used to compose the final set of chemical shifts for the query structure. If a fragment from the query structure is not found in the internal database, the most similar fragment is used.

^{1}H-NMR spectra

Prediction of chemical shifts

Ab-initio and semi empirical calculations

When a molecule is submitted to a static magnetic field, the nuclear spin energy level splits. An oscillating magnetic field can then induce transitions between these energy levels and produce the NMR spectrum. For a nucleus in a molecule, the magnetic field is due to the applied magnetic field, but also to the magnetic field produced by the electrons and other nuclei. The NMR chemical shift of a nucleus results from the difference in energy between the nuclear spin states, which can be calculated by ab-initio quantum mechanical methods, typically by solving the Schrodinger equation with approximations. The effects of the external magnetic field on the nucleus of interest are added into the equations as a “perturbation”. It is then possible to calculate the chemical shift, which is related to the total molecular energy, the applied magnetic field and the nuclear magnetic moment.

Ab-initio calculations are particularly useful for the prediction of chemical shifts of “unusual species”. In this context “unusual species” means chemical entities that are not frequently found in the available large database of chemical shifts, e.g. charged intermediates of reactions, radicals and structures containing elements other than H,C,O, N, S, P, halogens and a few common metals.

The Gaussian Program is one of the most popular tools for ab-initio calculation of NMR chemical shifts.

Database approaches

A useful empirical method for the prediction of chemical shifts and coupling constants relies on the informationcontained in databases of structures with the corresponding NMR data. Large databases with hundred thousands of chemical shifts are commercially available and are linked to predictive systems, which basically rely on database searching. Protons are internally represented by their structural environments, usually their HOSE codes. When a query structure is submitted, a search is performed to find the protons belonging to similar substructures. These are the protons with the same HOSE codes as the protons in the query molecule. The prediction of the chemical shift is calculated as the average chemical shift of the retrieved protons.

When common substructures cannot be found for a given proton interpolations are applied to obtain a prediction; proprietary methods are often used in commercial programs.

Increment based Methods

In this second empirical approach, which has also been used for ^{13}C NMR spectra, predictions are based on tabulated chemical shifts for classes of structures, and corrected with additive contributions from neighbouring functional groups or substructures. Several tables have been compiled for different types of protons. Increment rules can be found in nearly any textbook on NMR spectroscopy. In such tables, typical chemical shifts are assigned to standard structure fragments (eg. protons in a benzene ring). Substituent's in these blocks (eg. substituent's in ortho, meta and para positions) are assumed to make independent additive contributions to the chemical shift. Once the tables are defined, the method is easy implement, does not require databases and is extremely fast.

Tools prediction of ^{1}H NMR chemical shifts

Tools prediction of ^{1}H NMR chemical shifts of organic compounds are of great interest for automatic structure elucidation, for the analysis of combinatorial libraries and for assisting experimental chemists in the structural characterisation of small data sets of compounds.

A combination of physicochemical, topological and geometric information is used to encode the environment of a proton. The geometric information is based on Proton radial distribution function (RDF) descriptors and characterises the 3D environment of the proton.

Four different types of protons were treated separately according to their chemical environment: protons belonging to aromatic systems, to non aromaticπ – systems, to rigid aliphatic substructures and non-rigid aliphatic substructures. Each proton was represented by a fixed number of descriptors.

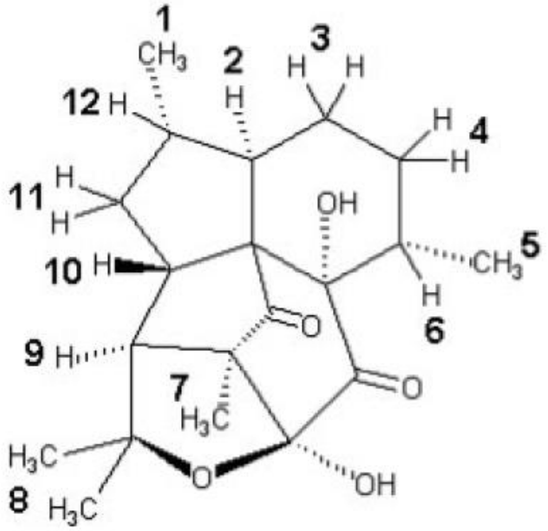
Representations of protons

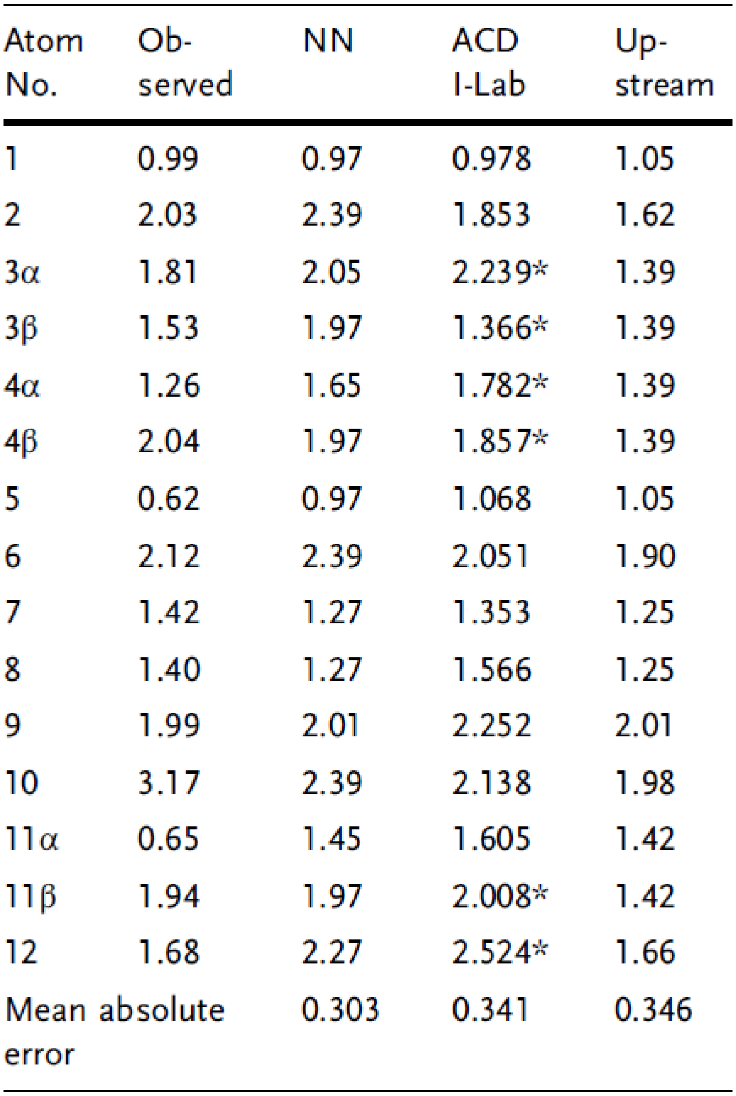
An extended set of physicochemical descriptors was used in this study, including for example partial atomic charge and effective polarazability of the protons, average of electronegativities of atoms two bonds away, or maximum π –atomic charge of atoms two bonds away.

Geometric descriptors were based on local RDF descriptors.

The minimum and maximum bond angles centered on the atom adjacent to the proton were also used as geometric descriptors. For aromatic and non-rigid aliphatic protons, these were the only two geometric descriptors used. In addition to these for non-aromatic π-protons the proton RDF descriptor and the nuber of non-hydrogen atoms at cis and trans positions were used as geometric descriptors. For rigid aliphatic protons, all the geometric descriptors were calculated.

Topoligical descriptors were based on the analysis of the connection table and atom properties. Some examples of topological descriptors are the number of carbon atoms in the second sphere centered on the proton, the number of oxygen atoms in the third sphere, and the number of atoms in the second sphere that belongs to an aromatic system. After selection of descriptors the best networks were applied to the prediction. Neural network was used for prediction.^{1}H NMR chemical shifts for a natural product was calculated and presented.





Atom No.	Observed	NN	ACD I-Lab	Upstream
1	0.99	0.97	0.978	1.05
2	2.03	2.39	1.853	1.62
3α	1.81	2.05	2.239*	1.39
3β	1.53	1.97	1.366*	1.39
4α	1.26	1.65	1.782*	1.39
4β	2.04	1.97	1.857*	1.39
5	0.62	0.97	1.068	1.05
6	2.12	2.39	2.051	1.90
7	1.42	1.27	1.353	1.25
8	1.40	1.27	1.566	1.25
9	1.99	2.01	2.252	2.01
10	3.17	2.39	2.138	1.98
11α	0.65	1.45	1.605	1.42
11β	1.94	1.97	2.008*	1.42
12	1.68	2.27	2.524*	1.66
Mean absolute error		0.303	0.341	0.346

Table: Predictions of ^{1}H NMR chemical shifts calculated based on Neural Networks, a database-centred method (ACD) and an increment-based method.

The performance of the Neural Network method is remarkable considering the relatively small data set on which it was based.

Infrared Spectra

Since IR spectroscopy monitors the vibration of atoms in a molecule in 3D space information on the 3D arrangement of the atoms should somehow be contained in an IR spectrum. However, the relationship between the 3D structure and the IR spectrum are rather complex, so no general attempt was made successfully in deriving the 3D structure of a molecule directly from the IR spectrum.

A series of monographs and correlation tables exist for the interpretation of vibrational spectra. However, the relationship of frequency characteristics and structural features is rather complicated and the number of known correlations between IR spectra and structures is very large. Many expert systems designed to assist the chemist in structural problem solving were based on the approach of characteristic frequencies.

Tools: Tele Spec-Online service

The identification of chemical compounds by IR spectroscopy is usually done by comparing an experimental spectrum of the compound with a reference spectrum. The TeleSpec system was developed to provide a method to relieve this difficulty by simulating an IR spectrum for a given input structure.

Approach

The correlation between a structure and its spectrum is rather complex. So a counter propagation network was chosen. The structure in the database are encoded using the radial distribution function (RDF) as a descriptor.

The coding is performed in three steps

1. 3D coordinates of the atoms are calculated (CORINA)
2. Physicochemical properties like charge distribution and polarizability were calculated (PETRA)
3. 3D information and Physicochemical properties are then used to code the molecule.

Mass spectra

Mass spectra are based on decomposition and reactions of organic molecules on their way from the ion source to the detector. The structure –MS correlation is basically a matter of relating reactions to the signals in a mass spectrum. The chemical structure information contained in mass spectra is difficult to extract because of the complicated relationship between MS data and chemical structures.

Identification of mass spectra is typically performed by searching for similarities of the measured spectrum to spectra stored in a library. Several MS databases and corresponding software products are used routinely. The use of correlation tables containing characteristic spectral data together with corresponding substructures has been successfully applied to spectroscopic methods.

Other methods consist of algorithms based on multivariate classification techniques or neural networks.

Possible Questions

Part A (Online multiple choice questions)

Part B (Each question carries two marks)

1. What should be the basic approach to the methods for the prediction of physicochemical properties.
2. What is meant by Quantitative structure-property relationship (QSPR)
3. What is meant by the Linear Free Energy Relatonship for the prediction of properties of compounds.
4. What are the three steps involved in the general procedure in the QPSR.
5. Why the set of calculated descriptors should not be used directly for the model building process.
6. Distinguish the two type of toxic effect measurements.
7. What is meant by an IR spectra
8. What information can be obtained from a mass spectra.
9. What is meant by a chemical shift in NMR spectra.
10. What is meant by TeleSpec online service.
11. How the coding is performed in when radial distribution function (RDF) is used as a descriptor in the prediction of NMR values..

Part C (Each question carries eight marks)

1. Discuss in detail about the Linear free energy relationships (LFER) for the prediction of properties
2. Discuss in detail about the Quantitative structure-property relationship (QSPR) for the prediction of properties.
3. Explain the structure representation, descriptor analysis and model building approaches in QSPR.
4. Explain in detail the Model building approach step in the QSPR.
5. Explain in detail the descriptor analysis approach step in the QSPR
6. Discuss in detail about the prediction of toxicity of compounds in cheminformatics.
7. What are the characteristic features of an ^{1}H-NMR spectra
8. What are the characteristic features of an ^{13}C-NMR spectra
9. Explain the methods involved in cheminformatics to predict the chemical shifts in ^{1}H-NMR spectra

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore – 641 021.

UNIT IV (Multiple choice Questions)

S.No	Question	A	B	C	D	Answer
1	The fundamental physical properties of a compound which determine its behavior in chemical, biochemical or environmental processes	Melting point	Boiling point	Specific gravity	polarity	polarity
2	The fundamental physical properties of a compound which determine its behavior in chemical, biochemical or environmental processes	Melting point	Boiling point	Specific gravity	Lipophilicity	Lipophilicity
3	To establish a quantitative relationship between the structural features of a compound and its properties	QSAR	QSPR	SAR	Molecular docking	QSPR
4	To establish a quantitative relationship between the structural features of a compound and its properties	QSAR	Linear free energy relationship	SAR	Molecular docking	Linear free energy relationship
5	The properties like the properties like partition	Quantitative structure-	Quantitative structure-	Structure activity relationship	Linear free energy relationship	Linear free energy relationship

	coefficients are determined by	property relationship	Activity relationship			
6	The properties like the properties like binding constants are determined by	Quantitative structure-property relationship	Quantitative structure-Activity relationship	Structure activity relationship	Linear free energy relationship	Linear free energy relationship
7	The properties like the properties like reaction rate constants are determined by	Quantitative structure-property relationship	Quantitative structure-Activity relationship	Structure activity relationship	Linear free energy relationship	Linear free energy relationship
8	Linear free energy relationship is used to determine	Reaction rate constants	Melting point	Boiling point	Specific gravity	Reaction rate constants
9	Linear free energy relationship is used to determine	binding constants	Melting point	Boiling point	Specific gravity	binding constants
10	Linear free energy relationship is used to determine	Partition coefficient	Melting point	Boiling point	Specific gravity	Partition coefficient
11	The basic assumption is that the influence of a structural feature on the free energy change of a chemical process is constant for a congeneric series of compounds	Quantitative structure-property relationship	Quantitative structure-Activity relationship	Structure activity relationship	Linear free energy relationship	Linear free energy relationship
12	Molecules are dissected into substructures and each substructure is seen to contribute a constant increment to the free energy based property	Quantitative structure-property relationship	Quantitative structure-Activity relationship	Structure activity relationship	Linear free energy relationship	Linear free energy relationship

13	In QPSR Descriptors have to be found representing the structural features which are related to the target property. It is related to	Structure representation	Descriptor analysis	Model building	Linear free energy	Structure representation
14	Descriptors used in QPSR for structure representation may range from	Simple atom	Molecule	Cluster of molecules	Nano material	Simple atom
15	Descriptors used in QPSR for structure representation may range from	functional group count	Molecule	Cluster of molecules	Nano material	functional group count
16	Descriptors used in QPSR for structure representation may range from	quantum chemical descriptors	Molecule	Cluster of molecules	Nano material	quantum chemical descriptors
17	The descriptors used for structure representation can be derived from	on the basis of the connectivity	On the basisPartition coefficient	On the basis of Reaction rate constant	On the basis of Binding constant	on the basis of the connectivity
18	The descriptors used for structure representation can be derived from	on the basis of the 3D structure	On the basisPartition coefficient	On the basis of Reaction rate constant	On the basis of Binding constant	on the basis of the 3D structure
19	The descriptors used for structure representation can be derived from	on the basis of the molecular surface (3D descriptors) of a structure	On the basisPartition coefficient	On the basis of Reaction rate constant	On the basis of Binding constant	on the basis of the molecular surface (3D descriptors) of a structure
20	In Descriptor analysis the set of calculated descriptors should not be used directly for the model building process, because	Different elements of the descriptor set may intercorrelate	Descriptors may not encode features that do not contribute to the property at all	The overall size of a descriptor set may not become unmanageable large	Of trans mapping	Different elements of the descriptor set may intercorrelate

21	In Descriptor analysis the set of calculated descriptors should not be used directly for the model building process, because	Different elements of the descriptor set may not intercorrelate	Descriptors may encode features that do not contribute to the property at all	The overall size of a descriptor set may not become unmanageable large	Of trans mapping	Descriptors may encode features that do not contribute to the property at all
22	In descriptor analysis the set of calculated descriptors should not be used directly for the model building process, because	Different elements of the descriptor set may not intercorrelate	Descriptors may encode features that do not contribute to the property at all	The overall size of a descriptor set may become unmanageable large	Of trans mapping	The overall size of a descriptor set may become unmanageable large
23	Completely independent descriptors are said to be orthogonal.	have a correlation coefficient of 0.0	have a correlation coefficient of 1.0	have a correlation coefficient of 10.0	have a correlation coefficient of 100.0	have a correlation coefficient of 0.0
24	In QPSR, Model building consists of	training, evaluation and testing steps	Only evaluation and testing steps	Only training and evaluation steps	Only testing step	training, evaluation and testing steps
25	The statistical measures which indicate the quality of a model	Mean	Median	regression coefficient	Mode	regression coefficient
26	The statistical measures which indicate the quality of a model	Mean	Median	Standard deviation	Mode	Standard deviation
27	The most frequently used statistical or neural network methods can be used to derive QSPR and QSAR models	Multiple Linear Regression Analysis	regression coefficient	Standard deviation	Median	Multiple Linear Regression Analysis
28	Counter propagation	^{1}H-NMR spectra	^{13}C-NMR	IR spectra	Mass spectra	IR spectra

	neural networks were used to make predictions of the full spectra from RDF codes of the molecules		spectra			
29	The spectra have been predicted using quantum chemistry calculations, data-base searches, additive methods, regressions and neural networks.	^{1}H-NMR spectra	^{13}C-NMR spectra	IR spectra	Mass spectra	^{1}H-NMR spectra
30	Correlations between structure and spectra were established on the basis of multivariate analysis of the spectra, database searching	^{1}H-NMR spectra	^{13}C-NMR spectra	IR spectra	Mass spectra	Mass spectra
31	^{13}C-NMR spectra is a powerful tool to determine	Structure of a molecule	Molecular weight of a molecule	Extinction coefficient	Particle size of a molecule	Structure of a molecule
32	^{13}C-NMR spectra is a powerful tool to determine	The carbon environment	Proton environment	The functional group	The nature of chromophore	The carbon environment
33	^{1}H-NMR spectra is a powerful tool to determine	The carbon environment in a molecule	Proton environment in a molecule	The functional group	The nature of chromophore	Proton environment in a molecule
34	IR spectra is a powerful tool to determine	The carbon environment in a molecule	Proton environment in a molecule	The functional group	The nature of chromophore	The functional group
35	Mass spectra is a powerful tool to determine	The carbon environment in a molecule	Proton environment in a molecule	The functional group	The molecular weight of a compound	The molecular weight of a compound
36	Characterized by the	Mass spectra	IR spectra	^{1}H-NMR spectra	^{13}C-NMR	^{1}H-NMR spectra

	chemical shift and coupling constants of signals.				spectra	
37	Characterised by the signals which provides the fragmentation pattern of a molecule	Mass spectra	IR spectra	^{1}H-NMR spectra	^{13}C-NMR spectra	Mass spectra
38	It is influenced by the 3D arrangement and bond types of the chemical environment of the atom and by its hybridization.	Chemical shift of a proton	Fragmentation pattern of a molecule	Molar extinction coefficient of a molecule	The functional groups present in a molecule	Chemical shift of a proton
39	The chemical shift in ^{1}H-NMR spectra is influenced by	3D arrangement of the molecule	2D arrangement of a molecule	Molecular weight of a molecule	Melting point of a molecule	3D arrangement of the molecule
40	The chemical shift in ^{1}H-NMR spectra is influenced by	bond types of the chemical environment of the atom and by its hybridization.	2D arrangement of a molecule	Molecular weight of a molecule	Melting point of a molecule	bond types of the chemical environment of the atom and by its hybridization.
41	The multiplicity of a signal depends on the coupling partners and on the bond types between atom.	the coupling partners and on the bond types between atom.	Functional groups	Stereochemistry of the molecule	Crystal structure of the molecule	the coupling partners and on the bond types between atom.
42	chemical shift values in are less sensitive to changes in solvent and experimental conditions	^{13}C NMR spectra	Mass spectra	IR spectra	^{1}H-NMR spectra	^{13}C NMR spectra
43	^{13}C NMR spectra are usually measured with the protons decoupled	Protons decoupled	Protons coupled	Carbon decoupled	Carbon coupled	Protons decoupled

44	the most successful approaches to systematically encoding substructures for ^{13}C NMR spectrum prediction	SMILES	MOLfile	SDfile	HOSE (Hierarchial Organisation of Spherical Environments)	
45	The HOSE code describes the environment of an atom in	several virtual spheres.	Coupled state	Crystal structure	A solution	several virtual spheres.
46	In an atom center fragment (ACF) concept the first layer is defined by	all the atoms that are one bond away from the central atom	includes the atoms within the two-bond distance	the atoms within the three-bond distance	the atoms within the four-bond distance	all the atoms that are one bond away from the central atom
47	In an atom center fragment (ACF) concept the second layer is defined by	all the atoms that are one bond away from the central atom	includes the atoms within the two-bond distance	the atoms within the three-bond distance	the atoms within the four-bond distance	includes the atoms within the two-bond distance
48	The ^{13}C NMR spectral signals are assigned to the	HOSE codes	Hash code	ROSDAL code	WLN code	HOSE codes
49	HOSE code ignores	Stereochemical information	Position isomers	Chain isomers	Functional group isomers	Stereochemical information
50	HOSE code ignores	Cis-trans isomer interaction	Position isomers	Chain isomers	Functional group isomers	Cis-trans isomer interaction
51	When a molecule is submitted to a static magnetic field	the nuclear spin energy level splits	the nuclear spin energy level degenerate	the nuclear spin energy level disappears	the nuclear spin energy level merges	the nuclear spin energy level splits
52	Ab-initio calculations are particularly useful for the prediction of chemical shifts of	Unusual species	Usual species	All heterocyclic compounds	Aliphatic compounds	Unusual species

53	Ab-initio calculations are particularly useful for the prediction of chemical shifts of	Charged intermediates of reactions	Usual species	All heterocyclic compounds	Aliphatic compounds	Charged intermediates of reactions
54	Ab-initio calculations are particularly useful for the prediction of chemical shifts of	Structures containing elements other than H,C,O, N, S, P, halogens	Usual species	All heterocyclic compounds	Aliphatic compounds	Structures containing elements other than H,C,O, N, S, P, halogens
55	The information is used to encode the environment of a proton	physicochemical, topological and geometric information	Nature of functional groups	Stereochemistry of the molecules	Crystal structure of the molecules	physicochemical, topological and geometric information
56	The geometric information of a proton is based on	Proton radial distribution function (RDF) descriptors	2D environment of a molecule	the analysis of the connection table and atom properties	Cis-trans isomers	Proton radial distribution function (RDF) descriptors
57	Topological information of a proton is based on	Proton radial distribution function (RDF) descriptors	2D environment of a molecule	the analysis of the connection table and atom properties	Cis-trans isomers	the analysis of the connection table and atom properties
58	The TeleSec on-line system was used to identify	IR spectrum for a given input structure.	^{1}H-NMR spectrum for a given input structure.	Mass spectrum for a given input structure.	^{13}C-NMR spectrum for a given input structure.	IR spectrum for a given input structure.
59	IR spectrum for a given input structure can be	The TeleSec on-line system	Ab-initio and semi empirical	HOSE code	CIF files	The TeleSec on-line system

	obtained from .		calculations			
60	NMR spectrum for a given input structure can be obtained from .	The TeleSec on-line system	Ab-initio and semi empirical calculations	HOSE code	CIF files	Ab-initio and semi empirical calculations

UNIT V

ComputerAssisted Structure elucidations; Computer Assisted Synthesis Design, Introduction to drugdesign; Target Identification and Validation; Lead Finding and Optimization; Analysis of HTS data; Virtual Screening; Design of Combinatorial Libraries; Ligand-Based and StructureBased Drug design; Application of Chemoinformatics in Drug Design.

The elucidation of a structure by the use of rule-based systems needs a technique for assembling a complete structure from substructure fragments that have been predicted. Several techniques and computer programes have been proposed.

CONGEN and GENOA: They can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures with more restrictive constraints eg. Isomers with specified molecular fragments.

DARC-EPIOS: could retrieve structural formulas from overlapping ^{13}C NMR data.

The above programs generate chemical structures by assembling atoms and /or molecular fragments. Another strategy is based on structure generation by removing bonds from a hyperstructure that initially contains all the possible bonds between all the required atoms and molecular fragments (structure reduction)

Eg.COCOA and GEN.

GENIUS: Implements genetic algorithms to approach the problem of finding structures consistent with an experimental ^{13}C NMR data and a molecular formula. A population of structures with the given formula is randomly generated. Then a neural network predicts a spectrum for each structure. Each structure is scored according to the similarity between the predicted spectrum and the experimental data. The fittest structures in a population survive, mutate and mate, while the less fit structures are discarded. The procedure is repeated in an iterative way to optimise the constitution of the molecules, until it produces the experimental spectra with a deviation as low as possible. The method avoids the exhaustive generation of all possible isomers and at the same time it was able to find the correct structures.

Attention has been paid to artificial neural network as a tool for spectral interpretation. The ANN approach applied to vibrational spectra allows the determination of adequate functional groups that can exist in the sample as well as the complete interpretation of spectra. Neural networks have been applied to IR spectrum interpreting systems in many variations and applications.

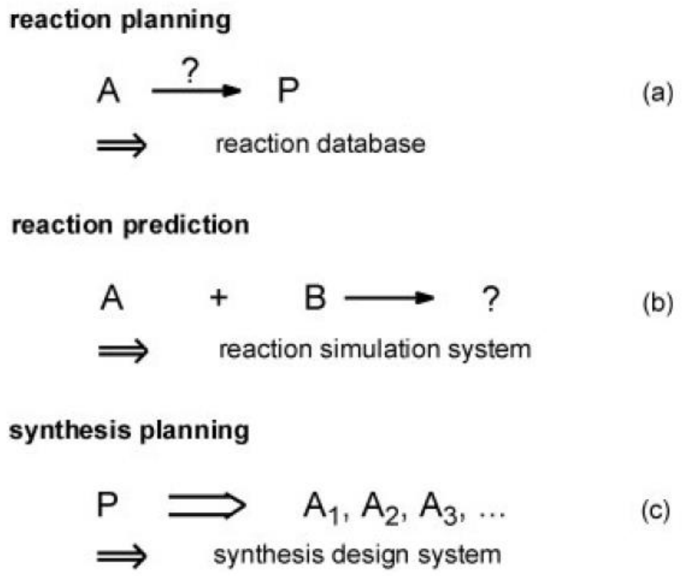
Computer Assisted Synthesis Design

The prediction of chemical reactions

While dealing with chemical reactions, the problems faced are

1. I want to transform a given starting material, A into a desired product P; how can I do this.
   This is a question of reaction planning. To answer such a problem reagents and reaction conditions have to be found that allow one to perform the desired transformation.
2. I have a set of starting material A and B; how they will react and which product will they give.
   This is a question of reaction prediction. To answer simulation of reactions based on knowledge gained from experience is the method of choice.
3. I want to obtain a certain chemical compound P; how can I make it. Which starting materials A1, A2, A3, etc. do I need to build this molecule P.

   Such questions are best answered by a query into a reaction database.



The prediction of the course and of the products of a chemical reaction is of fundamental interest as it concerns a problem with which chemists are constantly faced in their day to day work. They try to solve such questions by making predictions based on analogy, drawing from their experience acquired in their long training or gathered by making a series of experiments.

With the advent of reaction databases chemists now have a treasure full of informations on chemical reactions available at their finger tips. Thus searches in reaction databases might provide an answer to a chemists question on the product of a specific reaction. On top of that, reaction databases can also be used to derive knowledge on chemical reactions which can then be used for reaction prediction.

Two systems, CAMEO and EROS have been developed that a approach the task of reaction prediction in a broad and comprehensive manner.

CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)

Procedures were developed for

1. Base-catalysed and nucleophilic reactions
2. Acid-catalysed and electrophilic reactions
3. Pericyclic reactions
4. Oxidative and reductive reactions
5. Free-radical reactions
6. Carbenoid reactions

Emphasis was put on providing a sound physicochemical basis for determining a reaction mechanism. Methods were developed for the estimation of PKa values, bond dissociation energies, heat of formation, frontier molecular orbital energies and coefficients and steric hinderance.

EROS (Elaboration of reactions for Organic Synthesis)

The knowledge base is essentially two-fold; on one hand it consists of a series of procedures for calculating all important physicochemical effects such as heats of reaction, bond dissociation energies, chrge distribution, inductive, resonance and polarizability effects.

The other part of the knowledge base defines the reaction types on which the EROS system can work.

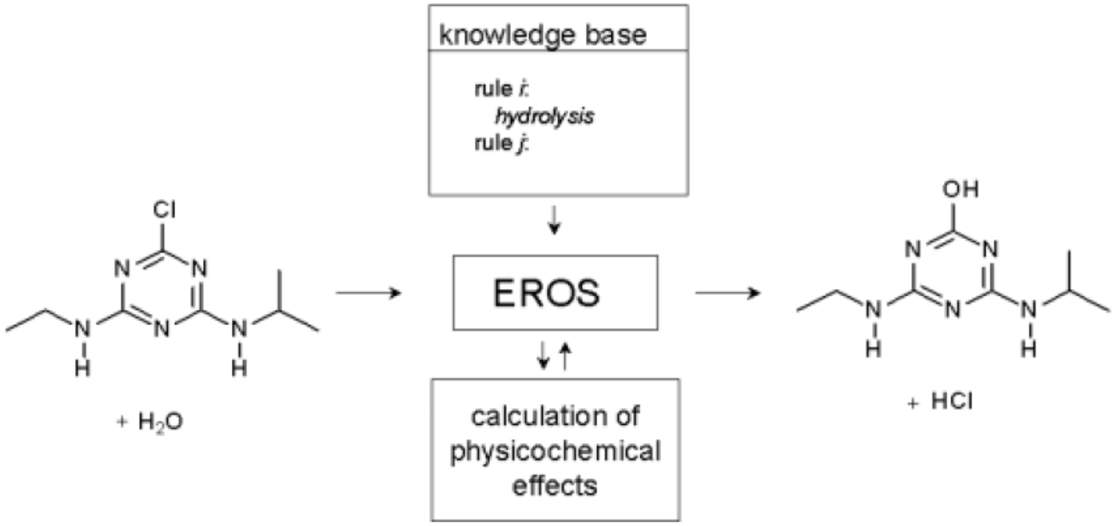


Fig: Outline of the EROS system

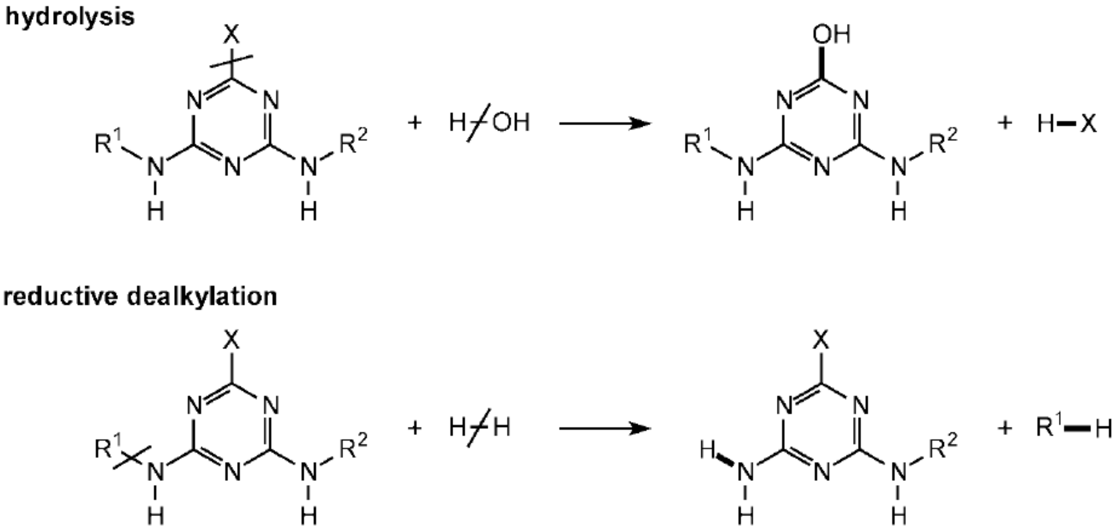


Figure 10.3-8. The two reaction types necessary for modeling the degradation of *s*-triazines in soil.

This latter part consists of an external file, the reaction rules. First, in the reaction rule header, specifications are made about the applicability of the individual reaction type that are to follow furthe down in the rule file. This header might specify that the reaction rules apply to degradation reactions in the environment to biochemical pathways, or to fragmentation and rearrangements in the mass spectrometer. Each reaction rule specifies the bond and electron rearrangement in the course of a reaction type, then gives constraints i.e. information on atoms and bonds to which the specified reaction scheme is applicable.

Computer Assisted Synthesis Design

How do chemists find a pathway to the synthesis of a new organic compound. They try to find suitable starting materials and powerful reactions for the synthesis of the target compound. Thus, synthesis design and chemical reactions are deeply linked, since a chemical reaction is the instrument by which chemists synthesize their compounds; synthesis design is a chemists major strategy to find the most suitable procedure for a synthesis problem.

The synthesis of each compound was considered as a specific task on its own. A suitable strategy for the synthesis of a target compound was mostly found on the basis of the intuition and experience of the acting chemists i.e. the planning of a synthesis ofa complex organic molecule was considered as an art form. No systematic approach was attempted to handle the strategic design of an organic synthesis.

In 1960, Corey introduced a general methodology for planning organic syntheses. The synthesis plan for a target molecule is developed by starting with the target structure and working backwards to available starting materials. The retrosynthetic analysis or disconnection of the target molecule in the reverse direction is performed by the systematic use of analytical rules which have been formulated by Corey.

Concepts for Computer-Assisted Organic Synthesis (CAOS)

The program systems for computer-assisted synthesis planning can be subdivided in to two groups:

1. Information-oriented and
2. logic oriented systems

In Logic-oriented approach generates reactions as bond breaking and bond making steps. These steps are often combined with mechanistic or thermodynamic considerations. In principle, logic oriented systems should not only be able to predict known reactions but should also generate novel reactions. This is both an advantage and a disadvantage. On one hand, such a system may suggest a reaction nobody has forseen and thus the system provides a new approach to the synthesis problem being considered. On the other hand, it may generate a huge number of chemically invalid reactions, which an experienced chemist would intuitively avoid. Therefore, the output of a logic oriented system has to be throughly verified by suitable evaluation techniques.

Information-oriented systems are based on a library of known retro-reactions which have been collected and evaluated by a group of chemists while coding them in electronic form. In addition, information on the scope and the expected yield under various conditions, as well as a strategic merit is usually stored. Such a reaction library is called a knowledge base. In synthesis design programes the knowledge base consists of a database of transforms. Each transform (retro-reaction) has been derived from a number of experimentally performed reactions.

Synthesis Design Systems

LHASA

Works in retrosynthetic manner. It contains a knowledge database of about 2200 transforms and 500 so-called tactical combinations. After a target structure is drawn, the user chooses a strategy for the retrosynthetic analysis. LHASA then searches its knowledge base for transforms which can be applied to the target structure i.e. those for which the retron for the corresponding transformk is present in the target structure. The major strategies are

Functional group bases or short-range strategies: The transform selection is guided by the presence and location of functional groups.

Topological strategies: A strategic bond is selected. The disconnection of a strategic bond should lead to a maximum of simplification. Different strategies can be invoked for different types of bonds, depending on the topology of the target structure (Cyclic strategic bonds, polyfused strategic bonds, ring appendages, acyclic strategic bonds, and manually designate bonds).

Transform-based or long-range strategies: Functional groups are introduced into target compound in order to establish the retron of a certain goal transform (e.g. the transform for the Diels-Alder reaction, Robinson annulation, Birch reduction, etc.)

Stereochemical strategies: The transform selection is guided by stereocenters.

Strategies based on starting materials: The analysis is directed towards a particular starting material.

SYNCHEM

A tool for discovering valid synthetic routes for organic molecules. It provides an environment for the application of artificial intelligence techniques. The user has to draw the target compound and to define some criteria for terminating an anlysis. The major parts of the system are the user interface, a knowledge base and an inference engine. The knowledge base is a library of about 1000 generalised transforms, a library of 5000 available starting materials. The inference engine should discover the most suitable and efficient synthesis path within all reasonable synthesis paths that could be generated by a systematic application of the knowledge base to the target compound.

SYNGEN

A simple mathematical model for the logical description of structures and reactions which was later on implemented in the program system SYNGEN. It generates all conceivable approaches to construct carbon skeleton of the target structure within defined constraints and without user intervention. A major difference from other synthesis design programs is that no external knowledge base is needed for defining transforms or reaction scheme for the retrosynthetic analysis. The central feature of a synthesis strategy generated by SYNGEN is the assembly of the carbon skeleton in the target from the skeletons of available starting materials.

The retrosynthetic analysis is performed in two steps:

In the first step it dissects the skeleton to find all fullyconvergent bondsets which utilises starting material skeletons found in two successive levels of cuts.

In the second step, the necessary functionality is generated for each of such bondsets at the respective carbon skeletons in order to make the synthesis feasible.

ALPHOS

It is an interactive system which performs the retrosynthetic analysis in a stepwise manner, determining at each step the synthesis precursors from the molecules of the preceding step. ALPHOS tries to combine the merits of a knowledge based approach with those of a logic centered approach.

Definition of terms

Lead Structure

A representative of a compound series with sufficient potential (as measured by potency, selectivity, pharmacokinetics, pysicochemical properties, absence of toxicity and novelty) to progress to a full drug development program.

Ligand: A ligand is a molecule binding to a biological macromolecule.

Enzyme: Enzymes are endogenous catalysts converting one or several substrates into one or several products.

Substrate: A substrate is the starting material of an enzymatic reaction

Inhibitor: A ligand preventing the binding of a substrate to its enzyme is called an inhibitor

Receptor: Receptors are membrane-bound or soluble proteins or protein complexes exerting a physiological effect after binding of an agonist.

Agonist: An agonist is a receptor ligand preventing the action of an agonist in a direct (competitive) or indirect (allosteric) manner.

Antagonist: An antagonist is a receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.

Role of Cheminformatics in Morden Drug Discovery

Recent chemical developments for drug discovery are generating a lot of chemical data which is referred as information explosion. This has created a demand to effectively collect, organize, analyse and apply the chemical information in the process of modern d rug discovery and development. The drug discovery process is aimed at discovering molecules that can be very rapidly developed for effective treatments to meet medical needs. The entanglement of chemistry and information management started in the mid of 1970s, applying in the area of prediction of protein structure, Fourier transform of X-ray crystallography, enzyme and chemical kinetics, analyse various types of spectroscopy data and binding of chemical compounds. During early 1980s, computer technology is considered as the core component by the medical chemist to solve chemical problems. For example, collecting crystal structures of small molecules in Cambridge Structural Database (CSD) provides a fertile resource for geometrical data on molecular fragments for calibration of force fields and validation of results from computational chemistry. The need of storing macromolecular data results in Protein Data Base (PDB). The needs and refinement on these approaches result in several tools and upgrading the process of solving the problems.

The traditional drug discovery process starts with a particular Disease, Identification of target, Identification of molecule effective against target and Preclinical testing. Identification of target and synthesis the molecule to increase their suitability takes more amounts of time and cost (in millions) discovery process of the drug. The development process starts with human clinical trials, approval from authority and delivers the product in the market. This process takes about 10-15 years to discover, develop and bring drug to the market.

The modern pharmaceutical drug discovery and development pipeline process, as shown in Fig. 7, starts with Disease selection, Target identification, Lead identification, Lead Optimization, Pre -clinical trial testing, Clinical trial testing, Approval and circulation (Drug in market). In traditional drug discovery phase, the process which cost more time and money is replaced with lead identification and lead optimization process in modern drug discovery system. Each phase has an interaction component that transfers data, knowledge and information to one another (shown in Fig. 8).

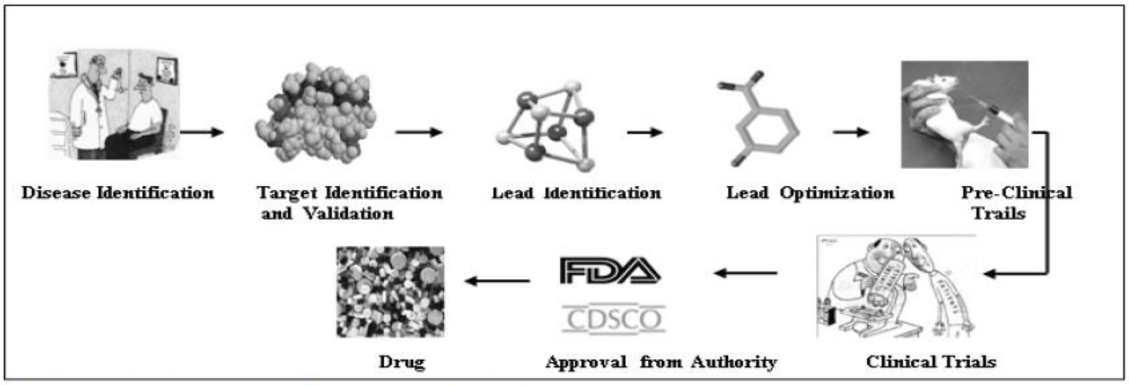


Fig. 7. Modern Drug Discovery and Development Life Cycle

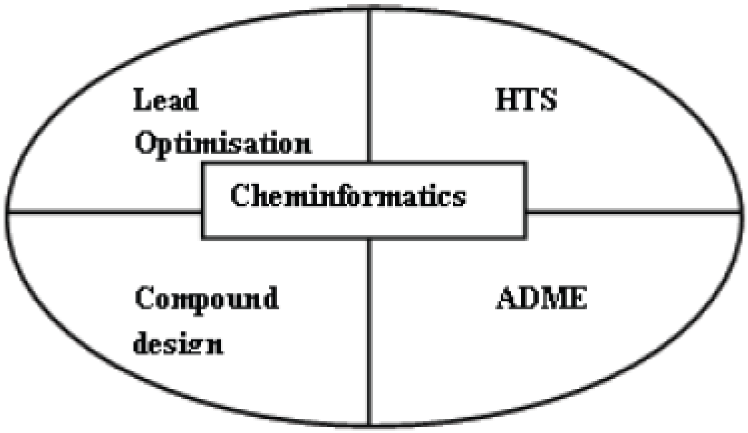


Fig. 8. Interaction Process

4.1 Pre Drug Discovery Process (Disease identification)

Before the discovery process starts, it is to understand the disease by knowing, how the genes can be altered, how it affects the protein, how these protein will react with each other in living organism, how the affected cells can change the specific tissues and how the disease affects the patient. All the phases in Pre-Discovery, Discovery and Development of Drug is shown in Fig. 9.

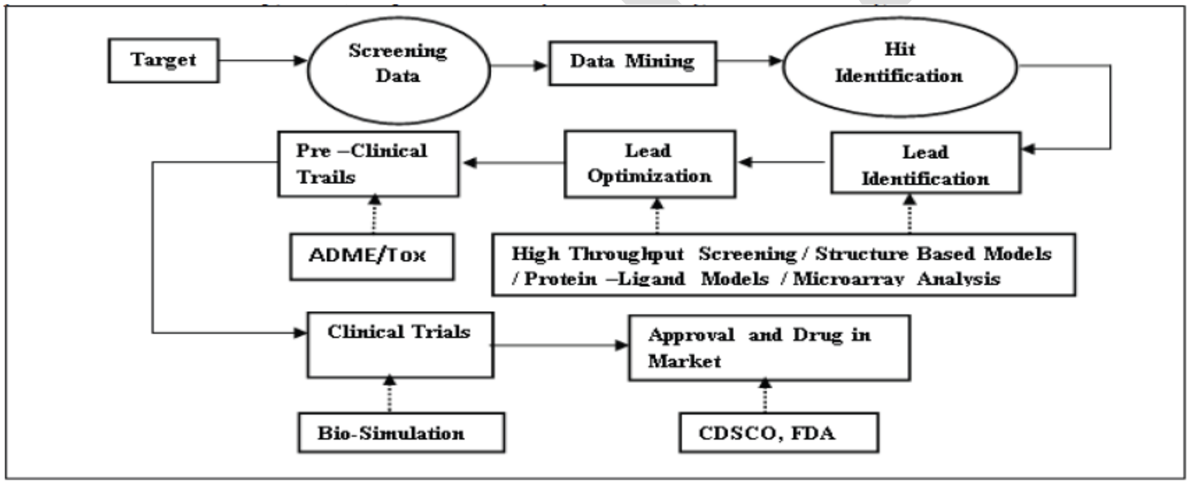


Fig. 9. Drug Discovery and Development Process

4.2 Morden Drug Discovery Process

The discovery process includes four important processes such as, target identification and validation, lead identification, lead optimization and pre-clinical trials.

a) Target Identification & Validation

Cheminformatics is used to identify target molecule which can be either gene or protein and could be a potential drug for the disease (Gene/Protein analysis). The Identified protein is separated, crystallized and ligand binding processes are done. Some approaches will inhibit the disease functionality by making the key molecule stop functioning. Another approach is by promoting specific molecule in the normal way which may have affected in the disease state. These approaches and different databases can be applied for the discovery of drug targets. After target Identification, validation phase starts by determining whether the modulation of the target will y ield a desired clinical outcome. This is base d on the results obtained between the cellular location and disease/health condition, potential expression and protein binding state.

b) Lead Identification.

Target -to-Screening (HTS) technique is applied where the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds. Lead identification also helps to see which molecules bind strongly to the target. Several similarity and diversity techniques can be applied for lead identification.

c) Lead Optimization

This phase results in finding the drug candidate from the lead identified compound. The goal is a process of refining the chemical structure of a confirmed hit to improve its drug characteristics. Several docking techniques can be applied to optimize the lead structures for target affinity and selectivity.

Different techniques and methods are used for Lead identification and Optimization process where some of the methods Virtual Screening, Molecular Database, Data mining, High-Throughput Screening (HTS), QSAR, Protein Ligand Models, Structure Based Models, Microarray analysis, Property Calculation and ADMET(adsorption, distribution, metabolism and elimination and Toxicity).

d) Pre- Clinical Trial

The preclinical stage is an important phase to check whether the compound can be made into a drug to treat specific disease which is not toxic and has minimum side effects. Toxicity tests are undertaken to show safety while pharmacokinetics testing is done to provide data on how a drug is absorbed, distributed, metabolised and excreted (ADME) from the body. Pre-clinical studies and testing can be done with or without animal testing method. In-vitro is a study, based on the test done in the clinical lab and the analysis based on living cell cultures and animal model can be referred as in –vivio method. This phase will be designed in a way such that it achieves risk-free, unproblematic and economic transition from pre-clinical to clinical trial in medical product development.

4.3 Development Process

Development process is another significant stage in the life cycle of finding new d rug. This stag e consists of three major phase such as clinical trial, approval from the authority and drug in market.

i) Clinical Trial

Clinical trial is the primary phase which will be fastest and safest way to find treatments which acts as the best solution for challenging health disease of human being. Patient with specific disease will be considered for clinical trial and relevant data will be co llected with respect to the time. Trials can be done in five ways such as, prevention trials, screening trails, diagnostic trails, treatment trial and quality of life trial.

ii) Approval from the Authority and Drug in Market

Based on the rules and regulation for new drug development in the country as well as in international market, research authorities check the safety and other parameters to approve the drug for marketing. Central Drugs Standard Control Organization (CDSCO) in India and Food and Drug Administration (FDA) in U.K. approves a new pharmaceutical compound for sales and marketing.

5. Conclusion

Average life span of Human being is gradually decreasing in the recent medical history due to the higher influence of new d iseases. Identifying and understanding structural and functional behaviour of chemical compounds/biomolecules are one of the challenging issues for medical researchers. Cheminformatics is an emerging field which is used for better understanding of biomolecules. This paper primarily focuses on cheminformatics and its applications on drug discovery, issues of traditional discovery and importance of modern drug discovery system. This in turn helps chemists and researchers for developing drugs without side effects.

Analysis of HTS data

A central problem in cheminformatics is the establishment of a relationship between a chemical structure and its biological activity. Huge amounts of data are gathered, particularly so through the synthesis of combinatorial libraries and subsequent high-throughput screening (HTS). A chemist synthesises about 50 compounds per year by traditional, organic synthesis. In combinatorial chemistry a series of homologs are synthesised. Thousands of compounds are accessible in a short period of time.

With these massive amounts of data produced in HTS for combinatorial libraries tools become necessary to make it possible to navigate through these data and to extract the necessary information to search- as is said quite often for a needle in a haystack.

In order to extract information from huge quantities of data and to gain knowledge from this information, the analysis and emploration have to be performed by automatic or semi-automatic methods.

Virtual screening

Virtual screening or *in silico* screening is defined as the selection of compounds by evaluating their desirability in a computational model. The desirability comprises high potency, selectivity, appropriate pharmacokinetic properties and favourable toxicology. Virtual screening assists the selection of compounds for screening in-house libraries and compounds from external suppliers. Two different strategies can be applied:

1. Diverse libraries can be used for lead finding by screening against several different targets. The selected compounds should cover the biological activity space well
2. Targeted or focussed libraries are suited for both lead finding nd optimisation. If knowledge of a lead compound is available, compounds with a similar structure are selected for the targeted library. Targeted libraries are focussed on a single target.

Virtual screening allows the scope of screening to be extended to external data bases. When this is done, increasingly diverse hits can be identified. The application of virtual screening techniques before or in parallel with HTS helps to reduce the assay-to-lead attrition rate observed from HTS. In addition, virtual screening is faster and less expensive than experimental synthesis and biological testing. Both ligand and structure based methods can be applied in virtual screening.

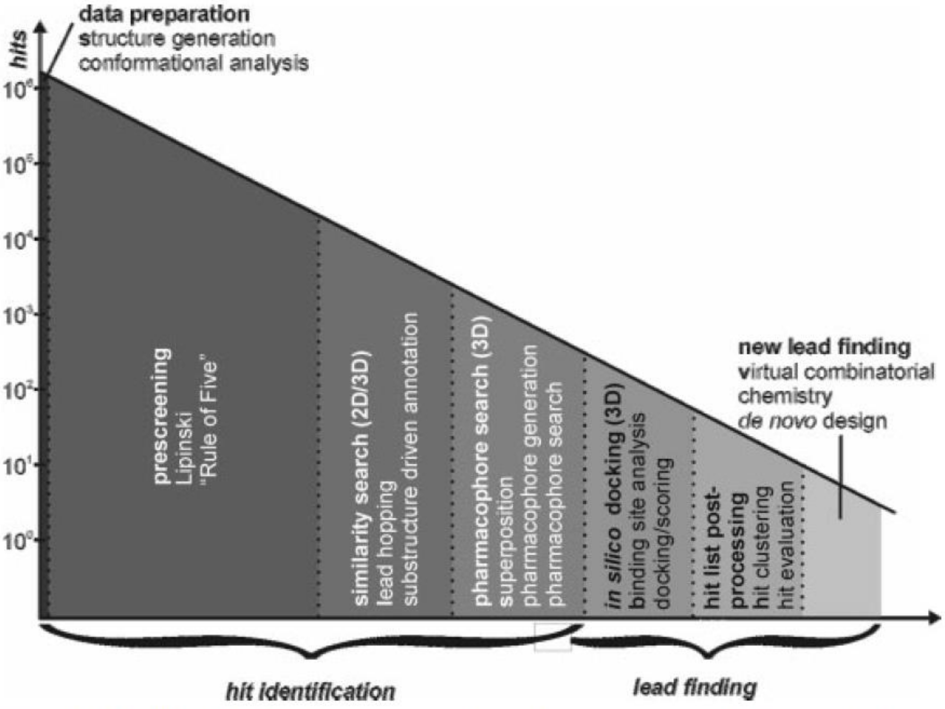


Figure 10.4-3. Work flow for virtual screening, from data preparation to finding new leads.

Design of Combinatorial Libraries

HTS data as wella as virtual screening can guide and direct the design of combinatorial libraries. A genetic algorithm (GA) can be applied to the generation of combinatorial libraries. The number of compounds accessible by combinatorial synthesis often exceeds the number of compounds which can be synthesised experimentally. To reduce the number of products a subset of fragments has to be chosen.

Ligand and structure based drug design

Depending on the information available about the protein structure and the ligands binding to a particular target, four different cases can be distinguished in drug design.

Table: Cases in the drug discovery process depending on knowledge of the receptor and ligand structure

	Ligand unknown	Ligand known

Protein structure unknown	Combinatorial chemistry and HTS	QSAR, pharmacophore models and hypothesis, similiarity search in databases
Protein structure known	*De novo* design, receptor based 3D searching	Structure based design, docking

The lead discovery process is as follows.

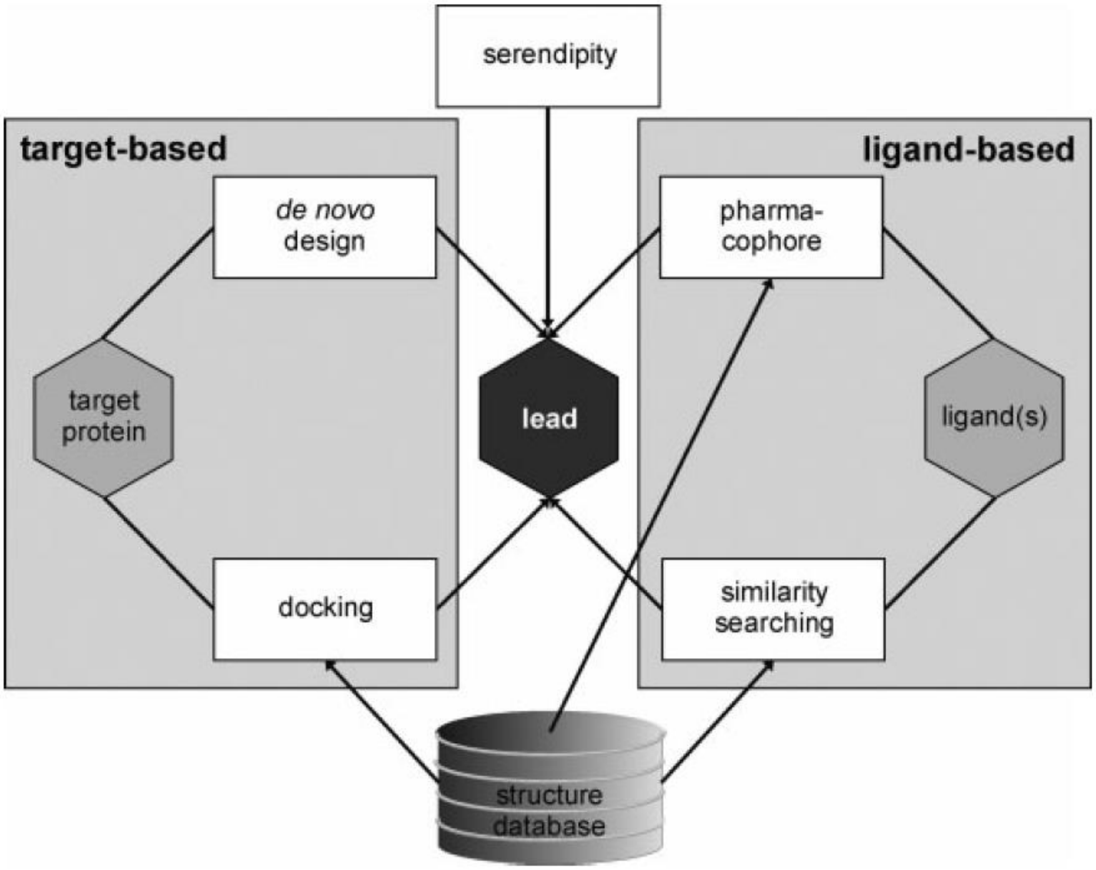


Fig: Lead discovery process

The lead structure can be discovered by serendipity. In rational drug design all information available about a target serves to direct the search for a new lead structure. If the 3D structure of the target of interest is known from X-ray crystallography, a number of methods in structure-based drug design can be applied. If an X-ray structure of the protein with a ligand is available, the binding

Mode of the ligand can be analysed. Docking of new ligands with the same binding mode, and ranking of these ligands by scoring functions, guide further drug development. Otherwise, one has to apply de novo design or perform a 3D search in a ligand database for a compound with a complementary shape and surface properties to the binding site of the receptor.

In an early stage of drug design, without structural information, about the target and without any knowledge about ligands binding to the target protein, one is obliged to screen combinatorial and proprietary libraries by HTS. As soon as some of the ligands binding to the target are known, they serve as a starting point for a similiarity search in a ligand database and for the perception of a pharmacophore by superposition of the ligands. Thus there is a distinction between ligand and structure based drug design.

Ligand based drug design

Prefiltering in Virtual Screening:

In general, the first step in virtual screening is the filtering by the application of Lipinski's "Rule of Five". A lead molecule should have

a. Molecular weight of less than 500 g/mol
b. A calculated lipophilicity (log P) of less than 5

c. Fewer than five H-bond donors
d. Fewer than 10H-bond acceptoprs (sum of all nitrogen and oxygen atoms)
e. Number of rotatable bonds is less than 10 or one of the four rules can be violated.

In a more recent study the physical properties of drugs in different development phases are compared. The molecular weight and lipophilicity are the properties showing the clearest influence on the successful passage of a candidate drug through the developmental process.

In Silico ADMET

Pharmaceutical companies evaluate the ADMET profiles (drug absorption, distribution, metabolism, excretion and toxicity) of potential leads at an earlier stage of the development process. For the consideration of ADMET properties in virtual screening, computational methods for their

Prediction is needed. Lipophilicity is a key property for estimation of the membrane permeability of a molecule. Programs to predict Log P are available and give reasonable results.

Similiarity Searches

Following the “similar structure-similar property” principle, structurally similar molecules are expected to exhibit similar properties or biological activities. In order to select compounds for focussed libraries, 2D and 3D similiarity searches are performed. The distance of all the ligands in a database from the ligands known to bind to the target of interest is calculated. Afterwards the ligands are ranked in reverse order of their distance. The ligands with a high rank are selected for the focussed library.

Structure based drug design

Fitting a ligand from a 3D structure database into the binding site of a target protein is called docking. The iterative building of new molecules in thebinding site of a receptor are called *de novo* design. The building approach is beginning with a single fragment and proceeding through the stepwise addition of further moieties. Alternatively small molecules are placed in the binding site of the protein and subsequently linked together (linking).

To end up with high-affinity ligands a high degree of steric and electronic complementarity of the ligand to the target protein is required. Further on an appropriate amount of the ligands hydrophobic surface should be buried in the complex. A certain degree of conformational rigidity is essential to ensure that the loss of entropy upon ligand biding is acceptable.

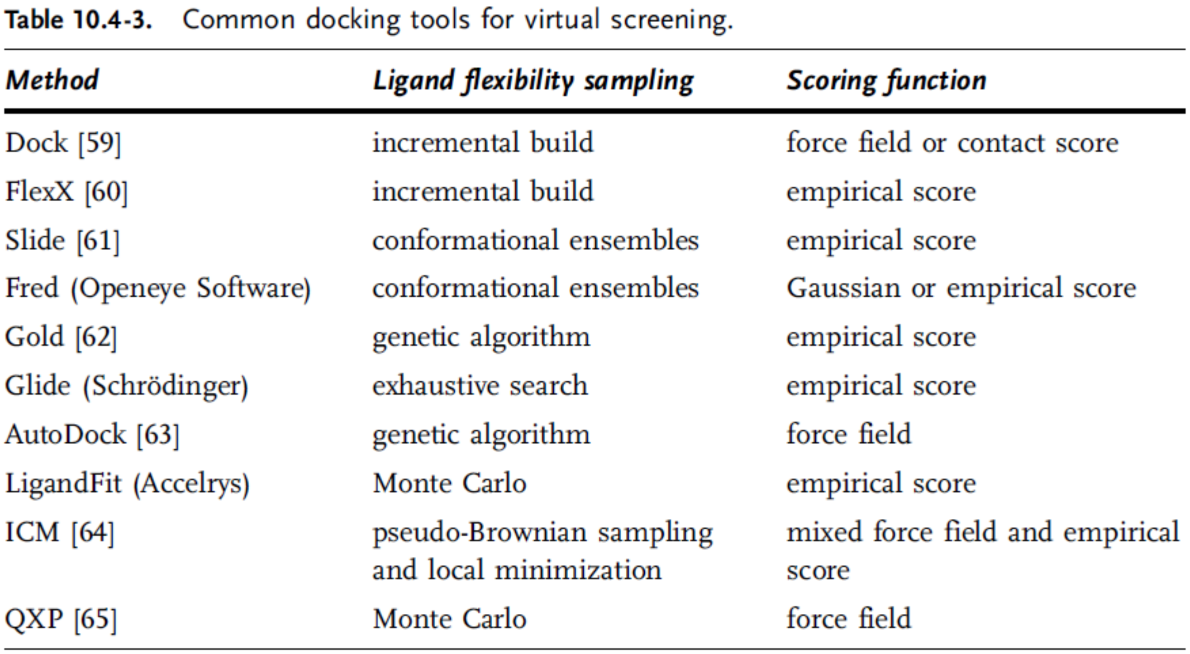


Table 10.4-3. Common docking tools for virtual screening.

Method	*Ligand flexibility sampling*	*Scoring function*
Dock [59]	incremental build	force field or contact score
FlexX [60]	incremental build	empirical score
Slide [61]	conformational ensembles	empirical score
Fred (Openeye Software)	conformational ensembles	Gaussian or empirical score
Gold [62]	genetic algorithm	empirical score
Glide (Schrödinger)	exhaustive search	empirical score
AutoDock [63]	genetic algorithm	force field
LigandFit (Accelrys)	Monte Carlo	empirical score
ICM [64]	pseudo-Brownian sampling and local minimization	mixed force field and empirical score
QXP [65]	Monte Carlo	force field

Possible Questions

Part A (Online multiple choice questions)

Part B (Each question carries two marks)

1. What is meant by CONGEN and GENOA techniques used in computer assisted structure elucidation
2. What are the procedures developed in CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions).
3. What is meant by a ligand in modern drug discovery
4. What is meant by a lead structure in modern drug discovery.
5. What is meant by an enzyme in modern drug discovery
6. What is meant by a substrate in modern drug discovery
7. What is meant by a inhibitor in modern drug discovery
8. What is meant by a receptor in modern drug discovery
9. Differentiate a agonist and a antagonist
10. How a drug target is identified.
11. What is meant by lead identification

12. What is meant by lead optimization
13. What is meant by HTS.
14. What is meant by ADMET profile of a drug.

Part C (Each question carries eight marks)

1. Explain the techniques used in computer assisted structure elucidations.
2. Explain the problems faced in dealing with chemical reactions in a computer assisted synthetic design.
3. Explain the CAMEO and EROS systems of computer assisted synthetic design.
4. Explain about the LHASA system of synthesis design system.
5. What are the major strategies followed in LHASA system of synthesis design system.
6. Explain in detail SYNCHEM and SYNGEN system of computer assisted synthetic designs.
7. Explain in detail the role of cheminformatics in drug discovery
8. Explain the process involved in the pre drug discovery process
9. Explain any four process involved in the modern drug discovery process
10. Explain the process taking place in the pre clinical trials.
11. Explain the terms (i) high-throughput screening (HTS) and (ii) virtual screening
12. Describe the Ligand and structure based drug design

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore – 641 021.

UNIT V (Multiple choice Questions)

S.No	Question	A	B	C	D	Answer
1	In computer assisted structure elucidations the programme can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures	CONGEN	DARC-EPIOS	GENIUS	ANN	CONGEN
2	In computer assisted structure elucidations the programme could retrieve structural formulas from overlapping ^{13}C NMR data.	CONGEN	DARC-EPIOS	GENIUS	ANN	DARC-EPIOS
3	In computer assisted structure elucidations the programme Implements genetic algorithms to approach the problem of finding structures	CONGEN	DARC-EPIOS	GENIUS	ANN	GENIUS
4	In computer assisted structure elucidations	CONGEN	DARC-EPIOS	GENIUS	ANN	ANN

	the programme allows the determination of adequate functional groups that can exist in the sample					
5	ANN Programme for the computer assisted structure elucidations	allows the determination of adequate functional groups that can exist in the sample	Implements genetic algorithms to approach the problem of finding structures	could retrieve structural formulas from overlapping ^{13}C NMR data.	can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures	allows the determination of adequate functional groups that can exist in the sample
6	GENIUS Programme for the computer assisted structure elucidations	allows the determination of adequate functional groups that can exist in the sample	Implements genetic algorithms to approach the problem of finding structures	could retrieve structural formulas from overlapping ^{13}C NMR data.	can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures	Implements genetic algorithms to approach the problem of finding structures
7	DARC-EPIOS Programme for the computer assisted structure elucidations	allows the determination of adequate functional groups that can exist in the sample	Implements genetic algorithms to approach the problem of finding structures	could retrieve structural formulas from overlapping ^{13}C NMR data.	can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures	could retrieve structural formulas from overlapping ^{13}C NMR data.

8	Procedures were developed for Base-catalysed and nucleophilic reactions	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)	EROS (Elaboration of reactions for Organic Synthesis)	**LHASA**	CONGEN	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
9	CONGEN Programme for the computer assisted structure elucidations	allows the determination of adequate functional groups that can exist in the sample	Implements genetic algorithms to approach the problem of finding structures	could retrieve structural formulas from overlapping ^{13}C NMR data.	can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures	can handle any structure and enumerate the isomers of a molecular formulae and were able to generate structures
10	Procedures were developed for Acid-catalysed and electrophilic reactions	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)	EROS (Elaboration of reactions for Organic Synthesis)	**LHASA**	CONGEN	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
11	Procedures were developed for Pericyclic reactions	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)	EROS (Elaboration of reactions for Organic Synthesis)	**LHASA**	CONGEN	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)

12	Procedures were developed for Oxidative and reductive	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)	EROS (Elaboration of reactions for Organic Synthesis)	**LHASA**	CONGEN	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
13	Procedures were developed for Free radical reactions	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)	EROS (Elaboration of reactions for Organic Synthesis)	**LHASA**	CONGEN	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
14	Procedures were developed for Carbenoid reactions	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)	EROS (Elaboration of reactions for Organic Synthesis)	**LHASA**	CONGEN	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
15	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions) can be used to measure	Physicochemical property PKa values	polarizability effects.	Inductive effects	Resonance effects	Physicochemical property PKa values
16	CAMEO (Computer Assisted Mechanistic Evaluation of Organic	heat of formation	polarizability effects.	Inductive effects	Resonance effects	heat of formation

	reactions) can be used to measure					
17	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions) can be used to measure	frontier molecular orbital energies	polarizability effects.	Inductive effects	Resonance effects	frontier molecular orbital energies
18	CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions) can be used to measure	steric hinderance.	polarizability effects.	Inductive effects	Resonance effects	steric hinderance.
19	EROS (Elaboration of reactions for Organic Synthesis) can be used to measure	steric hinderance.	frontier molecular orbital energies	polarizability effects.	heat of formation	polarizability effects.
20	EROS (Elaboration of reactions for Organic Synthesis) can be used to measure	steric hinderance.	Inductive effects	polarizability effects.	heat of formation	Inductive effects
21	EROS (Elaboration of reactions for Organic Synthesis) can be used to measure	steric hinderance.	Resonance effects	polarizability effects.	heat of formation	Resonance effects
22	LHASA synthetic design system	Works in retrosynthetic manner	provides an environment for the application	model for the logical	an interactive system which performs the	Works in retrosynthetic manner

			of artificial intelligence techniques	description of structures and reactions which was later on implemented in the program system	retrosynthetic analysis in a stepwise manner	
23	SYNCHEM synthetic design system	Works in retrosynthetic manner	provides an environment for the application of artificial intelligence techniques	model for the logical description of structures and reactions which was later on implemented in the program system	an interactive system which performs the retrosynthetic analysis in a stepwise manner	provides an environment for the application of artificial intelligence techniques
24	SYNGEN synthetic design system	Works in retrosynthetic manner	provides an environment for the application of artificial intelligence techniques	model for the logical description of structures and reactions which was later on implemented in the program system	an interactive system which performs the retrosynthetic analysis in a stepwise manner	model for the logical description of structures and reactions which was later on implemented in the program system
25	ALPHOS synthetic design system	Works in retrosynthetic manner	provides an environment for the application of artificial intelligence techniques	model for the logical description of structures and reactions which was later on	an interactive system which performs the retrosynthetic analysis in a stepwise manner	an interactive system which performs the retrosynthetic analysis in a stepwise manner

				implemented in the program system		
26	The synthetic design system Works in retrosynthetic manner	ALPHOS	SYNGEN	**SYNCHEM**	LHASA	LHASA
27	The synthetic design system provides an environment for the application of artificial intelligence techniques	ALPHOS	SYNGEN	**SYNCHEM**	LHASA	**SYNCHEM**
28	The synthetic design system provides model for the logical description of structures and reactions which was later on implemented in the program system	ALPHOS	SYNGEN	**SYNCHEM**	LHASA	SYNGEN
29	The synthetic design system provides an interactive system which performs the retrosynthetic analysis in a stepwise manner	ALPHOS	SYNGEN	**SYNCHEM**	LHASA	ALPHOS
30	A ligand is	a molecule binding to a biological macromolecule.	endogenous catalysts converting one or several substrates into one or several products.	the starting material of an enzymatic reaction	A substance preventing the binding of a substrate to its enzyme	a molecule binding to a biological macromolecule.

31	An enzyme is a	a molecule binding to a biological macromolecule.	endogenous catalysts converting one or several substrates into one or several products.	the starting material of an enzymatic reaction	A substance preventing the binding of a substrate to its enzyme	endogenous catalysts converting one or several substrates into one or several products.
32	A substrate is the	a molecule binding to a biological macromolecule.	endogenous catalysts converting one or several substrates into one or several products.	the starting material of an enzymatic reaction	A substance preventing the binding of a substrate to its enzyme	the starting material of an enzymatic reaction
33	An Inhibitor is	a molecule binding to a biological macromolecule.	endogenous catalysts converting one or several substrates into one or several products.	the starting material of an enzymatic reaction	A substance preventing the binding of a substrate to its enzyme	A substance preventing the binding of a substrate to its enzyme is called an inhibitor
34	A substance preventing the binding of a substrate to its enzyme is called	an inhibitor	A substrate	**An enzyme**	A ligand	an inhibitor
35	the starting material of an enzymatic reaction	an inhibitor	A substrate	An enzyme	A ligand	A substrate

36	endogenous catalysts converting one or several substrates into one or several products.	an inhibitor	A substrate	An enzyme	A ligand	**An enzyme**
37	a molecule binding to a biological macromolecule	an inhibitor	A substrate	An enzyme	A ligand	A ligand
38	membrane-bound or soluble proteins or protein complexes exerting a physiological effect after binding of a drug is called	Receptor	agonist	antagonist	Lead molecule	Receptor
39	A receptor ligand preventing the action of a antagonist in a direct (competitive) or indirect (allosteric) manner.	Receptor	agonist	antagonist	Lead molecule	agonist
40	a receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.	Receptor	agonist	antagonist	Lead molecule	Antagonist
41	A receptor is	soluble proteins exerting a	A receptor	a receptor ligand	a compound series with	soluble proteins exerting a

		physiological effect after binding of a drug	ligand preventing the action of a antagonist in a direct (competitive) or indirect (allosteric) manner.	preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.	sufficient potential for drug development	physiological effect after binding of a drug
42	An agonist is	soluble proteins exerting a physiological effect after binding of a drug	A receptor ligand preventing the action of a antagonist in a direct (competitive) or indirect (allosteric) manner.	a receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.	a compound series with sufficient potential for drug development	A receptor ligand preventing the action of a antagonist in a direct (competitive) or indirect (allosteric) manner.
43	An antagonist is	soluble proteins exerting a physiological effect after binding of a drug	A receptor ligand preventing the action of a antagonist in a direct (competitive) or indirect (allosteric) manner.	a receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.	a compound series with sufficient potential for drug development	a receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.

44	A lead molecule is	soluble proteins exerting a physiological effect after binding of a drug	A receptor ligand preventing the action of a antagonist in a direct (competitive) or indirect (allosteric) manner.	a receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.	a compound series with sufficient potential for drug development	a compound series with sufficient potential for drug development
45	In drug discovery process to identify target molecule which can be either gene or protein and could be a potential drug for the disease	Target identification	Lead identification	Lead optimization	Preclinical trial	Target identification
46	In drug discovery process where the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds.	Target identification	Lead identification	Lead optimization	Preclinical trial	Lead identification
47	In drug discovery finding the drug candidate from the lead identified compound	Target identification	Lead identification	Lead optimization	Preclinical trial	Lead optimization
48	Target identification means	to identify target molecule which	the protein targets are	finding the drug	phase to check whether the	to identify target molecule which

		can be either gene or protein and could be a potential drug for the disease	automatically screened against database of small-molecule o r cell-based assay compounds.	candidate from the lead identified compound	compound can be made into a drug to treat specific disease	can be either gene or protein and could be a potential drug for the disease
49	Lead identification in drug discovery means	to identify target molecule which can be either gene or protein and could be a potential drug for the disease	the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds.	finding the drug candidate from the lead identified compound	phase to check whether the compound can be made into a drug to treat specific disease	the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds.
50	Lead optimization means	to identify target molecule which can be either gene or protein and could be a potential drug for the disease	the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds.	finding the drug candidate from the lead identified compound	phase to check whether the compound can be made into a drug to treat specific disease	finding the drug candidate from the lead identified compound
51	Preclinical trial means	to identify target molecule which can be either gene or protein and could be a potential drug for the disease	the protein targets are automatically screened against database of small-molecule o r cell-based assay	finding the drug candidate from the lead identified compound	phase to check whether the compound can be made into a drug to treat specific disease	phase to check whether the compound can be made into a drug to treat specific disease

			compounds.			
52	the selection of compounds by evaluating their desirability in a computational model	Virtual screening	high-throughput screening	Design of Combinatorial Libraries	Lead discovery process	Virtual screening
53	a relationship between a chemical structure and its biological activity	Virtual screening	high-throughput screening	Design of Combinatorial Libraries	Lead discovery process	high-throughput screening
54	the selection of compounds by evaluating their desirability in a computational model	*In-silico* screening	high-throughput screening	Design of Combinatorial Libraries	Lead discovery process	*In-silico* screening
55	Which deviates from Lipinski's "Rule of Five	Molecular weight of more than 500 g/mol	lipophilicity (log P) of less than 5	Fewer than five H-bond donors	Fewer than 10H-bond acceptoprs	Molecular weight of more than 500 g/mol
56	Which deviates from Lipinski's "Rule of Five	Molecular weight of less than 500 g/mol	lipophilicity (log P) of more than 5	Fewer than five H-bond donors	Fewer than 10H-bond acceptoprs	lipophilicity (log P) of more than 5
57	Which deviates from Lipinski's "Rule of Five	Molecular weight of less than 500 g/mol	lipophilicity (log P) of less than 5	More than five H-bond donors	Fewer than 10H-bond acceptoprs	More than five H-bond donors

58	Which deviates from Lipinski's "Rule of Five	Molecular weight of less than 500 g/mol	lipophilicity (log P) of less than 5	Fewer than five H-bond donors	More than 10H-bond acceptoprs	More than 10H-bond acceptoprs
59	Which is not in ADMET profiles	drug absorption	Drug potency	Drug metabolism	Drug excretion	Drug potency
60	Which is not in ADMET profiles	drug absorption	Drug distribution	Drug stability	Drug toxicity	Drug stability

[16CHU501A] **Reg. No------------------**

Karpagam Academy of Higher Education

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore-21

(For the candidates admitted from 2016 onwards)

III B.Sc. Chemistry

Cheminformatics

Internal Test-I

Date: Maximum: 50 marks

Time: 2hrs

PART A (20 x 1 = 20 Marks)
Answer all the questions

1. The subject which plays a key role to maintain and access enormous amount of chemical data, produced by chemist is
   a. Statistics b. Bioinformatics c.**Cheminformatics** d. Vector algebra

2. The a process of generating and collecting data empirically (experimentation) or from theory is called as
   a. **Information Acquisition**, b. Information Management, c. Information use d. molecular simulation

3. The process which deals with storage and retrieval of chemical information
   a. Information Acquisition,b. **Information Management**, c. Information use d. molecular simulation

4. The process which includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences
   a. Information Acquisition,b. Information Management, c. **Information use** d. molecular simulation

5. The first information systems and services were paper based and the Chemical Abstracts was started in the year
   a. **1907** b. 1961 c. 1973 d. 1998

6. The computer-programme such as *Chemdraw* is used to form a
   a. **Image file** b. Graph theory c. Connection table d. Linear notation

7. An abstract structure that contains nodes connected by edges
a. Image file b. **Graph theory** c. Connection table d. Linear notation

8. A list of the bonds, stomic numbers, bonds, hybridization state and bond order constitute
a. Image file b. Graph theory c. **Connection table** d. Linear notation

9. A chemical tool box which interconverts chemical structures between different formats is
a. **Open Babel** b. Amber tools c. Chem draw d. Chem sketch

10. Cheminformatics mainly dealt with
a. **Small molecules** b. Macromolecules c. Polymers d. Nano particles

11. Bioinformatics has mainly dealt with
a. **Genes & Proteins** b. Macromolecules c. Enzymes d. Polymers

12. Chemical structure representations can be
a. **Linesr, 2D or 3D** b. Only Linear c. Only 2D d. Only 3D

13. SMILES is one of the following chemical structure representation
a. **Linear** b. 2D c. 3D d. Modelled.

14. The integration of technolopgies to quickly screen chemical compounds in search of a desired activity
a. **HTS** b. Docking c. drug discovery d. QSAR

15. A molecule editor developed by the cheminformatics
a. **Chemdraw** b. Chemwindow c. ISIS-draw d. Chem Reader

16. A study of interaction of drug molecule and a protein is called
a. HTS b. **Docking** c. drug discovery d. QSAR

17. The 1998 Nobel Chemistry Prize was awarded to Pople and Kohn for their work in
a. **Computational Chemistry and Molecular Modelling** b. Nanotechnology c. Green chemistry d. Electrochemistry

18. In a chemical database, structural data refers to
a. **2D representation of a molecule** b. Log P value c. Experimental notes d. Spectra or plots

19. In a chemical database, Numerical data refers to
a. 3D representation of a molecule b. **Log P value** c. Experimental notes d. Spectra or plots

20. The experimental notes that are associated with a structure or data point will be found in
    a. Structural data base b. Numerical data base c**. Annotation data base** d. Graphical data base

PART B (3 x 2 = 6 marks)
Answer all the questions

21. What is meant by cheminformatics.
22. Differentiate cheminformatics and Bioinformatics
23. Write notes on structure descriptors.

PART C (3 x 8 = 24 marks)
Answer all the questions

24.a. Explain the need and importance of cheminformatics

OR

24.b. What are the applications of cheminformatics.

25.a. Explain the tools used for cheminformatics

OR

25.b. Discuss the role of cheminformatics in Modern drug discovery

26.a. Write and explain some of the chemical structure representation of Caffine

OR

26.b. Explain the structure of MOL files and SD files.

b.

[16CHU501A] **Reg. No------------------**

Karpagam Academy of Higher Education

(Deemed to be University)

(Established Under Section 3 of UGC Act, 1956)

Coimbatore-21

(For the candidates admitted from 2016 onwards)

III B.Sc. Chemistry

Cheminformatics

Internal Test-I- ANSWER KEY

PART A (20 x 1 = 20 Marks)

Answer all the questions

1. c. **Cheminformatics**
2. **a. Information Acquisition**,
3. b. **Information Management**
4. c. **Information use**
5. a. **1907**
6. **Image file**
7. b. **Graph theory**
8. c. **Connection table**
9. a. **Open Babel**
10. a. **Small molecules**
11. **a. Genes & Proteins**
12. **a. Linesr, 2D or 3D**
13. a. **Linear**
14. **a. HTS**
15. **a. Chemdraw**
16. b. **Docking**
17. a. **Computational Chemistry and Molecular Modelling**
18. **a. 2D representation of a molecule**
19. b. **Log P value**
20. c**. Annotation data base**

PART B (3 x 2 = 6 marks)

Answer all the questions

21. What is meant by cheminformatics.

Cheminformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.

22. Differentiate cheminformatics and Bioinformatics

Cheminformatics has mainly dealt with small molecules, whereas bioinformatics addresses genes, proteins, and other larger chemical compounds (shown in Figure below). Chem and Bioinformatics complements each other for bimolecular process, like structure and function of proteins, the binding of a ligand to its binding site, the conversion of a substrate within its enzyme receptor, and the catalysis of a biochemical reaction by an enzyme.

23. Write notes on structure descriptors.

The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.

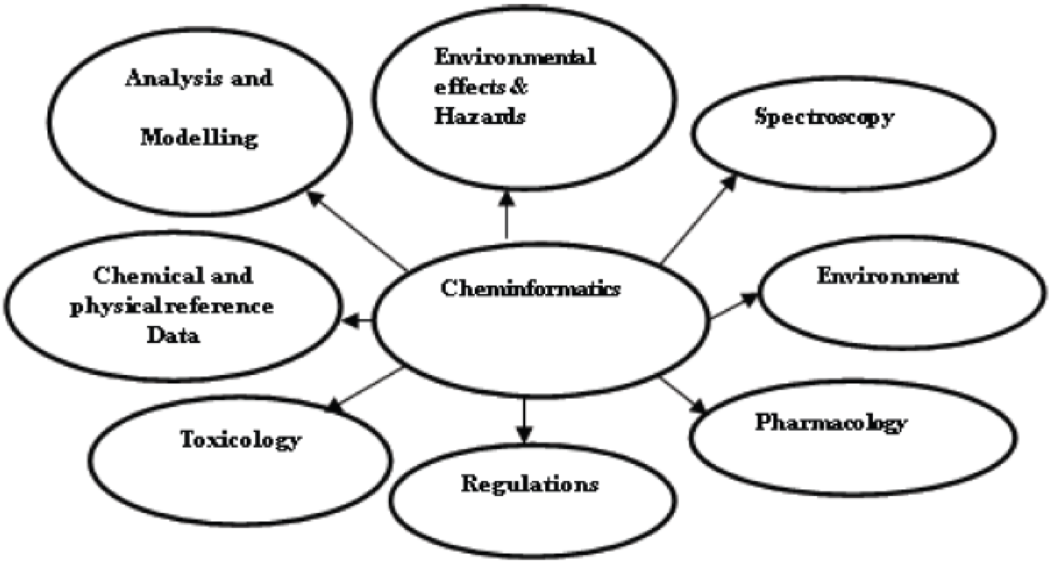
PART C (3 x 8 = 24 marks)
Answer all the questions

24.a. Explain the need and importance of cheminformatics

Need and importance of cheminformatics

Cheminformatics plays a key role to maintain and access enormous amount of chemical data, produced by chemist (more than 45 million chemical compounds are known and the number may increase in million every year,) by using a proper database. Also, the field of chemistry needs a novel technique for knowledge extraction from data to model complex relationships between the structure of the chemical compound and biological activity or the influence of reaction condition on chemical reactivity. Cheminformatics has wider range of application and Fig. 3. shows influence if cheminformatics in some specific research areas.

The need for cheminformatics



Three major aspects of Cheminformatics are

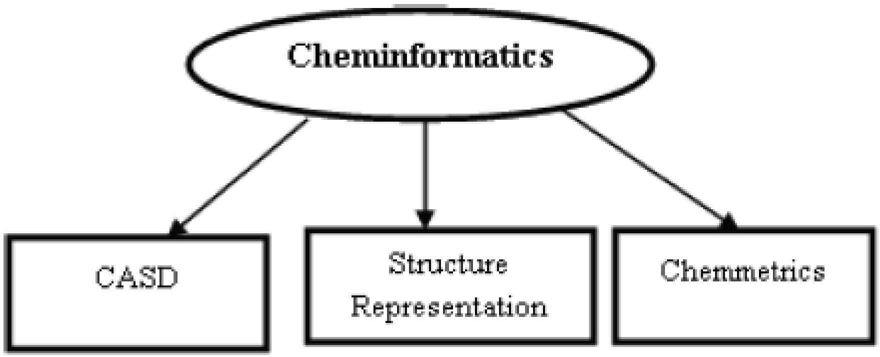
i) Information Acquisition, is a process of generating and collecting data empirically (experimentation) or from theory (molecular simulation)

ii) Information Management deals with storage and retrieval of information and

iii) Information use, which includes Data Analysis, correlation, and application to problems in the chemical and biochemical sciences.

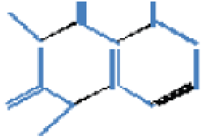
24.b. What are the applications of cheminformatics.

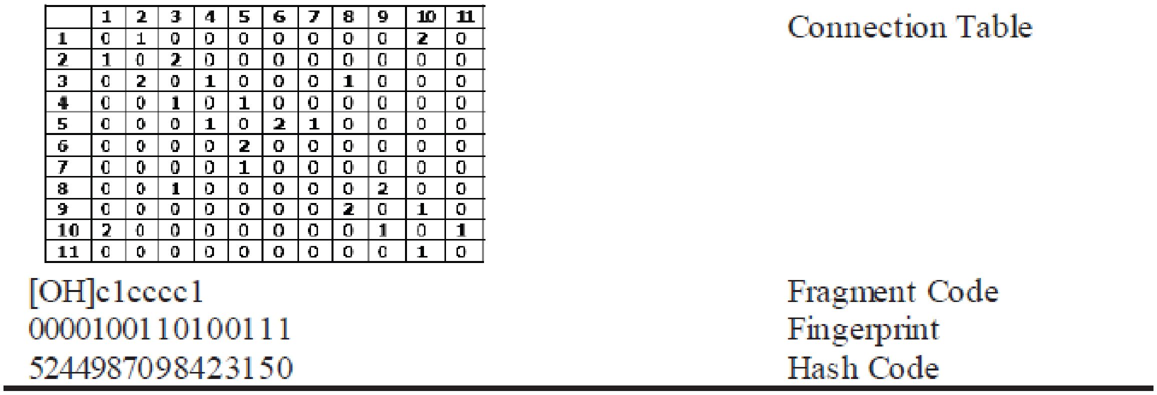
Cheminformatics is a significant application of information technology to help chemists for investigating new problems, organize, analyse, and understand scientific data in the development of novel compounds, materials and processes. Primary modules of cheminformatics are Computer-Assisted Synthesis Design, Structure representation and chemmetrics



Computer-Assisted Synthesis Design (CASD) is applied mainly where artificial intelligence technique can be applied. This technique is applied in various applications which included pharmaceutical, food industry, textile industry and agro industry. Various forms of machine readable chemical representation play basic property to design chemical database where the chemical information are stored for analysis and manipulation. The chemical structure representations can be linear, 2D or in 3D format. Some of the chemical structure representations are shown in Table 1. SMILES (Simplified Molecular Input Line Entry Specification) is one of the linear chemical notation format which is widely used among chemist [38] for various clinical and analysis purpose. Structure representation deals with Reaction Representation, Structure Descriptors, Molecular Modelling, Structure Searching, and Computer-Assisted Structure Elucidation (CASE).

Table 1: Some of the Chemical Structure Representation

Representation	Name
Caffine	Common Name
trimethylxanthine coffeine, theine, mateine,	Synonyms
$C_8H_{10}N_4O_2$	Empirical formula
3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione	IUPAC Name
58-08-2	CAS Registry Number
T56 BN DN FNVNVJ B1 F1 H1	WLN Notation
CN1C=NC2=C1C(=O)N(C(=O)N2C)C	SMILES
1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	Inchl
	Markush Structure



	1	2	3	4	5	6	7	8	9	10	11
1	0	1	0	0	0	0	0	0	0	2	0
2	1	0	2	0	0	0	0	0	0	0	0
3	0	2	0	1	0	0	0	1	0	0	0
4	0	0	1	0	1	0	0	0	0	0	0
5	0	0	0	1	0	2	1	0	0	0	0
6	0	0	0	0	2	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	2	0	0
9	0	0	0	0	0	0	0	2	0	1	0
10	2	0	0	0	0	0	0	0	1	0	1
11	0	0	0	0	0	0	0	0	0	1	0

Connection Table

[OH]c1cccc1	Fragment Code
0000100110100111	Fingerprint
5244987098423150	Hash Code

25.a. Explain the tools used for cheminformatics

Tools Used for cheminformatics

The development of software and tools for computer assisted organic synthesis are under vast development. This has resulted in many tools and representations for chemical structures. Some of the tools are listed below.

ISIS-Draw is a chemical structure drawing program for Windows, published by MDL Information Systems. It is the interfacial software to ISIS/Base database.

ChemDraw is a molecule editor developed by the cheminformatics company Cambridge Soft. ChemDraw is, along with Chem3D and ChemFinder, part of the ChemOffice suite of programs and is available for Macintosh and Microsoft Windows.

ChemWindow, is a chemical structure drawing program with several template. The template can be created by the customer can be saved in template folder and opened in preference dialogue box .

ChemSketch, is a chemical structure drawing program with predefined templates are available for drawing and it is more powerful and user friendly tool for structure analysis

ChemReader is a software developer toolkit for translating digital raster images of chemical structures into standard, chemical file formats that can be searched and analyzed with other open source or commercial cheminformatic software.

JME Molecular Editor is a Java applet which allows to draw / edit molecules and reactions (including generation of substructure queries) and to depict molecules directly within an HTML page

LogCHEM, an Inductive Logic Programming (ILP) based tool for discriminative interactive mining of chemical fragments.

PLSR (PLS-Regression), a simple chemmetrics tool, which relates two matrix X and Y through linear multivariate model and has the ability to analyse data with many, noisy, collinear, and even incomplete variables in both X and Y.

Wendi (Web Engine for Nonobvious Drug Information), a web based integrative data mining tool. It attempts to find non-obvious relationships between a query compound and scholarly publications, biological properties, genes and diseases using multiple information sources.

ChemMine tool is an online service for small molecule analysis. It provides an interface between cheminformatics and data mining tools for various analytical analyses in chemical genomics and drug discovery.

CML (Chemical Markup Language) is degined as combination of semantic text and nontextual information of chemical strucutre on the internet. It acts like HTML pages.

MyChemise (My Chemical Structure Editor) is a new 2D structure editor. It is designed as a Java applet that enables the direct creation of structures in the Internet using a web browser. MyChemise saves files in a digital format (.cse) and the import and export of .mol files using the appropriate connection tables is also possible.

PubChem is an open repository for small molecules and their experimental biological activity. It integrates and provides search, retrieval, visualization, analysis, and

Open Babel is a chemical tool box which interconverts chemical structures between different formats, over 110 formats.

AmberTools is used Biomolecular simulation and analysis of polymers, nucleotides, and synthetic organic structures.

Some other tools such as, CAS Draw, DIVA (Diverse Information, Visualization and Analysis), Structure Checker Accord, DS Accord Chemistry Cartridge, MarvinSketch PowerMV, TINKER, APBS, ArgusLab, Babel, ioSolveIT, ChemTK, Chimera, CLIFF, Dragon, gOpenMol, Grace, JOELib, Jmol, IA_LOGP, Lammps, MIPSIM, Mol2Mol, AMSOL, MOLCAS, Molexel, ICM-Pro, ORTEP, Packmol, Polar, XLOGP,PREMIER Biosoft, Q-chem, ALOGPS, Qmol, SageMD, ChemTK Lite, Transient, CLOGP,TURBOMOLE, UNIVIS, VMD, WHATIF, GCluto, COSMOlogic, KOWWIN are also used for similar kind of applications mentioned above.

25.b. Discuss the role of cheminformatics in Modern drug discovery.

The range of applications of cheminformatics is rich indeed; any field of chemistry can profit from its methods. The following lists different areas of chemistry and indicates some typical applications of cheminformatics.

a) Storing data generated through experiments or from molecular simulation Retrieval of chemical Structures from chemical database (Software libraries).

b) Prediction of physical, chemical and biological properties of chemical compounds.

c) Elucidation of the structure of a compound based on spectroscopic data.

d) Structure, Substructure, Similarity and diversity searching from chemical database

e) High Throughput Screening (HTS) is the integration of technologies (laborat ory automation, assay technology, micro plate based instrumentation, etc.) to quickly screen chemical compounds in search of a desired activity.

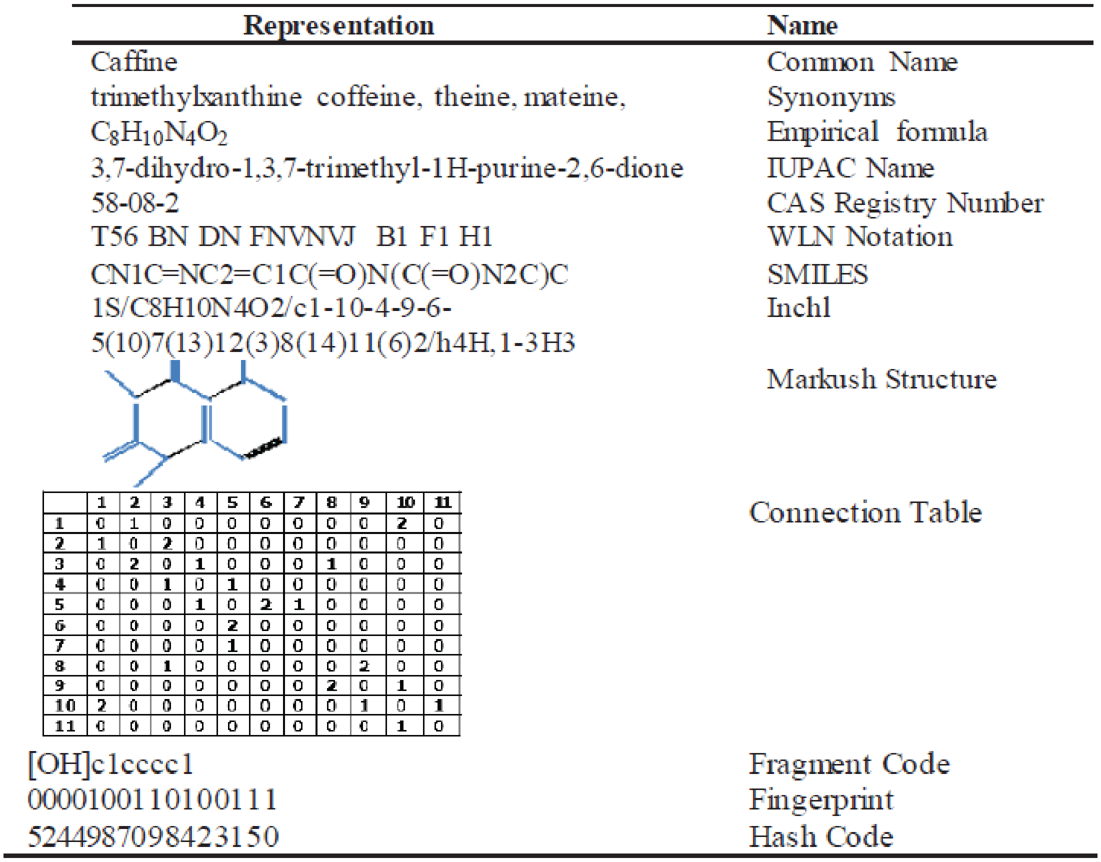
e) Docking - Interaction between two macromolecules.

f) Drug Discovery.

h) Molecular Science, Materials Science, Food Science (nutraceuticals), Atmospheric chemistry, Polymer chemistry, Textile Industry, Combinatorial organic synthesis (COS)

26.a. Write and explain some of the chemical structure representation of Caffine

Table 1: Some of the Chemical Structure Representation

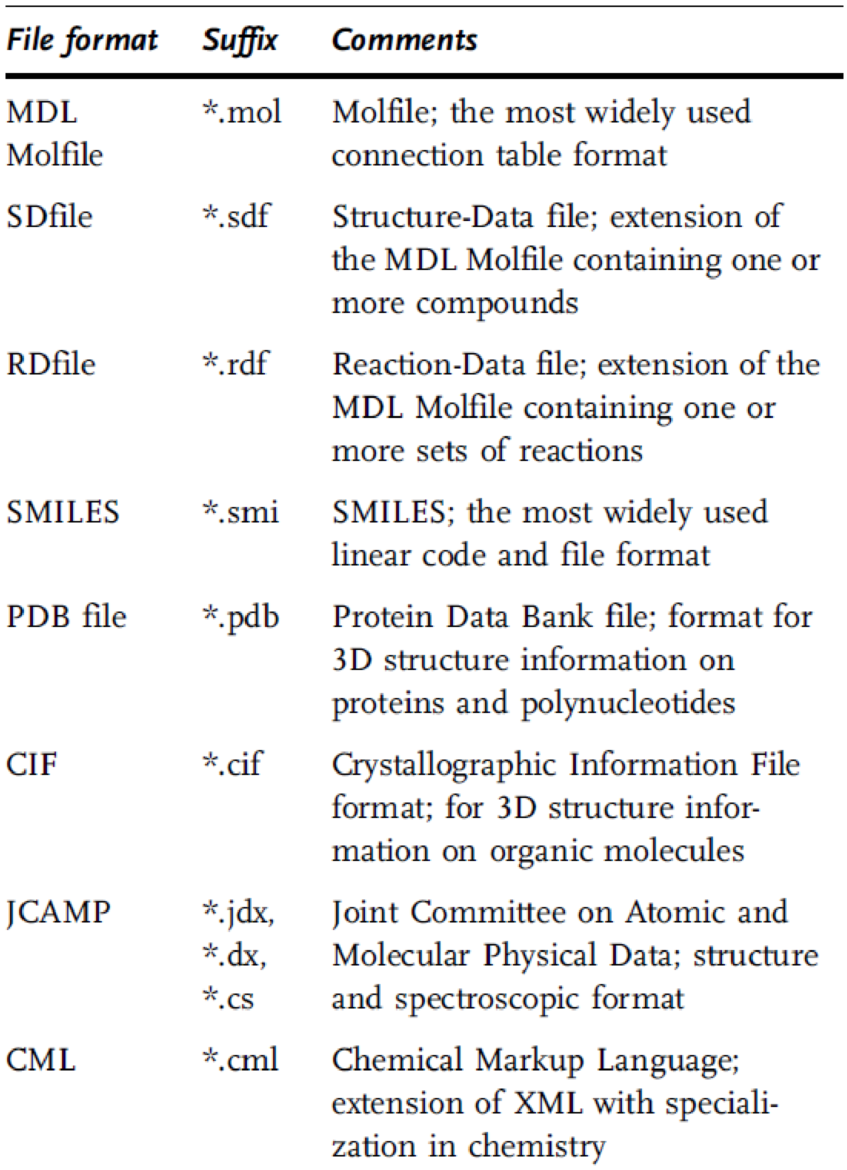


<table>
<tr><th>Representation</th><th>Name</th></tr>
<tr><td>Caffine</td><td>Common Name</td></tr>
<tr><td>trimethylxanthine coffeine, theine, mateine,</td><td>Synonyms</td></tr>
<tr><td>$C_8H_{10}N_4O_2$</td><td>Empirical formula</td></tr>
<tr><td>3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione</td><td>IUPAC Name</td></tr>
<tr><td>58-08-2</td><td>CAS Registry Number</td></tr>
<tr><td>T56 BN DN FNVNVJ B1 F1 H1</td><td>WLN Notation</td></tr>
<tr><td>CN1C=NC2=C1C(=O)N(C(=O)N2C)C</td><td>SMILES</td></tr>
<tr><td>1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3</td><td>Inchl</td></tr>
<tr><td></td><td>Markush Structure</td></tr>
<tr><td><table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th></tr>
<tr><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>0</td></tr>
<tr><td>2</td><td>1</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3</td><td>0</td><td>2</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>4</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>5</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>2</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>6</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>7</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>8</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td></tr>
<tr><td>9</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>10</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td></tr>
<tr><td>11</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
</table></td><td>Connection Table</td></tr>
<tr><td>[OH]c1cccc1</td><td>Fragment Code</td></tr>
<tr><td>0000100110100111</td><td>Fingerprint</td></tr>
<tr><td>5244987098423150</td><td>Hash Code</td></tr>
</table>

26.b. Explain the structure of MOL files and SD files.

Structure of MOL files

In chemistry, numerous software programs are available to handle structure information on molecules. The one task in common is to save data in a file. Many organizations have developed their own connection table format and a quite a fe have made provisions for the import or export of other file formats. 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 format of one software system has to be understood by another i.e. converted into its internal file format. MDL Molfile format developed at Molecular Design Limited became a standard file format. Several extensions have been made to the MDL Molfile format, leading to the SDfile, RGfile, Rxnfile or RDfile, with each one having special additional information on one or several molecules.

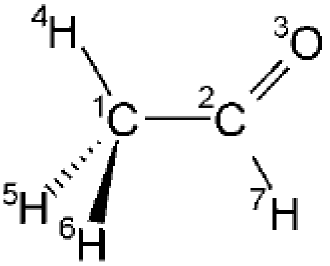


File format	*Suffix*	*Comments*
MDL Molfile	*.mol	Molfile; the most widely used connection table format
SDfile	*.sdf	Structure-Data file; extension of the MDL Molfile containing one or more compounds
RDfile	*.rdf	Reaction-Data file; extension of the MDL Molfile containing one or more sets of reactions
SMILES	*.smi	SMILES; the most widely used linear code and file format
PDB file	*.pdb	Protein Data Bank file; format for 3D structure information on proteins and polynucleotides
CIF	*.cif	Crystallographic Information File format; for 3D structure information on organic molecules
JCAMP	*.jdx, *.dx, *.cs	Joint Committee on Atomic and Molecular Physical Data; structure and spectroscopic format
CML	*.cml	Chemical Markup Language; extension of XML with specialization in chemistry

A Molfile describes a single molecular structure which can contain disjointed fragments. An SDfile (SD stands for structure-data) contains structure and data (properties) for any number of molecules, which makes it especially convenient for handling large sets of molecules. All the MDL file formats are referred together as CTfiles.

The structure of the MOLfile

The structure of ethanol molecule and the corresponding MOLfile is given below



1.	NSC7594 acetaldehyde	Header block	
2.	JTtclserve09180215543D 0 0.00000 0.00000NCI NS		
3.			
4.	7 6 0 0 0 0 0 0 0 0999 V2000	Counts line	Connection table (Ctab)
5.	0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0	Atom block	
6.	1.5000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0		
7.	2.1200 -1.0200 -0.0200 O 0 0 0 0 0 0 0 0 0 0 0 0		
8.	-0.3567 -0.4872 -0.8834 H 0 0 0 0 0 0 0 0 0 0 0 0		
9.	-0.3567 -0.5215 0.8636 H 0 0 0 0 0 0 0 0 0 0 0 0		
10.	-0.3567 1.0086 0.0198 H 0 0 0 0 0 0 0 0 0 0 0 0		
11.	2.0245 0.9324 0.0183 H 0 0 0 0 0 0 0 0 0 0 0 0		
12.	1 2 1 0 0 0 0	Bond block	
13.	2 3 2 0 0 0 0		
14.	1 4 1 0 0 0 0		
15.	1 5 1 0 0 0 0		
16.	1 6 1 0 0 0 0		
17.	2 7 1 0 0 0 0		
18.	M END	Properties block	

[16CHU501A] **Reg. No------------------**

Karpagam Academy of Higher Education
(Deemed to be University)
(Established Under Section 3 of UGC Act, 1956)
Coimbatore-21
(For the candidates admitted from 2016 onwards)
III B.Sc. Chemistry
Cheminformatics
Internal Test-II

Date: Maximum: 50 marks
Time: 2hrs

PART A (20 x 1 = 20 Marks)
Answer all the questions

1. In the Nomenclature of Inorganic compounds The first named element
   a. Oxygen b. Carbon c. **Electropositive element** d. Electronegative element

2. IUPAC name of phenylalanine is
   a. 3-amino-3-phenylpropanoic acid b. 3-amino-2-phenylpropanoic acid
   c. 2-amino-3-phenylpropanoic acid d. phenylpropanoic acid

3. Atoms are represented by their atomic symbols
   a. WLN notation b. **SMILES notation** c. ROSDAL notation d. SLN notation

4. The elements of a matrix contain values which specify the shortest distance between the atoms involved.
   a. adjacency matrix b. **distance matrix** c. incidence matrix d. bond matrix

5. Atom-by-atom searching is pertained to
   a. **backtracking algorithm** b. Substructure Search c. Full structure search
   d. Similiarity search

6. The structures are highly degenerate during the Full structure search is made using
   a. **empirical formula approach** b. molecular weights c. trival names
   d. various line notations

7. In full structure search Consisting of character strings, these representations are compact and easy to use
   a. **SMILES** b. ROSDAL c. SLN d. WLN

8. The basic strategy for the improvement of the performance of substructure search algorithms.
   a. **Optimisation of the hardware and software technologies used**
   b. By using WLN notation c. By using MOLfiles d. By using SDfiles

9. In the nomenclature of inorganic compounds the charges of ions are placed at

a. the top right-hand side next to the element symbol

b. the lower right hand side by index numbers.

c. the lower left hand side by index numbers.

d. As a superscript

10. Carbon atoms are shown in the notation only by digits in

a. WLN notation **b. ROSDAL Notation** c. SMILES notation d. SLN Notation

11. The WLN code for phenyl alanine is

a. **VQYZ1R** b. 1O-2=3O, 2-4-5N, 4-6-7=-12-7 c. NC(Cc1ccccc1)C(O)=O

d. C[1]H:CH:CH:CH:CH:C(:@1) CH2CH(NH2)C(=O)OH

12. The matrix is an n x m matrix where the nodes (atoms) define the columns (n) and the edges (bonds) correspond to the rows (m).

a. adjacency matrix b. distance matrix c. **incidence matrix** d. bond matrix

13. The simplest and more suitable way of storing chemical information is by using the character strings

a. empirical formula approach b. molecular weights c. trival names **d. Trade Names**

14. Hash codes are used as a charactering string for

a. Substructure Search b. **Full structure search** c. Similiarity search

d. Back tracking algorithm

15. To evaluate the performance of the descriptors one needs a database of compounds which the

a**. biological activities are known** b. Functional groups are known

c. Molecular formula are known d. Crystal structures are known

16. For similarity measure selection Queries are selected from therapeutic categories that

a. **Are fairly specific** b. Contain a small number of actives for reasonable statistics

c. Only one chemical class present in them d. Are fairly diversified

17. Matrices in which all elements are shown twice are called

a. **Redundant matrix** b. Non-redudant matrix

c. bond matrix d. Incidence matrix

18. SDfile describes

a. a single molecular structure which can contain disjointed fragments

b. **contains structure and data (properties) for any number of molecules**

c. A matrix contains each element only once

d. Matrices in which all elements are shown twice

19. The third strategy for optimization of substructure searching is done by a process
    a. **Screening** b. reduce the number of candidates atoms for mapping from GQ ---GT.
    c. partitioning d. coagulation

20. The similarity of reactions can be defined by
    a. **physicochemical parameters of the atoms,** b. Molecular weight
    c. Molecular formula d. Pattern Matching

PART B (3 x 2 = 6 marks)
Answer all the questions

21. What is meant by a line notation
22. What is meant by matrix representation
23. What are the different ways available for searching of a chemical structure.

PART C (3 x 8 = 24 marks)
Answer all the questions

24.a. Explain the rules to be followed in the nomenclature of inorganic and organic compounds

OR

24.b. Explain the structure of a MOLfile in detail.

25.a. Explain in detail the structure of a SDfile

OR

25.b. What are the different steps to be followed in similarity search approach.

26.a. Describe the substructure search in cheminformatics

OR

26.b. Explain the three dimensional structure search methods.

[16CHU501A] **Reg. No------------------**

Karpagam Academy of Higher Education
(Deemed to be University)
(Established Under Section 3 of UGC Act, 1956)
Coimbatore-21
(For the candidates admitted from 2016 onwards)
III B.Sc. Chemistry
Cheminformatics
Internal Test-II- ANSWER KEY

PART A (20 x 1 = 20 Marks)

Answer all the questions

1. c. Electropositive element
2. c. 2-amino-3-phenylpropanoic acid
3. b. SMILES notation
4. b. distance matrix
5. a. backtracking algorithm
6. a. empirical formula approach
7. a. SMILES
8. a. Optimisation of the hardware and software technologies used
9. a. the top right-hand side next to the element symbol
10. b. ROSDAL Notation
11. a. VQYZ1R
12. c. incidence matrix
13. d. Trade Names
14. b. Full structure search
15. a. biological activities are known
16. a. Are fairly specific
17. a. Redundant matrix
18. b. contains structure and data (properties) for any number of molecules
19. a. Screening
20. a. physicochemical parameters of the atoms

PART B (3 x 2 = 6 marks)

Answer all the questions

21. What is meant by a line notation

It represents the structure of chemical compounds as a linear sequence of letters and numbers. IUPAC nomenclature represents a kind of line notation. However it makes difficult to obtain additional information on the structure of a compound directly from its name.

A chemist trained in this line notation, could enter the code of large molecules faster than with a structure-editing program.

22. What is meant by matrix representation

The matrix of a structure with ‘n’ atoms consists of an array of n x n entries. A molecule with its different atoms and bond types can be represented in matrix form in different ways depending on what kind of entries are chosen for the atoms and bonds. Thus a variety of matrices has been proposed; adjacency, distance, incidence, bond, and bond-electron matrices.

23. What are the different ways available for searching of a chemical structure.

Full structure search

Sub structure search

Backtracking Algorithm
Similarity searching
Three Dimensional Structure Search

PART C (3 x 8 = 24 marks)
Answer all the questions

24.a. Explain the rules to be followed in the nomenclature of inorganic and organic compounds

Nomenclature of Inorganic compounds

1. Electropositive elements are listed first.

2. The stoichiometry of the elements is indicated at the lower right hand side by index numbers.
3. The charges of ions are placed at the top right-hand side next to the element symbol (eg S^{2-}).
4. In ions of complexes, the central atom is specified before the ligands are listed in alphabetical order, the complex ion is set in square brackets.

Nomenclature of organic compounds

1. The elements of an organic compound are listed in empirical formulas and the stoichiometry is indicated by index numbers.
2. Carbon and hydrogen atoms were positioned in the first and second places respectively and the hetero atoms following them in alphabetical order (eg. $C_9H_{11}NO_2$).
3. However since different compounds have the same empirical formulae, formulas were developed that indicate the presence of certain structural units and functional groups.

Eg. the empirical formulae of phenylalanine may be split into a more extended form that shows the presence of a phenyl ring, as well as an amino acid and a carboxylic acid group.

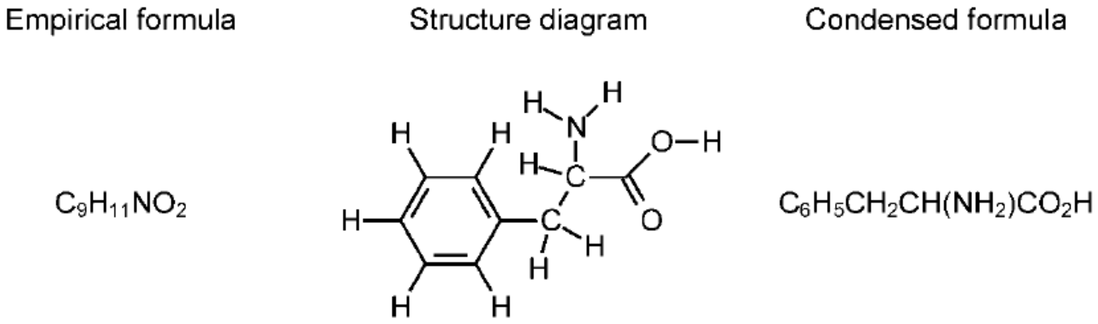


Fig: Different representations of phenylalanine.

Systematic IUPAC nomenclature of compounds

1. Trival names are short and simple to memorise.

2. IUPAC name can be quite long and cumbersome. The aim is to describe particular parts of the structure (fragments) in a systematic manner, with special expressions from a vocabulary of terms.

3. So the systematic nomenclature is used in Chemical Abstracts Service as index for chemical structures.

4. However this does not directly allow the extraction of additional information about the molecule, such as bond orders or molecular weight.

5. There are two basic rules for the nomenclature of organic compounds.
   a. No. of carbon atoms in the longest continuous aliphatic chain of carbon atoms has to be indicated. Branching of the skeleton, and the presence of rings, have to be specified by prefixes.
   b. Functional groups are to be specified by a prefix and /or suffix to indicate the family to which the compounds belongs.
   c. Substituents are listed in the name in alphabetical order.

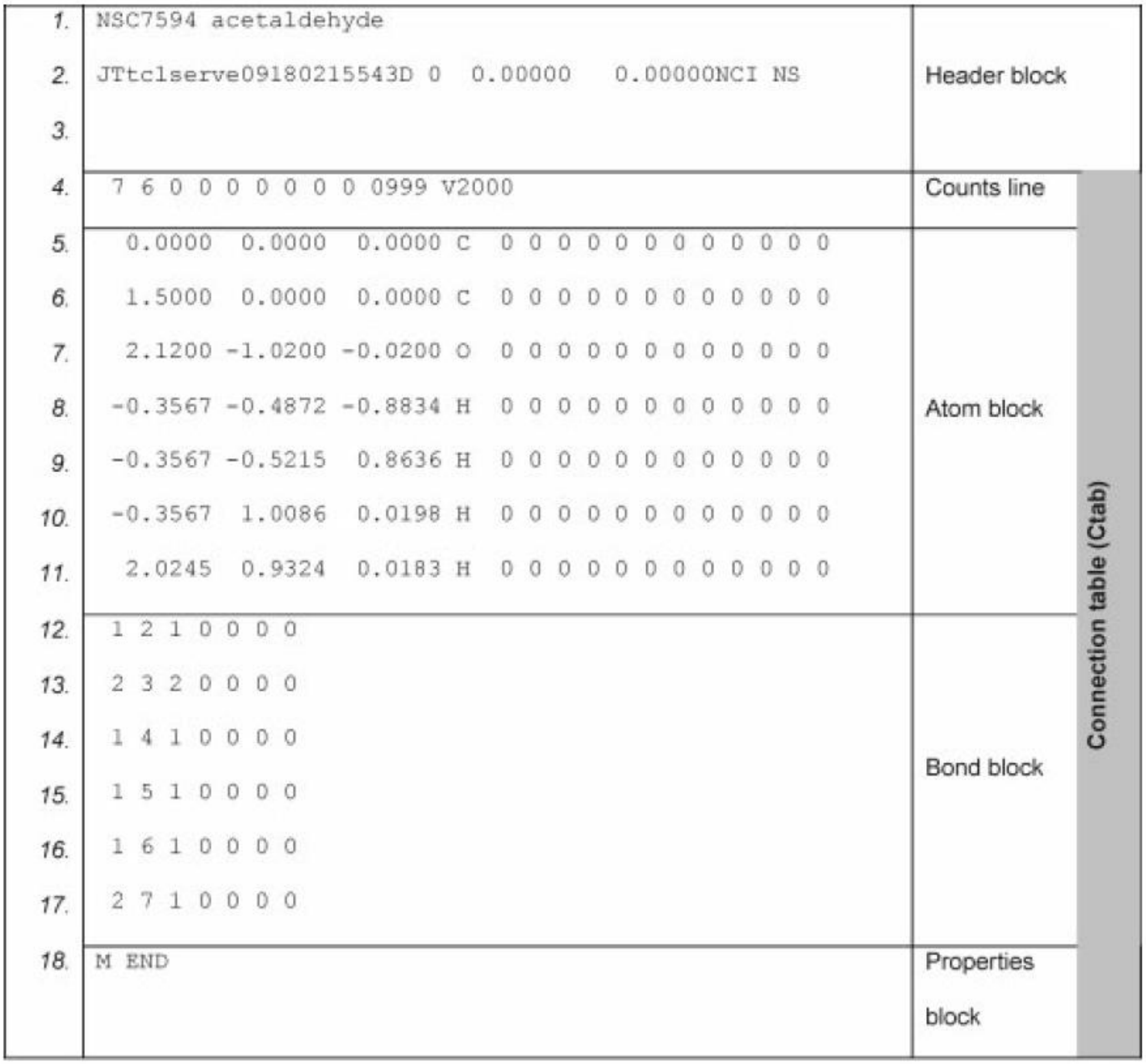
   Eg. The IUPAC name of phenylalanine is 2-amino-3-phenylprppanoic acid. It indicates a carbon chain of length three (propan-) of the acid (propanoic acid) with an additional two structural units, the phenyl and the amino group at different positions on the carbon chain: phenyl at carbon atom number 3 and amino at 2, with the counting beginning with the carbon atom of the acid group (COOH).

Neither a trival name not the IUPAC name, which both represent the structure as an alphanumerical (text) string, is ideal for computer processing. The reason is that various valid compound names can describe one chemical structure. As a consequence, the name/structure correlation is unambiguous but not unique. Nowadays programmes can translate names of structures and structures to names, to make published structures accessible in electronic journals

24.b. Explain the structure of a MOLfile in detail.

In chemistry, numerous software programs are available to handle structure information on molecules. The one task in common is to save data in a file. Many organizations have developed their own connection table format and a quite a fe have made provisions for the import or export of other file formats. 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 format of one software system has to be understood by another i.e. converted into its internal file format. MDL Molfile format developed at Molecular Design Limited became a standard file format. Several extensions have been made to the MDL Molfile format, leading to the SDfile, RGfile, Rxnfile or RDfile, with each one having special additional information on one or several molecules.

A Molfile describes a single molecular structure which can contain disjointed fragments. An SDfile (SD stands for structure-data) contains structure and data (properties) for any number of molecules, which makes it especially convenient for handling large sets of molecules. All the MDL file formats are referred together as CTfiles.



1.	NSC7594 acetaldehyde	Header block	
2.	JTtclserve09180215543D 0 0.00000 0.00000NCI NS		
3.			
4.	7 6 0 0 0 0 0 0 0 0999 V2000	Counts line	Connection table (Ctab)
5.	0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0	Atom block	
6.	1.5000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0		
7.	2.1200 -1.0200 -0.0200 O 0 0 0 0 0 0 0 0 0 0 0 0		
8.	-0.3567 -0.4872 -0.8834 H 0 0 0 0 0 0 0 0 0 0 0 0		
9.	-0.3567 -0.5215 0.8636 H 0 0 0 0 0 0 0 0 0 0 0 0		
10.	-0.3567 1.0086 0.0198 H 0 0 0 0 0 0 0 0 0 0 0 0		
11.	2.0245 0.9324 0.0183 H 0 0 0 0 0 0 0 0 0 0 0 0		
12.	1 2 1 0 0 0 0	Bond block	
13.	2 3 2 0 0 0 0		
14.	1 4 1 0 0 0 0		
15.	1 5 1 0 0 0 0		
16.	1 6 1 0 0 0 0		
17.	2 7 1 0 0 0 0		
18.	M END	Properties block	

Connection Table (Ctab) : (lines 4-18)

The first line of the header block contains the molecule name-No specific format (If no name is available the line is blank)

The second line, however, has a strict format and contains general information about the users name, the programme used to generate this file, and the date and time when the file was created.

a. The date and time information is formed of concatenated two digit values representing the month (09), day (18), year (02), hour (15) and minute (54) respectively.
b. It specifies also whether 2D or 3D atomic coordinates

The third line of the Header Block may be empty or may contain comments.

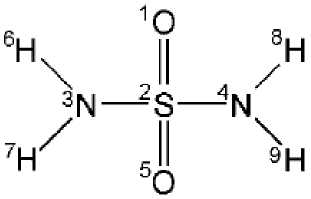
Lines 4-18 form the connection table (Ctab), containing the description of the collection of atoms constituting the given compound, which can be wholly or partially connected by bonds. Such a collection can represent molecules, molecular fragments, substructures, substituent groups, and so on.

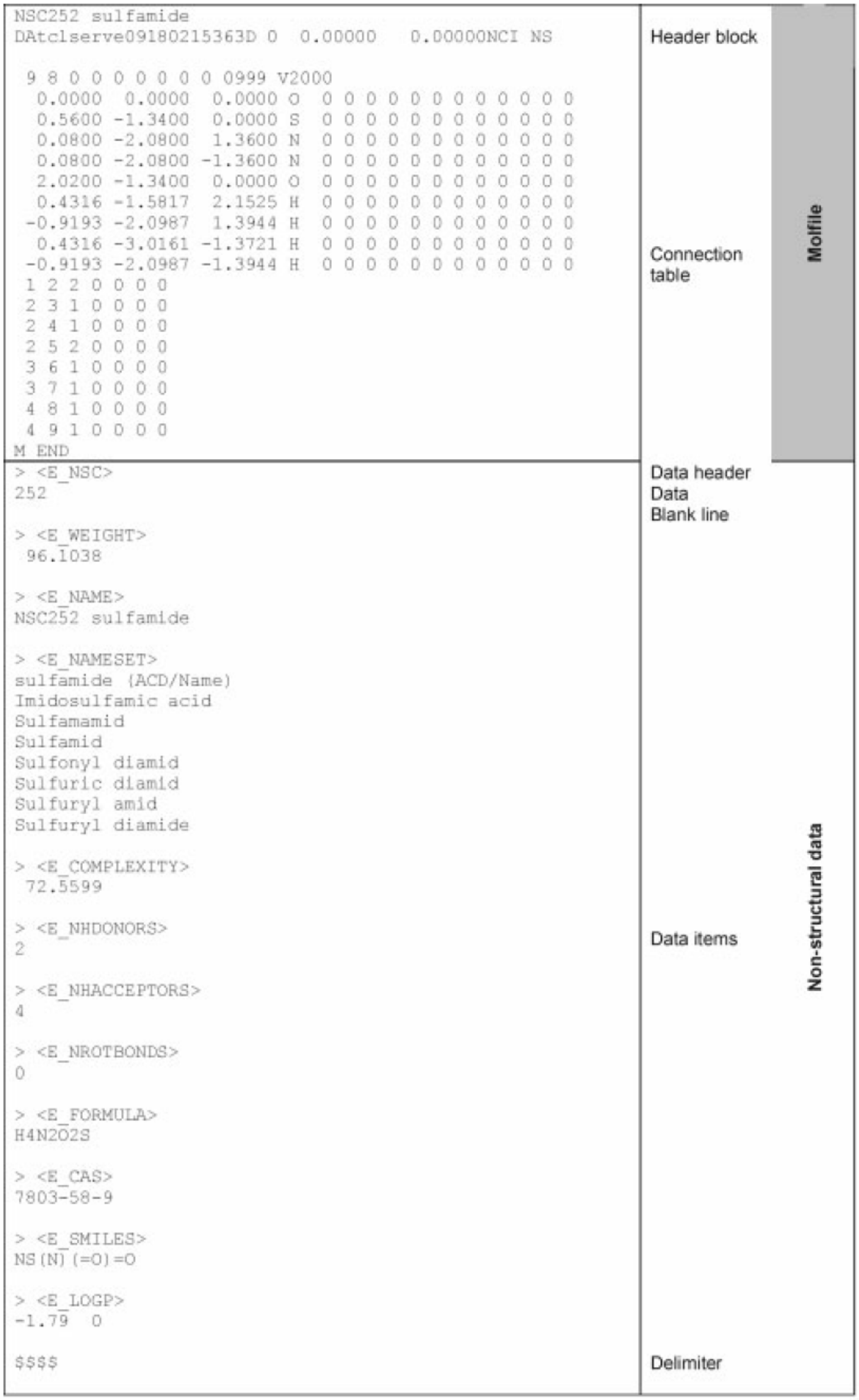
a. The first line is called the counts line, specifies how many atoms constitute the molecule represented by this file, how many bonds are within themolecule, whether this molecule is chiral (1 if it is chiral). The last but one entry is always set to 999. The last entry specifies the version of the Ctab (V2000 or V3000).

25.a. Explain in detail the structure of a SDfile

Structure of a SDfile

In the SDfile, each molecule is represented by its Molfile with additional data items describing its non-structural properties (Mol.weight, heat of formation, molecular descriptors, biological activity, etc.). The information on a molecule is terminated by a delimiter line (containing only “$$$$”). Each data item starts with a data header line, which reflects a molecular property name. Next one or more rows contain the actual data; they are terminated by an empty line.





```
NSC252 sulfamide
DAtclserve09180215363D 0   0.00000     0.00000NCI NS

  9 8 0 0 0 0 0 0 0 0999 V2000
    0.0000   0.0000   0.0000 O   0 0 0 0 0 0 0 0 0 0 0 0
    0.5600  -1.3400   0.0000 S   0 0 0 0 0 0 0 0 0 0 0 0
    0.0800  -2.0800   1.3600 N   0 0 0 0 0 0 0 0 0 0 0 0
    0.0800  -2.0800  -1.3600 N   0 0 0 0 0 0 0 0 0 0 0 0
    2.0200  -1.3400   0.0000 O   0 0 0 0 0 0 0 0 0 0 0 0
    0.4316  -1.5817   2.1525 H   0 0 0 0 0 0 0 0 0 0 0 0
   -0.9193  -2.0987   1.3944 H   0 0 0 0 0 0 0 0 0 0 0 0
    0.4316  -3.0161  -1.3721 H   0 0 0 0 0 0 0 0 0 0 0 0
   -0.9193  -2.0987  -1.3944 H   0 0 0 0 0 0 0 0 0 0 0 0
  1 2 2 0 0 0 0
  2 3 1 0 0 0 0
  2 4 1 0 0 0 0
  2 5 2 0 0 0 0
  3 6 1 0 0 0 0
  3 7 1 0 0 0 0
  4 8 1 0 0 0 0
  4 9 1 0 0 0 0
M  END
> <E_NSC>
252

> <E_WEIGHT>
 96.1038

> <E_NAME>
NSC252 sulfamide

> <E_NAMESET>
sulfamide (ACD/Name)
Imidosulfamic acid
Sulfamamid
Sulfamid
Sulfonyl diamid
Sulfuric diamid
Sulfuryl amid
Sulfuryl diamide

> <E_COMPLEXITY>
 72.5599

> <E_NHDONORS>
2

> <E_NHACCEPTORS>
4

> <E_NROTBONDS>
0

> <E_FORMULA>
H4N2O2S

> <E_CAS>
7803-58-9

> <E_SMILES>
NS(N)(=O)=O

> <E_LOGP>
-1.79   0

$$$$
```

25.b. What are the different steps to be followed in similarity search approach.

Similarity search process

The most common objects of interest to a chemist are molecules. Some sources of drug like compounds are the MDL Drug Data Report (MDDR) a licensed database compiled from the patent literature containing about 115 000 compounds, as well as the database of the National Cancer Institute (NCI), containing about 250 000 compounds. The biological data base is formed of three files, which contain data from different types of measurements – TGI, LC50 and GI50.

An example of a fragment based search space is

"… a large set of diverse fragments together with generic definitions of how the fragments can be combined to molecules … The space contains about 17 000 fragments which can be connected to each other via 12 different link types.

Reactions an be considered as composite systems containing reactant and product molecules as well as reaction sites. The similarity of chemical structures is defined by generalized reaction types and by gross structural features. The similarity of reactions can be defined by physicochemical parameters of the atoms and bonds at the reaction site.

Descriptor selection and encoding

The atom pair ap, and topological torsion tt descriptor may be selected. Two other atomic properties are – atomic log P contribution and partial atomic charges.

Similarity Measure Selection

In general, different similarity measures yield different rankings, except when they are monotonic. Improved results are obtained by using fusion methods to combine the rankings resulting from different coefficients. Empirically, the Dice coefficient has worked better than cosine similarity in retrieving actives and is the standard choice for use with the ap and tt descriptors.

To evaluate the performance of the descriptors one needs a database of compounds which the biological activities are known. Queries are selected that are typical of drug-like molecule and from therapeutic categories that

1. Contain a large enough number of actives (eg.> 50) for reasonable statistics.
2. Have several chemical classes present in them
3. Are fairly specific

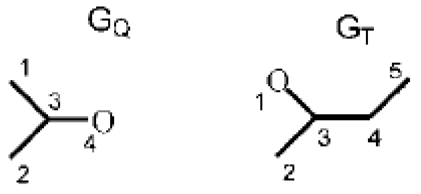
26.a. Describe the substructure search in cheminformatics

Substructure Search

A substructure search algorithm is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures such as identification of equivalent atoms, determination of maximal common substructure, ring detection, calculation of topological indices, etc.

The search for structural fragments (substructures) is very important in medicinal chemistry, QSAR, spectroscopy and many other fields in the process of pharmacophore, chromophore or other –phore perceptions.

Substructure searching is the process of identifying parts of a given structure that are equivalent to a specified query substructure. In graph-theoretical terms substructure searching is the task of checking whether the query graph (GQ) is isomorphic with a subgraph of another target graph (GT). Sometimes the target graph is called a reference graph.



There are several basic strategies for the improvement of the performance of substructure search algorithms.

1. Optimisation of the hardware and software technologies used
2. Usage of various methods to improve the perception of the substructure isomorphic with the query graph or with the rejection of inappropriate target structure candidates as early as possible.
3. Pre-processing of the most time-consuming operations that are independent of the query structure and storing them as an integral part of the data base which can be used at search time.

26.b. Explain the three dimensional structure search methods.

Chemists know that 2D representation of molecular moieties gives a very rough picture of their real-world structure. While for some practical applications this representation is sufficient for most modern investigations in all areas of molecular design, 3D structure representation and 3D structure search are highly mandatory.

The first task was the creation of large 3D chemical structure database. A subsequent step was the development of fast 3D search approaches (follows the existing 2D methods such as atomly atom mapping, maximal common substructure, 2D keys, fragments, etc., by substituting them with 3D counterparts). 3D similarity search methods are quite well developed.

3D substructure search in usually known as pharmacophore searching in QSAR. Generally speaking there are two major approaches to it: topological and chemical function queries. In all the 3D search methods the conformational flexibility creates considerable difficulties.

Structure descriptor

It is a mathematical representation of a molecule resulting from a procedure transforming the structural information encoded within a symbolic representation of a molecule.

Molecules can be represented by structure descriptors in a hierarchial manner with respect to

a. Descriptor data type

b. Molecular representation of the compound

The information content of a structure descriptor on two major factors

a. Molecular representation of the compound
b. The algorithm which is used for the calculation of the descriptor.

Reg. No. : ------------------- **[16CHU501A]**

KARPAGAM ACADEMY OF HIGHER EDUCATION
COIMBATORE-21
(For the candidates admitted from 2016 & onwards)
B.Sc. DEGREE EXAMINATION, SEPTEMBER 2018
Fifth Semester
Chemistry
INTERNAL TEST - III
CHEMINFORMATICS

Time: 2 Hours **Maximum:50 marks**

PART- A

Answer All the Questions **(20 x 1 = 20 Marks)**

1. The process in which the atoms in GQ and GT are separated into different classes
   a. partitioning b. Union c. Intersection d. Adding

2. To establish a quantitative relationship between the structural features of a compound and its properties
   a. QSAR b. Linear free energy relationship c. SAR d. Molecular docking

3. In QPSR Descriptors have to be found representing the structural features which are related to the target property. It is related to
   a. Structure representation b. Description analysis c. Model building
   d. Linear Free Energy Relationship

4. The descriptors used for structure representation can be derived from
   a. On the basis of the connectivity b. On the basisPartition coefficient
   c. On the basis of Reaction rate constant d. On the basis of Binding constant

5. In QPSR, Model building consists of
   a. training, evaluation and testing steps b. Only evaluation and testing steps
   c. Only training and evaluation steps d. Only testing step

6. Procedures were developed for Base-catalysed and nucleophilic reactions
   a. CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
   b. EROS (Elaboration of reactions for Organic Synthesis)
   c. LHASA d. CONGEN

7. CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions) can be used to measure
   a. Physicochemical property PKa values b. polarizability effects.
   c. Inductive effects d. Resoncnce effects.

8. A molecule binding to a biological macromolecule

a. an inhibitor b. Ligand c. A substrate d. An enzyme

9. In drug discovery process to identify target molecule which can be either gene or protein and could be a potential drug for the disease

a. Target identification b. Lead identification
c. Lead optimisation d. Preclinical training

10. Linear free energy relationship is used to determine

a. binding constants b. Melting point c. Boiling point d. Specific gravity

11. Descriptors used in QPSR for structure representation may range from

a. functional group count b. Molecule c. Cluster of molecules
d. Nano material

12. Completely independent descriptors are said to be orthogonal.

a. have a correlation coefficient of 0.0 b. have a correlation coefficient of 1.0
c. have a correlation coefficient of 10.0 d. have a correlation coefficient of 100.0

13. The spectra have been predicted using quantum chemistry calculations, data-base searches, additive methods, regressions and neural networks.

a. ^{1}H-NMR spectra b. ^{13}C-NMR spectra c. IR spectra d. Mass spectra

14. The synthetic design system Works in retrosynthetic manner

a. ALPHOS b. SYNGEN c. SYNCHEM d. LHASA

15. The starting material of an enzymatic reaction

a. an inhibitor b. A substrate c. An enzyme d. A ligand

16. A receptor ligand preventing the action of an agonist in a direct (competitive or indirect (allosteric) manner.

a. Receptor b. agonist c. antagonist d. Lead molecule

17. In drug discovery process where the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds.

a. Target identification b. Lead identification c. Lead optimization
d. Preclinical trial

18. The similarity of reactions can be defined by

a. physicochemical parameters of the atoms b. Molecular weight
c. Molecular formula d. Pattern Matching

19. The basic strategy for the improvement of the performance of substructure search algorithms.

a. Optimisation of the hardware and software technologies used
b. By using WLN notation c. By using MOLfiles d. By using SDfiles

20. The integration of technolopgies to quickly screen chemical compounds in search of a

desired activity

a. HTS b. Docking c. Drug discovery d. QSAR

PART- B

Answer All the Questions **(3 X 2= 6 Marks)**

21. What is meant by the Linear Free Energy Relatonship for the prediction of properties of compounds.

22. What is meant by a ligand in modern drug discovery

23. What is meant by a lead structure in modern drug discovery.

PART- C

Answer All the Questions **(3 X 8= 24 Marks)**

24.a. Discuss in detail about the Quantitative structure-property relationship (QSPR) for the prediction of properties.

OR

24.b. What are the characteristic features of an ^{1}H-NMR spectra.

25.a. Explain the problems faced in dealing with chemical reactions in a computer assisted synthetic design.

OR

25.b. Describe the Ligand and structure based drug design

26.a. Explain any four process involved in the modern drug discovery process

26.b. Discuss in detail about the prediction of toxicity of compounds in cheminformatics.

Reg. No. : ------------------- **[16CHU501A]**

KARPAGAM ACADEMY OF HIGHER EDUCATION
COIMBATORE-21
(For the candidates admitted from 2016 & onwards)
B.Sc. DEGREE EXAMINATION, SEPTEMBER 2018
Fifth Semester
Chemistry
INTERNAL TEST - III
CHEMINFORMATICS

PART- A

Answer All the Questions **(20 x 1 = 20 Marks)**

1. a. partitioning
2. b. Linear free energy relationship
3. a. Structure representation
4. a. On the basis of the connectivity
5. a. training, evaluation and testing steps
6. a. CAMEO (Computer Assisted Mechanistic Evaluation of Organic reactions)
7. a. Physicochemical property PKa values
8. b. Ligand
9. a. Target identification
10. a. binding constants
11. a. functional group count
12. a. have a correlation coefficient of 0.0
13. a. ^{1}H-NMR spectra
14. d. LHASA
15. b. A substrate
16. c. antagonist
17. b. Lead identification
18. a. physicochemical parameters of the atoms
19. a. Optimisation of the hardware and software technologies used
20. a. HTS

PART- B

Answer All the Questions **(3 X 2= 6 Marks)**

21. What is meant by the Linear Free Energy Relatonship for the prediction of properties of compounds.

LFER methods are widely used for the properties like partition coefficients, binding constants, or reaction rate constants. This is based on the pioneering work of Hammett for the prediction of chemical reactivity. The basic assumption is that the influence of a structural feature on the free energy change of a chemical process is constant for a congeneric series of compounds. A property φ that is linearly dependant on a free energy change can then be calculated by the property of the basic element of this series, the so-called parent element, and the constant φx for the structural feature X.

22. What is meant by a ligand in modern drug discovery

A ligand is a molecule binding to a biological macromolecule. The drug molecules which are used for docking studies in order to determine the binding interactions with a protein are called ligands.

23. What is meant by a lead structure in modern drug discovery.

A **lead compound** (i.e. a "leading" compound, not to be confused with various compounds of the metallic element lead) in drug discovery is a chemical compound that has pharmacological orbiological activity likely to be therapeutically useful, but may nevertheless have suboptimal structure that requires modification to fit better to the target; lead drugs offer the prospect of being followed by back-up compounds.

A representative of a compound series with sufficient potential (as measured by potency, selectivity, pharmacokinetics, pysicochemical properties, absence of toxicity and novelty) to progress to a full drug development program.

PART C

Answer All the Questions (3 X 8= 24 Marks)

24.a. Discuss in detail about the Quantitative structure-property relationship (QSPR) for the prediction of properties.

The general procedure in a QSPR approach consists of three steps

a. Structure representation
b. Descriptor analysis
c. Model building

Structure representation

Descriptors have to be found representing the structural features which are related to the target property. This is the most important step in QSPR, and the development of powerful descriptors is of central interest in this field. Descriptors can range from simple atom or functional group counts to quantum chemical descriptors. They can be derived on the basis of the connectivity (2D descriptors), the 3D structure, or the molecular surface (3D descriptors) of a structure. Which kind of descriptors should or can be used is primarily dependent on the size of the data set to be studied and the required accuracy.

Descriptor Analysis

In general, the set of calculated descriptors should not be used directly for the model building process, mainly because of three problems

1. Different elements of the descriptor set may intercorrelate ie. Different descriptors basically encode the same structural aspect.
2. Descriptors may encode features that do not contribute to the property at all
3. The overall size of a descriptor set may become unmanageable large.

Each case requires a pre-processing of the descriptor set such that the essential information is extracted into a reduced descriptor set with higher information density related to the target property. Mainly, two statistical measures are used to judge the quality of the descriptors: the variance and the correlation coefficient among the descriptors. The former is a measure of the variation of a descriptor across a data set. A low variance indicates little information content of a descriptor. The latter is a measure of internal redundancy. Completely independent descriptors have a correlation coefficient of 0.0 and are said to be orthogonal. This ideal case is of course hardly ever found, and the correlation of two descriptors should normally be not greater than 0.6, but reports of acceptable correlation coefficients between descriptors have ranged from less than 0.4 to 0.9 in the literature. The descriptor set can then be reduced by eliminating candidates that show such bad characteristics.

Model Building

The model building step deals with the development of mathematical models to relate the optimized set of descriptors with the target property. Two statistical measures indicate the quality of a model, the regression coefficient, r, or its square r^2, and the standard, σ.

Model building consists of three steps; training, evaluation and testing. In the ideal case the whole training data set is divided into three portions, the training, the evaluation set, and the test set. A wide variety of statistical or neural network methods can be used to derive QSPR and

QSAR models. The most frequently used methods are Multiple Linear Regression Analysis (MLRA) and feed forward neural networks with back propagation of errors. Once a model has been derived with the training set, the evaluation set can be used to test the predictive power of the resulting models, ie, to predict the target property for compounds yet unknown to the model and to optimize model parameters thereby. In many cases the data set is too small to allow its splitting. Therefore cross-validation techniques are applied for the evaluation step. In k-fold cross validation the training set is split in k subsets. Then k-1 subsets are used as a training set and one subset as test set. This procedure is repeated k times. As we have now a prediction for each compound, we can calculate cross-validated errors of the predictions. Values of k usually range from 5 to 10. If k equals N, the number of cases in the training set, the procedure is called leave-one-out cross-validation.

24.b. What are the characteristic features of an ^{1}H-NMR spectra.

Prediction of chemical shifts

Ab-initio and semi empirical calculations

When a molecule is submitted to a static magnetic field, the nuclear spin energy level splits. An oscillating magnetic field can then induce transitions between these energy levels and produce the NMR spectrum. For a nucleus in a molecule, the magnetic field is due to the applied magnetic field, but also to the magnetic field produced by the electrons and other nuclei. The NMR chemical shift of a nucleus results from the difference in energy between the nuclear spin states, which can be calculated by ab-initio quantum mechanical methods, typically by solving the Schrodinger equation with approximations. The effects of the external magnetic field on the nucleus of interest are added into the equations as a “perturbation”. It is then possible to calculate the chemical shift, which is related to the total molecular energy, the applied magnetic field and the nuclear magnetic moment.

Ab-initio calculations are particularly useful for the prediction of chemical shifts of “unusual species”. In this context “unusual species” means chemical entities that are not frequently found in the available large database of chemical shifts, e.g. charged intermediates of reactions, radicals and structures containing elements other than H,C,O, N, S, P, halogens and a few common metals.

The Gaussian Program is one of the most popular tools for ab-initio calculation of NMR chemical shifts.

Database approaches

A useful empirical method for the prediction of chemical shifts and coupling constants relies on the informationcontained in databases of structures with the corresponding NMR data. Large databases with hundred thousands of chemical shifts are commercially available and are linked to predictive systems, which basically rely on database searching. Protons are internally represented by their structural environments, usually their HOSE codes. When a query structure is submitted, a search is performed to find the protons belonging to similar substructures. These are the protons with the same HOSE codes as the protons in the query molecule. The prediction of the chemical shift is calculated as the average chemical shift of the retrieved protons.

When common substructures cannot be found for a given proton interpolations are applied to obtain a prediction; proprietary methods are often used in commercial programs.

Increment based Methods

In this second empirical approach, which has also been used for ^{13}C NMR spectra, predictions are based on tabulated chemical shifts for classes of structures, and corrected with additive contributions from neighbouring functional groups or substructures. Several tables have been compiled for different types of protons. Increment rules can be found in nearly any textbook on NMR spectroscopy. In such tables, typical chemical shifts are assigned to standard structure fragments (eg. protons in a benzene ring). Substituent’s in these blocks (eg. substituent’s in ortho, meta and para positions) are assumed to make independent additive contributions to the chemical shift. Once the tables are defined, the method is easy implement, does not require databases and is extremely fast.

25.a. Explain the problems faced in dealing with chemical reactions in a computer assisted synthetic design.

Computer Assisted Synthesis Design

How do chemists find a pathway to the synthesis of a new organic compound. They try to find suitable starting materials and powerful reactions for the synthesis of the target compound. Thus, synthesis design and chemical reactions are deeply linked, since a chemical reaction is the instrument by which chemists synthesize their compounds; synthesis design is a chemists major strategy to find the most suitable procedure for a synthesis problem.

The synthesis of each compound was considered as a specific task on its own. A suitable strategy for the synthesis of a target compound was mostly found on the basis of the intuition and experience of the acting chemists i.e. the planning of a synthesis ofa complex organic molecule was considered as an art form. No systematic approach was attempted to handle the strategic design of an organic synthesis.

In 1960, Corey introduced a general methodology for planning organic syntheses. The synthesis plan for a target molecule is developed by starting with the target structure and working backwards to available starting materials. The retrosynthetic analysis or disconnection of the target molecule in the reverse direction is performed by the systematic use of analytical rules which have been formulated by Corey.

Concepts for Computer-Assisted Organic Synthesis (CAOS)

The program systems for computer-assisted synthesis planning can be subdivided in to two groups:

1. Information-oriented and
2. logic oriented systems

In Logic-oriented approach generates reactions as bond breaking and bond making steps. These steps are often combined with mechanistic or thermodynamic considerations. In principle, logic oriented systems should not only be able to predict known reactions but should also generate novel reactions. This is both an advantage and a disadvantage. On one hand, such a system may suggest a reaction nobody has forseen and thus the system provides a new approach to the synthesis problem being considered. On the other hand, it may generate a huge number of chemically invalid reactions, which an experienced chemist would intuitively avoid. Therefore, the output of a logic oriented system has to be throughly verified by suitable evaluation techniques.

Information-oriented systems are based on a library of known retro-reactions which have been collected and evaluated by a group of chemists while coding them in electronic form. In addition, information on the scope and the expected yield under various conditions, as well as a strategic merit is usually stored. Such a reaction library is called a knowledge base. In synthesis design programes the knowledge base consists of a database of transforms. Each transform (retro-reaction) has been derived from a number of experimentally performed reactions.

25.b. Describe the Ligand and structure based drug design

Depending on the information available about the protein structure and the ligands binding to a particular target, four different cases can be distinguished in drug design.

Table: Cases in the drug discovery process depending on knowledge of the receptor and ligand structure

	Ligand unknown	Ligand known

Protein structure unknown	Combinatorial chemistry and HTS	QSAR, pharmacophore models and hypothesis, similiarity search in databases
Protein structure known	*De novo* design, receptor based 3D searching	Structure based design, docking

The lead discovery process is as follows.

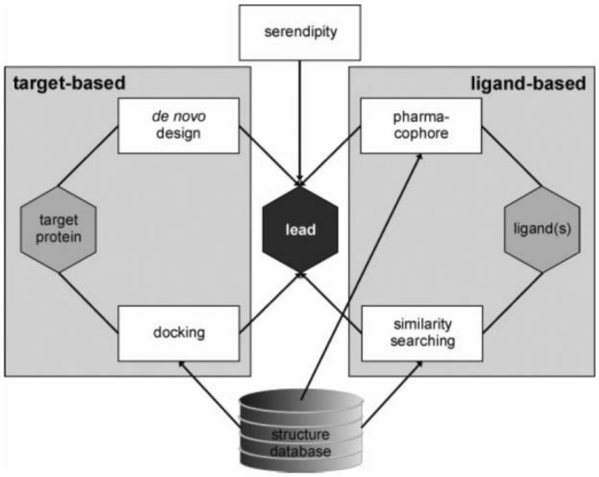


Fig: Lead discovery process

The lead structure can be discovered by serendipity. In rational drug design all information available about a target serves to direct the search for a new lead structure. If the 3D structure of the target of interest is known from X-ray crystallography, a number of methods in structure-based drug design can be applied. If an X-ray structure of the protein with a ligand is available, the binding.

Structure based drug design

Fitting a ligand from a 3D structure database into the binding site of a target protein is called docking. The iterative building of new molecules in thebinding site of a receptor are called *de novo* design. The building approach is beginning with a single fragment and proceeding through the stepwise addition of further moieties. Alternatively small molecules are placed in the binding site of the protein and subsequently linked together (linking).

To end up with high-affinity ligands a high degree of steric and electronic complementarity of the ligand to the target protein is required. Further on an appropriate amount of the ligands hydrophobic surface should be buried in the complex. A certain degree of conformational rigidity is essential to ensure that the loss of entropy upon ligand biding is acceptable.

26.a. Explain any four process involved in the modern drug discovery process

a) Target Identification & Validation

Cheminformatics is used to identify target molecule which can be either gene or protein and could be a potential drug for the disease (Gene/Protein analysis). The Identified protein is separated, crystallized and ligand binding processes are done. Some approaches will inhibit the disease functionality by making the key molecule stop functioning. Another approach is by promoting specific molecule in the normal way which may have affected in the disease state. These approaches and different databases can be applied for the discovery of drug targets. After target Identification, validation phase starts by determining whether the modulation of the target will y ield a desired clinical outcome. This is base d on the results obtained between the cellular location and disease/health condition, potential expression and protein binding state.

b) Lead Identification.

Target -to-Screening (HTS) technique is applied where the protein targets are automatically screened against database of small-molecule o r cell-based assay compounds. Lead identification also helps to see which molecules bind strongly to the target. Several similarity and diversity techniques can be applied for lead identification.

c) Lead Optimization

This phase results in finding the drug candidate from the lead identified compound. The goal is a process of refining the chemical structure of a confirmed hit to improve its drug characteristics. Several docking techniques can be applied to optimize the lead structures for target affinity and selectivity.

Different techniques and methods are used for Lead identification and Optimization process where some of the methods Virtual Screening, Molecular Database, Data mining, High-Throughput Screening (HTS), QSAR, Protein Ligand Models, Structure Based Models, Microarray analysis, Property Calculation and ADMET(adsorption, distribution, metabolism and elimination and Toxicity).

d) Pre- Clinical Trial

The preclinical stage is an important phase to check whether the compound can be made into a drug to treat specific disease which is not toxic and has minimum side effects. Toxicity tests are undertaken to show safety while pharmacokinetics testing is done to provide data on how a drug is absorbed, distributed, metabolised and excreted (ADME) from the body. Pre-clinical studies and testing can be done with or without animal testing method. In-vitro is a study, based on the test done in the clinical lab and the analysis based on living cell cultures and animal model can be referred as in –vivio method. This phase will be designed in a way such that it achieves risk-free, unproblematic and economic transition from pre-clinical to clinical trial in medical product development.

26.b. Discuss in detail about the prediction of toxicity of compounds in cheminformatics.

In general, the first step in virtual screening is the filtering by the application of Lipinski's "Rule of Five". A lead molecule should have

a. Molecular weight of less than 500 g/mol
b. A calculated lipophilicity (log P) of less than 5
c. Fewer than five H-bond donors
d. Fewer than 10H-bond acceptoprs (sum of all nitrogen and oxygen atoms)
e. Number of rotatable bonds is less than 10 or one of the four rules can be violated.

In a more recent study the physical properties of drugs in different development phases are compared. The molecular weight and lipophilicity are the properties showing the clearest influence on the successful passage of a candidate drug through the developmental process.

***In Silico* ADMET**

Pharmaceutical companies evaluate the ADMET profiles (drug absorption, distribution, metabolism, excretion and toxicity) of potential leads at an earlier stage of the development process. For the consideration of ADMET properties in virtual screening, computational methods for their

Prediction is needed. Lipophilicity is a key property for estimation of the membrane permeability of a molecule. Programs to predict Log P are available and give reasonable results.

Till now, many computational models have been developed for drug safety assessment, which could be generally divided into three categories: qualitative classification, quantitative regression and read-across. As the first step of drug safety assessment, we only need to know a compound is toxic or non-toxic, highly toxic or slightly toxic, rather than its exact toxicity value, so classification models can be used. For a small number of chemical analogs, quantitative structure-toxicity relationship (QSTR) models can be derived for prediction of exact toxicity values. For those unique compounds, read-across is also a feasible approach to deduce certain toxicity endpoint from their similar structures with experimental toxicity values. These models have high accuracies especially in a local chemical space, and sometimes they can replace in vitro or in vivo assays for certain endpoints. Furthermore, structural alerts (SAs) can be derived from the models as keys for a compound to cause adverse effects on organs, which can be used in structural modification to reduce the risk by chemists.