

# **Automatic Verification of Small Molecule Structure with One Dimensional Proton Nuclear Magnetic Resonance Spectrum**

DOCTORAL THESIS

FOR THE DEGREE OF A  
DOCTOR OF INFORMATICS

AT THE FACULTY OF ECONOMICS,  
BUSINESS ADMINISTRATION AND  
INFORMATION TECHNOLOGY  
OF THE  
UNIVERSITY OF ZURICH

by  
JIWEN LI

from  
V.R. China

Accepted on the recommendation of  
PROF. DR. A. BERNSTEIN  
PROF. DR. K. BALDRIDGE

2010

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich herewith permits the publication of the aforementioned dissertation without expressing any opinion on the views contained therein.

Zurich, April 14. 2010

The Vice Dean of the Academic Program in Informatics: Prof. Dr. H. C. Gall

# Contents

Acknowledgements.....	viii
Abstract.....	ix
List of Figures .....	x
List of Tables .....	xii
Part I.....	1
Introduction, Background and the Proposal.....	1
<i>Chapter 1</i> Motivation.....	3
1.1 Quantification .....	4
1.2 Compound Structure Verification .....	5
1.3 Applying NMR in Compound Library Management QC/QA.....	6
1.3.1 1H NMR versus 13C NMR.....	6
1.3.2 1D 1H NMR versus 2D 1H NMR .....	6
1.3.3 Applying 1D 1H NMR for Structure Verification and Quantification .....	8
1.3.4 Automating 1D 1H NMR Molecule Structure Verification .....	9
1.3.5 Expert Systems and their Applications .....	10
1.3.6 Using Artificial Intelligence for Molecule Structure NMR Verification .....	11
1.4 Structure of the Thesis.....	12
<i>Chapter 2</i> Background.....	13
2.1 Human Structure Verification Procedure with 1D 1H NMR Spectrum .....	13
2.1.1 1H 1D NMR Spectroscopy, NMR Sample and NMR Solvent .....	13
2.1.2 Basic NMR/Chemical Concepts used in Molecule Structure 1D 1H NMR Verification. ....	14
2.1.2.1 Chemical shift.....	14
2.1.2.2 Integration .....	16
2.1.2.3 J-coupling .....	17
2.1.2.4 Coupling Constant and Connectivity.....	19
2.1.2.5 Second-order Coupling .....	20
2.1.2.6 Magnetic Inequivalence.....	21
2.1.3 Human Process of Molecular Structure 1D 1H NMR Verification- an Example.....	22

2.1.3.1 Identifying Peak Clusters.....	23
2.1.3.2 Identifying Solvent .....	24
2.1.3.3 Computing Proton Numbers of Peak Clusters .....	26
2.1.3.4 Verifying Consistency between the Molecular Structure and the Peak Clusters with Proton Number .....	29
2.1.3.5 Further Verifying the Consistency between Peak Clusters and Function Groups by Coupling Analysis .....	32
2.1.4 Summary of the Human Logic for 1D 1H NMR Molecular Structure Verification .....	36
2.2 Current Automatic NMR Spectrum Molecule Structure Consistency Analysis System .....	39
2.2.1 General System Architecture .....	41
2.2.1.1 Molecular Interpreter .....	41
2.2.1.1.a Identifying Chemical Equivalent Functional Groups .....	42
2.2.1.1.b Predicting Chemical Shift .....	42
2.2.1.1.c Predicting Number of Couplings and Coupling Constant.....	43
2.2.1.1.d Count Total Number of Protons within a Molecule.....	44
2.2.1.2 NMR Spectrum Interpreter .....	44
2.2.1.2.a Automatically Identifying Peaks in Spectrum .....	44
2.2.1.2.b Grouping Symmetric Peaks into Peak Clusters .....	45
2.2.1.2.c Estimating Multiplicities and Coupling Constants for Each Peak Cluster.....	45
2.2.1.3 Consistency Analyzer .....	45
2.2.2 Difference between Human Structure Verification Logic and Techniques used in the Structure Verification System .....	47
2.2.2.1 Differences in Molecular Interpretation.....	47
2.2.2.2 Differences in NMR Spectrum Interpreter.....	48
2.2.2.3 Differences in Consistency Analysis.....	50
2.3 NMR Structure Verification Technique beyond 1D 1H NMR Spectra .....	51
2.4 Conclusion .....	51
<i>Chapter 3 The Proposal</i> .....	52
3.1 Implementation Plans .....	52
3.2 Possible Challenges .....	53
Part II.....	55
Automatic 1D 1H NMR Molecule Structure Verification Software Architecture, Methods and Evaluation .....	55
<i>Chapter 4 Automatic 1D 1H NMR Molecule Structure Verification Architecture and Methods</i> .....	57



4.1 System Architecture.....	57
4.2 Molecular Interpreter .....	59
4.3 NMR Spectrum Interpreter .....	62
4.3.1 Peak Hypothesis Generator (Deconvolution Method + Derivative Method) .....	64
4.3.2 Peak Cluster Hypothesis Generation .....	65
4.3.3 Experimental Multiplet Hypothesis Interpreter .....	66
4.4 Consistency Analyzer - Searching Consistent Peak Cluster List .....	66
4.4.1 Solvent Detection.....	70
4.4.1.1 DMSO detection.....	70
4.4.1.2 H2O Detection.....	72
4.4.2 Determine Integration Proton Ratio (Integration per Proton) .....	75
4.4.3 Matching of Experimental Peak Cluster Hypotheses and Structural Multiplet Distributions.....	78
4.4.3.1 Searching Module Architecture .....	80
4.4.4 Quantification Module .....	84
4.4.5 Creating a Structure Verification Report .....	85
Chapter 5 A Probabilistic Explanation of the System Architecture.....	86
5.1 Probabilistic Model of the Search Module .....	86
5.2 Searching Heuristics.....	87
5.3 Estimating Probability with Chemical and NMR Knowledge .....	89
5.3.1 Computing $\theta_{x_{i,cs}}^{y_j}$ .....	91
5.3.2 Computing $\theta_{x_{i,pn}}^{y_j}$ .....	92
5.3.3 Computing $\theta_{x_{i,M}}^{y_j}$ .....	92
5.3.4 Computing Coupling Constant Measure $\theta_{x_{i,J}}^{y_j}$ .....	93
5.3.5 Computing Coupling Connectivity Measure $\theta_{x_{i,con}}^{y_j, x_i, y}$ .....	94
5.3.6 Spectrum Fitting Score $\theta_{x_{i,sf}}$ and $\theta'_{x_{i,sf}}$ .....	95
5.3.7 Reliability Score $\theta_{x_{i,rel}}$ .....	96
5.3.8 Solvent likelihood $\theta_{x_{i,so}}$ .....	97
Chapter 6 Experiments.....	100
6.1 Experimental Setup.....	100
6.1.1 Evaluation Criteria.....	101
6.1.2 Evaluation Data .....	102
6.1.2.1 Real Compounds and Their Spectra .....	102

6.1.2.2 Simulated Spectra and Theoretical Multiplet Distribution Lists .....	109
6.1.3 Experimental Design to Compute $FN'$ , $FP'$ and $CR'$ .....	113
6.1.3.1 An approach to Compute $FN'$ .....	113
6.1.3.2 An approach to compute $FP'$ .....	113
6.1.3.3 An approach to compute $CR'$ .....	113
6.2 Experimental Results.....	114
6.2.1 Experimental Results of Estimating False Negative Rate(FN) .....	114
6.2.1.1 Experimental Result on Real Compound Dataset .....	114
6.2.1.2 Experimental Results of Simulated Dataset (Easy Setup) .....	115
6.2.1.3 Experimental Results of Simulated Dataset (Difficult Setup) .....	115
6.2.2 Experimental Results of Estimating False Positive Rate(FP) .....	116
6.2.2.1 Experimental Results of Real Compound Dataset .....	116
6.2.2.2 Experimental Results of Simulated Dataset (Easy Setup) .....	116
6.2.2.3 Experimental Results of Simulated Dataset (Difficult Setup) .....	117
6.2.3 Experimental Results of Estimating Consistent Rate (CR) .....	117
6.3 Discussion of the Experimental Results .....	121
6.3.1 Decision Accuracy .....	121
6.3.2 Time Complexity .....	122
6.3.3 Assignment Quality, Consistency between the System and Spectroscopists.....	124
Part III.....	131
Contribution, Limitation, Future Work and Conclusion.....	131
Chapter 7 Contribution .....	133
7.1 Impact for NMR and Pharmaceutical Industries.....	133
7.1.1 Impact on the NMR industry.....	133
7.1.2 Impact on the Pharmaceutical Industry.....	134
7.2 Contribution to Computer Science .....	135
7.2.1 Human Logic Based Optimization – a Demonstration .....	135
7.2.2 Human Logic Based Optimization versus Classical Optimization .....	138
7.2.2.1 Representation of Problem as Graph Search.....	138
7.2.2.2 Difference between Human Logic Based Optimization and Best First Search .....	140
7.2.2.3 Difference between Stochastic Optimization and Human Logic Based Optimization..	142
7.2.3 Summary of Human Logic Based Optimization.....	144
Chapter 8 Limitation .....	146

8.1 Limitation in Technology .....	146
8.1.1 Problems of Isomere, Conformer, and Hetero Coupling .....	146
8.1.2 Keeping Improving Assignment Accuracy .....	147
8.1.3 Adding 2D <sup>1</sup> H NMR and 1D <sup>13</sup> C NMR Interpretation .....	147
8.1.4 Combining the Structure Verification of NMR Spectrum with Mass Spectrum .....	147
8.2 Limitation of the Experiment .....	148
8.2.1 Limited Representativeness of Simulated Dataset .....	148
8.2.2 Limited Representativeness of Real Compound Dataset .....	148
8.3 Limitation in Industrialization .....	149
8.3.1 NMR Automation Hardware .....	149
8.3.2 Link to Compound Library Management Automation .....	149
<i>Chapter 9</i> Future Work .....	151
9.1 Future Work in NMR/Pharmaceutical Industry .....	151
9.2 Future Work in Applied Computer Science.....	152
<i>Chapter 10</i> Conclusion .....	153
Part IV.....	156
Appendix .....	156
A. Glossary .....	157
B. References .....	161
C. List of Detailed Assignments of 81 Spectrum-Structure Pairs .....	167
D. Curriculum Vitae.....	258

# Acknowledgements

I would like to gratefully acknowledge the enthusiastic supervision of Prof. Abraham Bernstein of my dissertation. Avi, I sincerely thank you for giving me the opportunity to address an important problem in practice and write the dissertation about it. Despite my lengthy wander in research, you insisted on guiding me into the right direction and helping me to build confidence in my research with great patience. This makes me appreciate the “real” meaning of research and finding an appropriate balance between theory and practice. I also thank you for giving me a wide angle of views to look at the research. Only with your help, I can “jump” into the “general” science to look at computer science, and convert myself into an interdisciplinary researcher. In addition, thank you for giving me enough time to read research books since this helped me to build a strong basis to support the research. I would also like to thank Prof. Kim Baldrige and Prof. Burkhard Stiller for agreeing to be members of my Ph.D. committee. Kim, I especially thank you for many discussions on my thesis, and all the knowledge you gave to me in both computer science and organic chemistry.

Furthermore, I would like to thank Dr. Isabelle Guyon in helping me to understand how to research the problem from the machine learning perspective. I thank Prof. Wang Yuandi to help me build a laid architecture of the modern mathematics, and having conceptive understanding of the architecture. I cherish the year I have spent with above two persons in Zurich, and their help is extremely valuable in my “growing-up” in research. I’d also like to thank Prof. Konstantin Pervushin, Dr. James Masse, Dr. Till Kuehn, Dr. Sandra Loss, Dr. Bjoern Heitmann, Dr. Michael Fey, Dr. Jochen Klages for four years cooperation in the project which my thesis is based upon. All of you have research experiences from different academic backgrounds, and these often helped me to consider the problem from an angle beyond what I can learn from a computer science department. I thank all of you to give me the deep knowledge in organic chemistry and nuclear magnetic resonance.

I am grateful to all my friends from the Informatics Department, the University of Zurich, for being the “family” during the many years I stayed there and for their continued support thereafter. Jonas Luell, I especially thank you for all advises you gave to me at both the research level and the personal level. It is always full of fun to have a conversation with you, and thank you for your care and attention.

Finally, I am forever indebted to my parents and my wife for their understanding, endless patience and encouragement when it was most required.

# Abstract

Small molecule structure one dimensional (1D) proton ( $^1\text{H}$ ) Nuclear Magnetic Resonance (NMR) verification has become a vital procedure for drug design and discovery. However, the inefficient throughput of human verification procedure has limited its application only to an arbitral instrument for molecular structural identification. Considering NMR's unimpeachable advantages in molecular structural identification tasks (compared to other techniques), to popularize NMR technology into routine molecular structural verification procedures (especially in compound library management of the pharmaceutical industry), will dramatically increase the efficiency of drug discovery procedures. As a result, some automatic NMR structure verification software approaches were developed, described in the literature and are commercially available. Unfortunately, all of them are limited in principal (e.g. they heavily depend on the chemical shift prediction) and are shown not to be working in practice.

Driven by the strong motivation from the industry, we propose a new approach as an alternative to approach the problem. Specifically, we propose to utilize approaches from artificial intelligence (AI) to mimic the spectroscopist's NMR molecular structure verification procedure. Guided by this strategy, a human-logic based optimization (i.e. heuristic search) approach is designed to mimic the spectroscopist's decision process. The approach is based on a probabilistic model that is used to unify the human logic based optimization approach under maximum likelihood framework. Furthermore, a new automatic 1D  $^1\text{H}$  NMR molecular structural verification system is designed and implemented based on the optimization approach proposed earlier.

In order to convince vast NMR spectroscopists and molecular structural identification participators, comprehensive experiments are used to evaluate the system's decision accuracy and consistency to the spectroscopists. The results of the experiments demonstrate that the system has very high performance in terms of both accuracy and consistency with the spectroscopists on the test datasets we used<sup>1</sup>. This result validates both the correctness of our approach and the feasibility of building industrialized software based on our system to be used in practical industrial structural verification environments. As a result, commercial software based on our system is under development by a major NMR manufacture, and is going to be released to the pharmaceutical industry.

Finally, the thesis also discusses similarities and differences between the human logic based optimization and other typically used optimization approaches, and especially focuses on their applicability. Through these discussions, we hope that the human logic based optimization could be used as a reference by other practical computer science participants to solve other automation problems from different domains.

---

<sup>1</sup> To be convenient for the evaluation of vast molecular structural identification practitioners, detail structural verification reports of 81 compounds generated by the system are cataloged in the thesis' appendix.

# List of Figures

Fig 1 Anatomy of 1D NMR Experiment, and Sample 1D NMR Spectrum .....	7
Fig 2 Anatomy of 2D NMR Experiment, and Sample 2D NMR Spectrum .....	8
Fig 3 A NMR spectrometer, a NMR sample and Structure of DMSO .....	14
Fig 4 1D <sup>1</sup> H NMR spectrum and Molecule Structure of Ethanol .....	16
Fig 5 Pascal Triangle .....	17
Fig 6 First Order Multiplet Pattern (a) .....	17
Fig 7 First Order Multiplet Pattern (b) .....	18
Fig 8 Coupling Constant and Connectivity of Ethanol.....	19
Fig 9 Roof Top Effect .....	20
Fig 10 Molecule Structure and Magnetic Inequivalent Multiplet Pattern in the 1D <sup>1</sup> H NMR spectrum of 1,2-dichlorobenzene.....	21
Fig 11 Molecule Structure and 1D <sup>1</sup> H NMR Spectrum of +-Pseudoephedrin.....	22
Fig 12 Peak Clusters Identified from 1D <sup>1</sup> H NMR Spectrum of +-Pseudoephedrin.....	23
Fig 13 H <sub>2</sub> O and DMSO Patterns in 1D <sup>1</sup> H NMR Spectrum of +-Pseudoephedrin.....	26
Fig 14 Identify Functional Group and Proton Numbers from +-Pseudoephedrin .....	27
Fig 15 Proton Numbers of Peak Clusters of +-Pseudoephedrin.....	28
Fig 16 Assignments between Peak Clusters and Functional Groups with Proton Number on +- Pseudoephedrin.....	32
Fig 17 Protons in Aromatic Ring and Their Complex Peak Cluster Patterns.....	34
Fig 18 Identify Functional Group and Proton Numbers from Pseudoephedrin.....	34
Fig 19 One-to-one assignments between peak clusters and functional groups of +-Pseudoephedrin.....	35
Fig 20 Human Logic for 1D <sup>1</sup> H NMR Structure Verification .....	38
Fig 21 Structure of NMR Structure Verification System .....	41
Fig 22 Chemical Equivalent Protons in Ethanol .....	42
Fig 23 (a) High order multiplet and (b) overlap of first order multiplets.....	49
Fig 24 Example of missing experimental multiplet interpretations.....	49
Fig 25 System Flow Chart.....	58
Fig 26 Molecular Interpreter Module Flow Chart.....	61
Fig 27 NMR Spectrum Interpreter Module Flow Chart.....	63
Fig 28 Peak Pick Routine Flow Chart .....	65
Fig 29 Searching Peak Cluster List Module .....	69
Fig 30 Solvent Detection .....	75
Fig 31 Integration Proton Ratio Computation Flowchart .....	77
Fig 32 Searching Module Out Loop Flow Chart.....	81
Fig 33 Searching Module Internal Loop Flow Chart.....	83
Fig 34 The Order to Build the Peak Cluster List .....	88
Fig 35 An example of a simulated spectrum (the first setup) .....	109
Fig 36 An example of a simulated theoretical multiplet distribution list (the first setup) .....	110
Fig 37 An example of a simulated spectrum (the second setup).....	111
Fig 38 An example of a simulated theoretical multiplet distribution list (the second setup) .....	112
Fig 39 Automatic assignments between NMR spectrum and structure of +-Pseudoephedrin .....	118

Fig 40 Average Time Expenses on Different Datasets .....	123
Fig 41 Wrong assignments by the system, and their corrections on Essigsaeurelinallylester .....	126
Fig 42 Wrong assignments by the system, and their corrections on Benzonitril .....	127
Fig 43 Peak Cluster should Split on Linalool.....	128
Fig 44 Peak Cluster should not Split on 1-Octyne .....	129
Fig 45 Demo –Input (a).....	136
Fig 46 Demo –Step1 (b).....	136
Fig 47 Demo –Step2 (c) .....	136
Fig 48 Demo –Step3 (d).....	136
Fig 49 Demo –Step4 (e).....	137
Fig 50 Demo –Step5 (f).....	137
Fig 51 Demo –Step6 (g).....	137
Fig 52 Demo –Output (h) .....	137
Fig 53 Problem Setup .....	139
Fig 54 Search Space Structure I.....	139
Fig 55 Search Space Structure II.....	139
Fig 56 Search Space Structure III.....	140

# List of Tables

Table 1 Typical Chemical Shift Ranges for Various Functional Groups.....	15
Table 2 Coupling Constant Ranges for Various Functional Groups in Common Use.....	19
Table 3 Deuteration Degree of DMSO .....	24
Table 4 List of Compounds Used in Evaluation.....	108
Table 5 Experimental Result of Estimating FN on Real Compound Dataset.....	114
Table 6 Experimental Result of Estimating FN on Simulated Dataset (The First Setup).....	115
Table 7 Experimental Result of Estimating FN on Simulated Dataset (The Second Setup) .....	115
Table 8 Experimental Result of Estimating FP on Real Compound Dataset .....	116
Table 9 Experimental Result of Estimating FP on Simulated Dataset (The First Setup) .....	116
Table 10 Experimental Result of Estimating FP on Simulated Dataset (The Second Setup) .....	117
Table 11 Experimental Result of Estimating CR on Real Compound Dataset.....	121



## **Part I**

### **Introduction, Background and the Proposal**



## **Chapter 1 Motivation**

In medicine, biotechnology and pharmacology, drug discovery is the process by which drugs are discovered and/ or designed. In the past, most drugs have been discovered either by identifying the active ingredient from traditional remedies or by serendipitous discovery. In contrast to this, modern drug discovery processes focus on understanding how disease and infection are controlled at the molecular and physiological level, and targeting specific macromolecules (proteins or nucleic acids in most cases) based on this knowledge. This change is due to the scientific conclusion that the effectiveness of the drug in the human body is mediated by specific interactions of the drug molecule with biological macromolecules. As a result, in the modern era of pharmacology, pure chemicals, instead of crude extracts, become standard drugs. And drug discovery becomes the process to identify organic molecules<sup>2</sup> that could effectively interact with specific macromolecules in the human body.

The process of finding a new molecule against a chosen target (macromolecule) for a particular disease usually involves high-throughput screening (HTS) (Bailing, et al., 2004) (Burbaum, et al., 1997) (Hann, et al., 2004), wherein large libraries of molecules are tested for their ability to modify the target. For example, if the target is a novel G protein-coupled receptor (GPCR), molecules will be screened for their ability to inhibit or stimulate that receptor. If the target is a protein kinase, the molecules will be tested for their ability to inhibit that kinase. Another important function of HTS is to show how molecules are selective for the chosen target, but not for other related macromolecules. This cross-screening is also important since the more unrelated targets a molecule hits, the more likely that off-target toxicity will occur with that molecule once it reaches the clinic. A drug discovery process normally requires several iterative HTSs, in which it hopes that the properties of the new compound will be found and (or) improved. Once a compound has been found with sufficient target potency and selectivity, it will be proposed for drug development.

HTS's in the drug discovery use compound libraries, wherein a large collection of organic compounds are stored, and each compound also has associated information such as the molecular structure, purity, quantity, and other physiochemical characteristics of the compound stored in the database. Chemical compounds are usually designed by organic chemists and computational chemists and synthesized by organic chemists and medicinal chemists. Because of the expense and the effort involved in chemical synthesis, the compounds must be correctly stored for later use to prevent early degradation. In a typical chemical library, each chemical has a particular shelf life and storage requirement, and there is a timetable by which library compounds are to be disposed of and replaced on a regular basis. Since quantity of all possible organic compounds is large and increases exponentially with the size of the molecule, the inventory of a compound library could easily reach up into millions of compounds, which makes the management of even a modest-sized compound library a full-time endeavor. To relief the quantity of the routine workload, robots have been used to automate the compound storage (Chan, et al., 2002).

---

<sup>2</sup> Note, in the scope of the thesis, without special explanation, the term "molecule" means small molecule. Here a small molecule is a low molecular weight organic compound which is by definition not a polymer.

Compound library is the test object of HTS's. Therefore, the output of HTS's relies on the quality of the compound library. To guarantee the effectiveness of HTS's, a quality control (QC) and quality assurance (QA) system is established in chemical library management, where library entities need to be determined and rechecked on their analytical characterization in a regular basis during their shelf-life. Typical jobs involved in identifying analyte characterizations include the compound's molecular structure verification, quantification, purity determination, etc.

## 1.1 Quantification

In the scope of this thesis, quantification is defined as the procedure to determine the molar concentration of the main chemicals in a liquid sample, whereupon the solvent (e.g. Deuterated Dimethyl Sulfoxide (DMSO)<sup>3</sup>) and impurities that are connected to the solvent (e.g. water (H<sub>2</sub>O) in DMSO) are not considered as main chemicals.

Chemical concentration is indispensable information for HTS. During HTS, the decision whether one of the compounds is further investigated as a potential drug candidate for a specific disease or not is based on a binding experiment of the substance to a certain target. However, the accuracy of the binding constant provided by these studies strongly depends on the accuracy of the molar concentration of the compound, which can change drastically over time due to degradation or fallout of the solution (Popa-Burke, et al., 2004). Thus, the concentration of these chemicals in the library needs to be determined and revalidated on a regular basis to prevent false positive hits.

The traditional approach to quantify a compound is to weigh the dry compound on the scale. This approach is inaccurate and sensitive to the amount of impurity in the sample. To relief the problem, two instrumental analysis techniques have begun to be used to address the issue of quantitative analysis- chemiluminescent nitrogen detector (CLND) (Corens, et al., 2004) and evaporative light-scattering detector (ELSD) (Fang, et al., 2000).

The principle of CLND is based on measuring nitrogen content of a sample. With the knowledge of the number of nitrogen atoms in a molecule of analyte, one can determine the sample quantity. Literature has shown that this approach is very promising for the quantitative analysis of combinatorial compound libraries (Taylor, et al., 2002) (Sepetov, et al., 1999). However, it requires that compounds contain nitrogen and does not allow the use of any nitrogen-containing solvent during analysis.

ELSD, as another instrumental analysis technique, creates an aerosol from a sample, and then determines the sample concentration by measuring the amount of diffused light on the aerosol. Note, the relationship between the amount of diffused light and the amount of analyte can be precisely described by the mathematical formula. However, chemical practice has shown the approach is not very accurate in quantification.

---

<sup>3</sup> **DMSO** – a solvent often used to store organic compounds of compound libraries in the liquid phase.

As a result, organic chemists still keep looking for better sample quantification approaches for compound library management.

## 1.2 Compound Structure Verification

In the scope of this thesis, compound structure verification is defined as the process to check if a given molecule structure is consistent with the spectroscopically measured structural information. Specifically, mainly two spectrometric techniques are used for compound structure verification: Mass spectroscopy, and Nuclear Magnetic Resonance (NMR) spectroscopy, with some additional confirmation of the structure provided by IR spectroscopy and X-Ray crystallography (Pretsch, et al., 2009).

Mass spectroscopy is based on the measurement of a fundamental characteristic of the compound: mass-to-charge ratio of the molecule, after ionization of the molecule. These mass-charge-ratio patterns can give chemists hints to “guess” the potential structure of given compounds. Despite the inability of mass/charge patterns to discriminate the subtle difference of the molecule structure, this technique has been identified as the method of choice for the high-throughput structure confirmation of compounds in compound library management (Sepetov, et al., 1999). There are a few reasons for this choice. For example, the method does not depend on the presence of chromophores<sup>4</sup> or any functional group in a molecule. High sensitivity is another advantage of mass spectrometry: as little as femto-molars of a compound can be easily measured. In addition, mass spectrometry is a fast method, with the measurement time approximately several seconds, and it can be easily automated. Unfortunately, mass spectrometry cannot be used to determine the concentration of the compound, since mass experiment begins with ionization of the analyte, and compounds with the same concentration in the analyte may have different abilities to be ionized, and thus give substantially different response in mass spectra (Sepetov, et al., 1999).

Compared to mass spectrometry, NMR spectrometry is the most informative method for characterization of organic compounds. It yields peaks in nuclear magnetic resonance spectrum with individual hydrogen and carbon atoms in the molecular structure, which allows detailed reconstruction of the molecule’s architecture. However, NMR is a relatively insensitive and slow method, it requires homogenous samples, and consumes expensive deuterated solvents. As a result, NMR has been limited to be used mainly for the structural identification of “interesting” compounds found during HSTs, and has not been used in routine quality control (QC) and quality assurance (QA) of compound library management.

---

<sup>4</sup> A chromophore is part (or moiety) of a molecule responsible for its color.

## 1.3 Applying NMR in Compound Library Management QC/QA

In this subsection, we first briefly introduce the major NMR experiments which are used to identify structures of molecules, and analyze their applicability to compound library management. We conclude that the bottleneck of applying NMR to compound library management lies on automating the spectroscopist's 1D  $^1\text{H}$  NMR spectra interpretations. Next, we demonstrate several examples of how modern artificial intelligence (AI) technologies are used to automate domain expert's decision making procedures in various application fields. Referring to these successful stories, finally we propose to utilize AI technologies to mimic the NMR spectroscopist's spectra interpretation process in order to automate this human procedure.

### 1.3.1 $^1\text{H}$ NMR versus $^{13}\text{C}$ NMR

As we introduced in 1.2, in principal, NMR techniques supply chemists with more detailed information about compound structure, which make it a potential technology to improve current compound library management QC/QA. Specifically, there are two types of NMR techniques mainly involved in compound structure verification:  $^1\text{H}$  NMR and  $^{13}\text{C}$  NMR.

$^1\text{H}$  NMR (also called Proton NMR or Hydrogen NMR) is the application of NMR spectroscopy with respect to hydrogen-1 nuclei within the molecules of a substance, in order to determine the structure of its molecule.

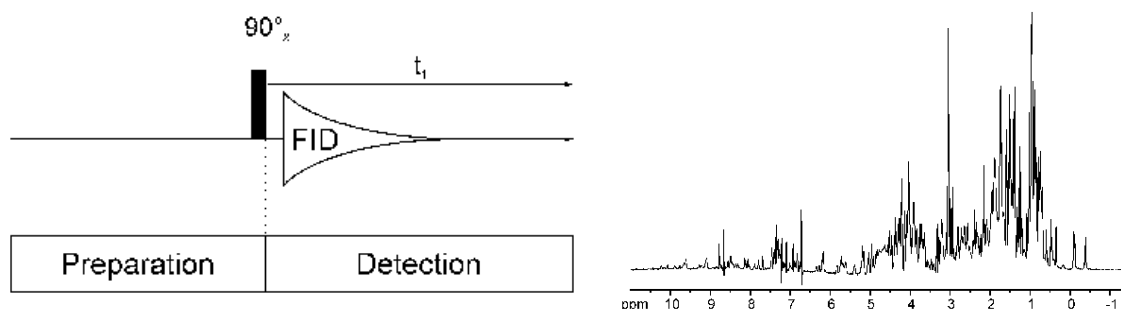
Comparably,  $^{13}\text{C}$  NMR (also called Carbon-13 NMR) is the application of NMR spectroscopy with respect to carbon. It is analogous to  $^1\text{H}$  NMR and allows the identification of carbon atoms in an organic molecule to determine molecule structure.

However,  $^{13}\text{C}$  NMR detects only the  $^{13}\text{C}$  isotope of carbon, whose natural abundance is only about 1.1% (exiguous), while the main carbon isotope,  $^{12}\text{C}$  is not detectable by NMR. This makes  $^1\text{H}$  NMR a lot more sensitive compared to  $^{13}\text{C}$  NMR, where  $^1\text{H}$ 's nature abundance is more than 99.9% (abundant). As a result,  $^1\text{H}$  NMR becomes the main approach for compound structure elucidation, while  $^{13}\text{C}$  NMR is used as an accessorial approach to supplement  $^1\text{H}$  NMR.

### 1.3.2 1D $^1\text{H}$ NMR versus 2D $^1\text{H}$ NMR

Multiple types of  $^1\text{H}$  NMR experiments could be generated by NMR spectroscopy, where two types of experiments are generally used in small molecule structure verification. They are one dimensional  $^1\text{H}$  NMR ( $^1\text{H}$  1D NMR) experiments and two dimensional  $^1\text{H}$  NMR ( $^1\text{H}$  2D NMR) experiments. Note,

both inventions of 1D and 2D experiments were acknowledged by Nobel prizes. Fig 1<sup>5</sup> shows the anatomy of the 1D NMR experiment and an example of how a resulting 1D NMR spectrum looks like. 1D NMR experiment consists of two sections: preparation and detection. During preparation, by giving a radio frequency pulse (for example 90 degree pulse), the spin systems of the molecule is set to a defined state. Then during detection, the resulting nuclear magnetic resonance, named free induction decay (FID), is recorded during time interval  $t_1$ . After that, the FID signal is Fourier transformed to yield the 1D NMR spectrum.



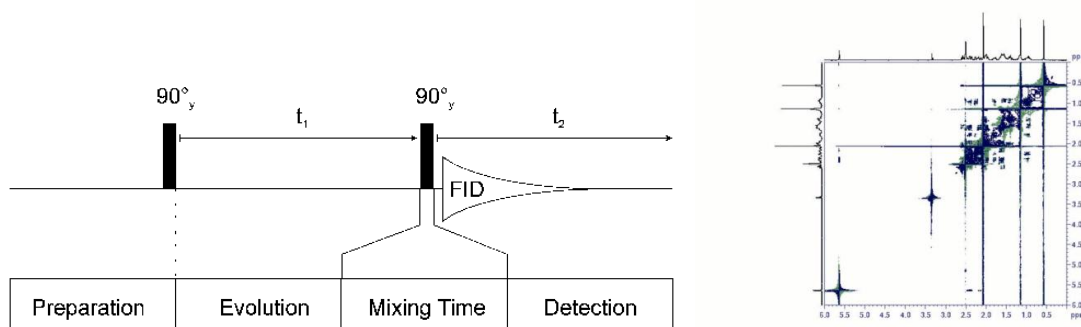
**Fig 1 Anatomy of 1D NMR Experiment, and Sample 1D NMR Spectrum**

In contrast, a 2D NMR experiment can be understood as a series of 1D NMR experiments. Each experiment consists of a sequence of radio frequency pulses with delay periods between them. It is the timing, pulse frequencies, and intensities of these pulses that distinguish different 2D NMR experiments from each other. During the decays, the nuclear spins are allowed to freely rotate for a determined length of time known as the evolution time. The frequencies of the nuclei are detected after the final pulse. By incrementing the evolution time in successive experiments, a two-dimensional data set is generated from a series of one-dimensional experiments. An example of a 2D NMR experiment is the homo-nuclear correlation spectroscopy (COSY) sequence, which consists of a pulse ( $p_1$ ) followed by an evolution time ( $t_1$ ) followed by a second pulse ( $p_2$ ) followed by a measurement time ( $t_2$ ). Then, a 2 dimensional Fourier transform is performed along dimensions of  $t_1$  and  $t_2$  to generate the 2D NMR spectrum. The anatomy of COSY 2D NMR experiment and a sample spectrum are shown in Fig 2<sup>6</sup>.

2D NMR spectra can provide additional information about the structure of a molecule, which 1D NMR spectra cannot supply, and these are especially useful in determining the structure of a molecule that are too complicated to be interpreted with 1D NMR experiments alone. For example, cross peaks – points that are symmetric along the diagonal from the bottom left to the upper right of the 2D spectrum in Fig 2 give us additional information about which peaks (that represent different nucleus ) in 1D NMR are coupled (interacted). Note, the principal of 2-dimensional (high-dimensional) NMR experiments concerns complex physical procedures, which are beyond the scope of this thesis. Therefore, we refer the interested readers to a NMR textbook for example (Keeler, 2005).

<sup>5</sup> Images in Fig 1 are sourced from PPS2 projects for the determination of protein structure by NMR spectroscopy from Birkbeck, University of London. The original images are located at <http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/1dnmr.htm>.

<sup>6</sup> Images in Fig 2 are sourced from PPS2 projects for the determination of protein structure by NMR spectroscopy from Birkbeck, University of London. The original images are located at <http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/2dnmr.htm>.



**Fig 2 Anatomy of 2D NMR Experiment, and Sample 2D NMR Spectrum**

In practice, 2D NMR experiments are composed of a set of 1D type NMR experiments, which makes it often 2 orders slower than 1D NMR experiments. As we have explained in 1.2, even 1D NMR experiments are shown slower compared with mass experiments in molecule structure verification tasks, the dramatically large time consumption of 2D experiments makes them incompetent for routine QC/QA of compound library management.

As a result, improving the sensitivity and acquisition speed of 1D <sup>1</sup>H NMR becomes the only possibility of pushing the NMR application into compound library management QC/QA.

### 1.3.3 Applying 1D <sup>1</sup>H NMR for Structure Verification and Quantification

Powered by the requirement of better QC/QA of compound library management, new NMR probes (Macnaughtan, et al., 2003) (Wang, et al., 2004) and automation techniques (e.g. automatic sample changers or flow-injection systems) keep emerging in the NMR engineering field to improve the sensitivity and spectrum acquisition speed of 1D <sup>1</sup>H NMR. With these new breakthroughs in NMR spectroscopy (especially in probe technology), for example, one can acquire quantitative 1D <sup>1</sup>H NMR spectra of 5 Mikroliter (μl) of a 10 millimolar (mM) solution in non-deuterated DMSO within two minutes with commercial NMR spectrometer (e.g. from Bruker Biospin AG<sup>7</sup>). These technical breakthroughs make a 1H 1D NMR experiment only two orders of magnitudes slower compared to the mass experiment. In addition, in the practical compound library management environment, mass spectrometry is linked with liquid chromatography (HPLC), and utilizes HPLC as the pre-device to separate and quantify compounds. This preprocess of HPLC is often slow, which makes the total acquisition time of 1D <sup>1</sup>H NMR spectrum shorter than the corresponding time consumption of HPLC-MS, (which often takes 8 minutes). Thus, it becomes possible to shift part of the structure verification tasks in compound library management QC/QA from mass spectroscopy to more accurate 1H 1D NMR spectroscopy, assuming that 1D <sup>1</sup>H NMR spectrum interpretation is not a time consuming task.

<sup>7</sup> Bruker is one of leading NMR manufactures.



Another advantage of the 1D  $^1\text{H}$  NMR technique is that it could be used to determine the concentration of the analyte, while mass technology has to be combined with HPLC for measuring concentration. This technique provides an alternative method for determining the molar concentration of compounds in solution without prior knowledge of their molecular weight, which makes it particularly useful when sub-milligram quantities of compound are to be analyzed and applicable to compound library management (Pierens, et al., 2008). Specifically, the NMR approach of quantification consists of two parts:

- (1) Identifying a signal (which is possibly featuring a fine structure) of the main substance in the 1D  $^1\text{H}$  NMR spectrum, determining the number of  $^1\text{H}$  nuclei that generate this signal, and measuring the signal area underneath this signal.
- (2) This signal is quantitatively compared to a reference signal for which the number of  $^1\text{H}$  nuclei, the area underneath the signal, and the molar concentration is known resulting in the wanted molar concentration of the main substance.

1D  $^1\text{H}$  NMR quantification is a lot more accurate than other methods we discussed in 1.1. Using a good signal to noise spectrum with correct phasing and baseline correction, it has been shown that 1D  $^1\text{H}$  NMR quantification can have accuracy of less than 5% relative deviation from the real concentration (Pinciroli, et al., 2001). In addition, the principle that it uses  $^1\text{H}$  nuclei signal for concentration determination makes it universal for all kinds of organic compounds. However, quantification with 1D  $^1\text{H}$  NMR relies on molecule structure NMR verification. It has a risk that a signal from the NMR spectrum which belongs to impurities instead of main substance is used to compute the concentration. To guarantee the accuracy of quantification, a complete molecule structure NMR structure verification process has to be carried out to select the correct NMR signal (which is generated from the main substance). In other words, quantification with 1D  $^1\text{H}$  NMR is in fact a byproduct of the molecule structure 1D  $^1\text{H}$  NMR spectrum verification. Therefore, in the scope of the thesis, we focus on the explanation of the structure verification process itself.

### 1.3.4 Automating 1D $^1\text{H}$ NMR Molecule Structure Verification

Technical breakthrough in NMR hardware (especially in probe technology) has shifted the bottleneck of extending 1D  $^1\text{H}$  NMR application from NMR spectrum acquisitions to interpretation of 1D  $^1\text{H}$  NMR spectra. The 1D  $^1\text{H}$  NMR spectrum interpretation is an empirical procedure and consumes human effort (Detail see 2.1). Surveys on NMR spectroscopists show that a top structure NMR verification expert has maximal capacity of interpreting only 100 1D  $^1\text{H}$  NMR spectra per day, and with this interpretation speed, he/she gets quickly exhausted. This natural slowness of NMR spectrum interpretation creates a new bottleneck, and continues to keep NMR out of routine QA/QC of compound library management, where a scale of million compounds needs to be identified towards their molecular structures.

Driven by the motivation of popularizing NMR in molecular structure verification tasks, during the past 20 years, several approaches are proposed in academic world and/or implemented as commercial software to automate 1D  $^1\text{H}$  NMR spectrum interpretation. The majority of these approaches focuses on 1D  $^1\text{H}$  NMR spectrum prediction, followed by comparison of the predicted spectrum and measured spectrum (Castiglione, et al., 1998) (Griffiths, 2000) (Griffiths, 2001)

(Griffiths, et al., 2002) (Griffiths, et al., 2004) (Griffiths, 2005) (Griffiths, et al., 2005) (Jansma, et al., 2005) (Golotvin, et al., 2006). Unfortunately, these approaches and the corresponding software have been shown unreliable for structural verification tasks in the practical application environment. As a result, they have not been applied to compound library management (for detailed explanation see 2.2). Recently, relatively new approaches are proposed to improve the previous systems by supplementing 1D <sup>1</sup>H NMR structural verification with 2D <sup>1</sup>H NMR structural verification, in which additional information about peak correlations is supplied (Golotvin, et al., 2007) (Schröder, et al., 2000). However, structure verification accuracies of these new approaches are not convincing, either. In addition, the strategy of turning to 2D <sup>1</sup>H NMR technology dramatically increases the acquisition time of the NMR system (see 1.3.2), and this in turn diminishes the advantage of NMR to HPLC-MS in time expense of acquisition. As a result, the pharmaceutical industry still relies on human 1D <sup>1</sup>H NMR spectrum verification approaches as their major resort for molecule structure NMR identification/ verification process. Due to the low throughput of the human interpretation procedure for QC/QA in compound library management, they still rely on mass spectrum based analysis technology.

### 1.3.5 Expert Systems and their Applications

Artificial intelligence (AI) is the branch of computer science which aims to create intelligent machine. After half a century's development, unfortunately, AI research is still far away from its original goal – to build a general intelligent system. However, the technologies created in AI research have been adopted in a wide range of fields (e.g. medical diagnosis, stock trading, robot control, scientific discovery, etc), and are often used as elements of larger information systems (Kurzweil, 2006) (Committee on Innovations in Computing and Communications: Lessons from History, 1999). In the field of AI, a sub-domain named expert system is particularly oriented toward the application domain. Specifically, an expert system is a computer application that solves complicated problems that would otherwise require extensive human expertise. To do so, it mimics the human reasoning process of applying the domain knowledge to solve the specific problem in the domain, for which the process itself would normally require human intelligence.

Many expert systems have been developed to solve problems in multiple domains. For example, in the financial domain, an expert system named Mavent Expert System (Steinmann, et al., 1991) has been built for the Federal National Mortgage Association (FNMA) to assist with mortgage application. Specifically, a set of mortgage application rules are captured from loan officers, and it is used to (1) judge whether all conditions for granting a particular type of loan to a given client have been satisfied, (2) calculate the required term of repayment according to the borrower's, (3) and evaluate means and the security to be obtained from the client. It has been proven that the system can produce results which are correct in 80-90% of all cases, and due to this accuracy it supplies a significant amount of assistance to the bank branch. In addition, the explanation facilities of the system of how it reaches its decisions are built in a way to make the decision process visible so it can be confirmed by the loan officers.

Another successful application domain of expert system is medical diagnosis. In medical diagnosis, it is difficult for physicians to transfer their knowledge into distinct rules. Instead, they apply the rules with a certain amount of uncertainty. To adapt to the characteristics of the diagnostic process as carried out by the physician, the expert systems in medical diagnosis often adopt probabilistic reasoning techniques such as Bayesian network (Pearl, 2000), and Bayesian logic (Berger, 1993) to deal with the uncertainty embedded in medical diagnosis. For example, DXplain (Barnett, et al., 1987) (2009) is a Clinical Decision Support System (CDSS) (Berner, 1998) designed by the Laboratory of Computer Science at the Massachusetts General Hospital that assists clinicians by generating stratified diagnoses based on user input of patient signs and symptoms, laboratory results, and other clinical findings. Evidential support for each differential diagnosis is presented along with recommended follow-up that may be conducted by the clinician to arrive at a more definitive diagnosis. DXplain generates ranked diagnoses associated with the symptoms using a modified form of Bayesian logic. Specifically, each clinical finding entered into DXplain is assessed by determining the importance of the finding and how strongly the finding supports a given diagnosis for each disease in the knowledge base. Using this criterion, DXplain generates ranked differential diagnoses with the most likely diseases yielding the highest rank. Using stored information regarding each disease's prevalence and significance, the system differentiates between common and rare diseases. Analysis of accuracy has shown promise in DXplain. In a preliminary trial investigation of 46 benchmark cases with a variety of diseases and clinical manifestations, the ranked differential diagnoses generated by DXplain were shown to be in alignment with a panel of five board-certified physicians (Feldman, et al., 1991). In another study investigating how well decision support systems work at responding to a bioterrorism event, an evaluation of 103 consecutive internal medicine cases showed that Dxpain correctly identified the diagnosis in 73% of cases, with the correct diagnosis averaging at a rank of 10.7 (Bravata, et al., 2004). As a result, usage of DXplain as a tool for medical consultation has been common to some institutions since it fills a gap, particularly for medical students in clinical rotations, which are not adequately covered by textbook literature (London, 1998). The large knowledge base of the system combined with its ability to formulate diagnostic hypotheses have made it a popular education tool for US-based medical schools, and by 2005 DXplain was supporting more than 33,189 total users (Barnett, 2004).

### 1.3.6 Using Artificial Intelligence for Molecule Structure NMR Verification

Previous successes of Expert Systems in various application domains and their substantial backbone – mimicking human logic – propose a new strategy to approach the problem of automating the molecule structure 1D <sup>1</sup>H NMR verification procedure. In addition, the fact that human molecule structure NMR verification processes have been proven to be the only reliable structure verification process reinforces the motivation. As a result, in this thesis we explore and discuss how to utilize technologies developed in the artificial intelligence domain (especially in expert system domain) to build an automatic molecule structure 1D <sup>1</sup>H NMR verification system.

## 1.4 Structure of the Thesis

With the motivation declared above,

in Chapter 2, we explain the human molecule structure NMR spectrum verification process in detail and introduce current available automatic molecule structure NMR spectrum verification technologies and systems.

In Chapter 3, we propose our view of how to solve the problem, and explain our goal.

In Chapter 4, we explain in detail about our system design.

In Chapter 5, we give a probabilistic explanation of the system design, and further explain the computational details of the system.

In Chapter 6, we describe our evaluation approach and experiment result.

In Chapter 7, we conclude the contribution of our work to the pharmaceutical industry, and further discuss the contribution of our new optimization principal to applied computer science research.

In Chapter 8, we analyze the limitation of our current system.

In Chapter 9, we propose the directions to further improve the current system.

And Chapter 10, we give the conclusion.

## **Chapter 2 Background**

In this chapter, we first provide some background knowledge about 1D  $^1\text{H}$  NMR spectra, and explain the human interpretations of 1D  $^1\text{H}$  NMR spectra with an example in section 2.1. Consecutively, we focus on introducing current automatic NMR spectra analysis technologies in section 2.2.

### **2.1 Human Structure Verification Procedure with 1D $^1\text{H}$ NMR Spectrum**

In this subsection, first we give a short explanation of NMR spectroscopy, NMR samples and NMR solvent. After that, we introduce the background knowledge to NMR that spectroscopists use to interpret 1D  $^1\text{H}$  NMR spectra. Consecutively, we utilize an example to demonstrate the human structure verification procedure. Finally, we summarize this human process with a flowchart.

#### **2.1.1 $^1\text{H}$ 1D NMR Spectroscopy, NMR Sample and NMR Solvent**

1D  $^1\text{H}$  NMR spectroscopy is an instrumentation to apply nuclear magnetic resonance technology with respect to the isotope  $^1\text{H}$  of hydrogen (hydrogen-1 or proton) nuclei within the molecules of a substance, in order to determine the structures of its molecules. Typical analytes of 1D  $^1\text{H}$  NMR spectroscopy are organic compounds, in which the isotope  $^1\text{H}$  of hydrogen (hydrogen-1) universally exists (this is due to the high nature abundance ( $> 99.9\%$ ) of the isotope  $^1\text{H}$ ). Ubiquity of natural hydrogen in organic compounds guarantees that 1D  $^1\text{H}$  NMR technology is universally applicable for structural determination tasks of vast chemicals stored in compound libraries.

1D  $^1\text{H}$  NMR spectra are recorded in solution samples, and obviously solvent protons must not be allowed to interfere with the NMR signals from the target compound. Therefore, solvents without hydrogen, such as carbon tetrachloride or trifluoroacetic acid are often used. More commonly, deuterated (deuterium =  $^2\text{H}$ , often symbolized as D) solvents are especially popular to be used in NMR experiments. For example, deuterated dimethyl sulfoxide (DMSO), which has structure  $(\text{CD}_3)_2\text{SO}$ , forms the most widely used solvent in NMR experiments.

To avoid straying away from the point, we leave the readers who are interested in the principal of NMR to NMR textbooks for example (Keeler, 2005). Instead we give readers a simplistic cognition of NMR by

showing pictures of a modern NMR spectrometer, an NMR sample, and 2 dimensional structure of DMSO in Fig 3<sup>8</sup>.

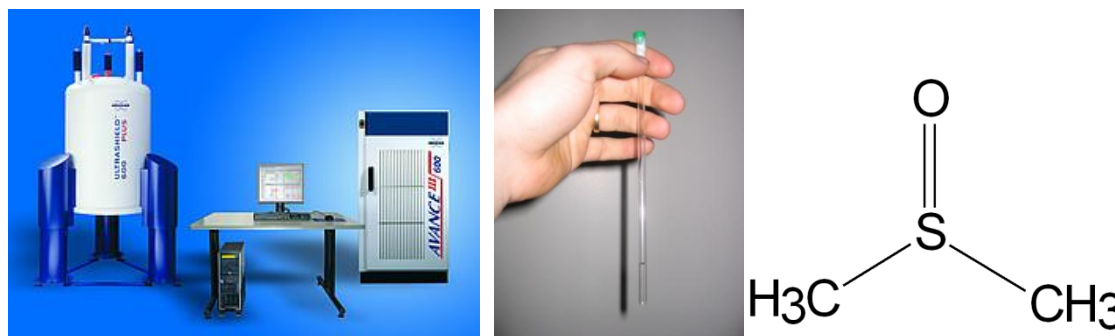


Fig 3 A NMR spectrometer, a NMR sample and Structure of DMSO

### 2.1.2 Basic NMR/Chemical Concepts used in Molecule Structure 1D <sup>1</sup>H NMR Verification.

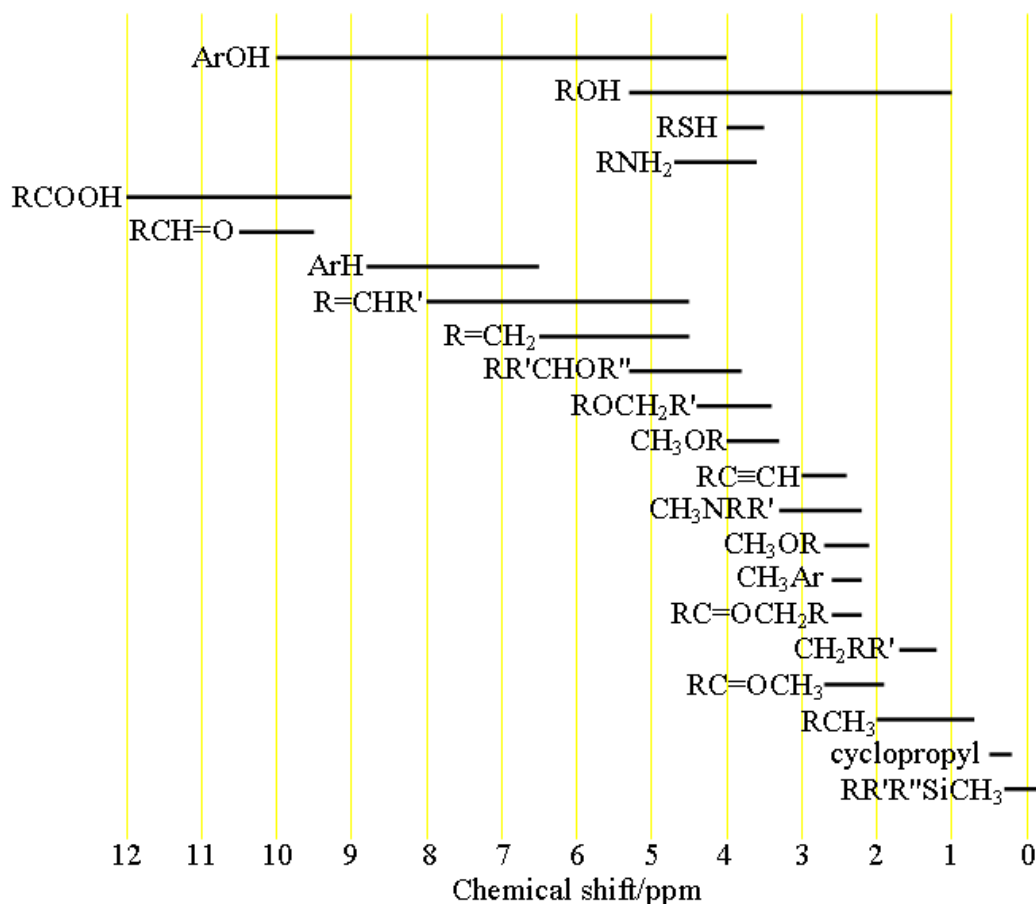
Several information in the 1D <sup>1</sup>H NMR spectrum is used to characterize the structure of an organic compound. They are chemical shift (in the range +12 to -4ppm), integration curve, J-coupling, coupling constant, connectivity, second order coupling, magnetic inequivalence, etc<sup>9</sup>.

#### 2.1.2.1 Chemical shift

Under an external magnetic field, depending on the local chemical environment, different protons in a molecule resonate at slightly different frequencies. Since both this frequency shift and the fundamental resonant frequency are directly proportional to the strength of the magnetic field, the shift is converted into a field-independent dimensionless value known as the chemical shift. The chemical shift is reported as a relative measure from some reference resonance frequency (e.g. the hydrogen-1 of tetramethylsilane (TMS) is commonly used as a reference). This difference between the frequency of the signal and the frequency of the reference is divided by frequency of the reference signal to give the chemical shift. The frequency shifts are extremely small in comparison to the fundamental NMR frequency. A typical frequency shift might be 100 Hz, compared to a fundamental NMR frequency of 400 MHz, so the chemical shift is generally expressed in parts per million (ppm) (Keeler, 2005).

<sup>8</sup> The image of the NMR spectrometer is sourced from the official website of Bruker Cooperation at <http://www.bruker-biospin.com/avanceiii.html>. The image of the NMR sample is sourced from [http://www.absoluteastronomy.com/topics/NMR\\_spectroscopy](http://www.absoluteastronomy.com/topics/NMR_spectroscopy).

<sup>9</sup> Some more subtle information in 1D <sup>1</sup>H NMR spectra could help in identifying molecular structures. However, the usage of this information is often diversified among NMR spectroscopists. This makes it difficult to model these usages in cyberspace. Therefore we skip the introduction of this information in the thesis.



**Table 1 Typical Chemical Shift Ranges for Various Functional Groups**

Through understanding different chemical environments, the chemical shift can be used to obtain some structural information about the molecule in a sample. Specific to the structural verification task, different chemical environments in the molecule are usually organized as chemically equivalent functional groups<sup>10</sup>, while the protons in the same functional group have the same chemical shift. And different functional groups often produce NMR signals at different chemical shift ranges. This physical phenomenon supplies NMR spectroscopists with important evidence to assign protons in a molecule to its spectrum. For example, for the 1D <sup>1</sup>H NMR spectrum of ethanol (CH<sub>3</sub>CH<sub>2</sub>OH), one would expect three specific signals at three specific chemical shift ranges: one for the CH<sub>3</sub> group, one for the CH<sub>2</sub> group and one for the OH group. A typical CH<sub>3</sub> group has a shift range around 0.8-2ppm, a CH<sub>2</sub> attached to an OH has a shift range around 3.5-4.5ppm, and an OH has a wide shift range around 4-10ppm depending on the solvent used (see Fig 4). For assigning protons of a molecule to the spectrum, spectroscopists normally use a chemical shift table to identify chemical shift ranges of typical chemically functional groups. Table 1<sup>11</sup> gives an example of such tables.

<sup>10</sup> In the scope of the thesis, without special annotation, the term “chemically equivalent functional group” is shortened as “functional group”.

<sup>11</sup> The image of Table 1 is sourced from the NMR tutorial of the NMR lab webpage of the Institution of Chemistry in Hebrew University at <http://chem.ch.huji.ac.il/nmr/techniques/1d/row1/h.html>.

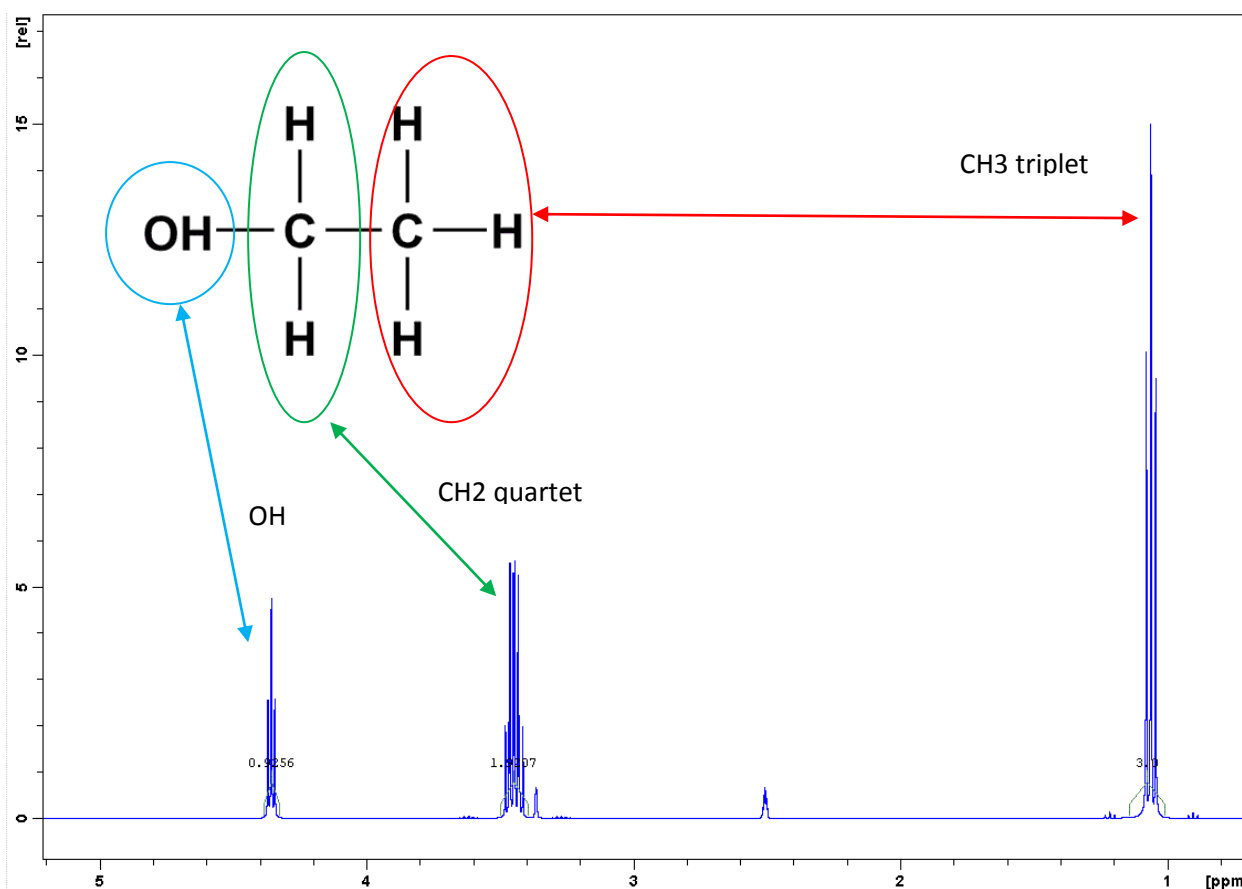


Fig 4 1D  $^1\text{H}$  NMR spectrum and Molecule Structure of Ethanol

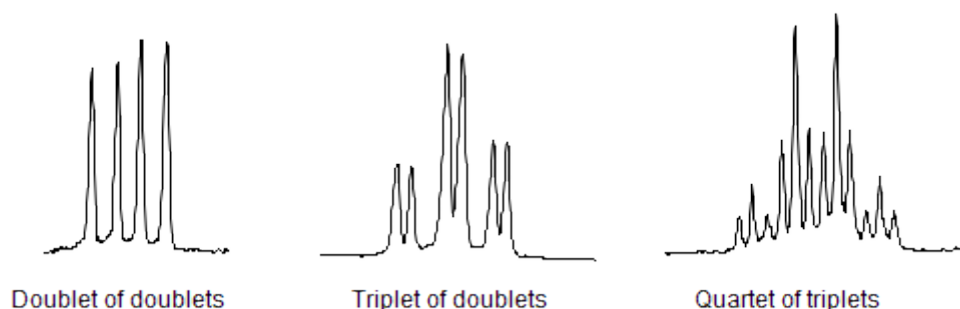
#### 2.1.2.2 Integration

Beside the chemical shift, sizes of NMR signals are indicators of the chemical structure too. In fact, the size of the NMR signal represents the quantity of protons belonging to a certain functional group. In other words, sizes of NMR signals are proportional to the number of protons in the functional groups. For example, in the proton spectrum of ethanol ( $\text{CH}_3\text{CH}_2\text{OH}$ ), the signals from  $\text{CH}_3$  group would be three times as large as the signals from  $\text{OH}$  group since  $\text{CH}_3$  has 3 protons and  $\text{OH}$  only has one. Similarly, the signals of the  $\text{CH}_2$  group would be twice the size of the signals from  $\text{OH}$  but only  $2/3$  of the size of the signals from the  $\text{CH}_3$ . To simplify human interpretation, modern NMR analysis software allows analysis of the size of NMR signals to understand how many protons give rise to a given signal. This is known as integration – a mathematical process which calculates the area under a curve. Note, though calculation of integration is done automatically, identification of individual NMR signals from the spectrum is left for human interpretation. Another note is that analysts determine the size of an NMR signal by integrating the signal instead of measuring its height in amplitude. This is due to that the signal's size depends both on its height and its width, and therefore can only be accurately measured by integrating the whole signal.





groups, one contains 2 protons and the other contains 3 protons, will lead to a quarter of triplets (qt), etc (see Fig 7<sup>13</sup> for the multiplet patterns).



**Fig 7 First Order Multiplet Pattern (b)**

Further rules for identifying J-coupling of a molecule include that couplings between protons in the same functional group have no effects on NMR spectra, couplings between protons that are distant (usually more than 3 chemical bonds apart in molecules) are usually too small to cause observable splitting, long-range couplings over more than three chemical bonds can often be observed in cyclic and aromatic compounds, leading to more complex splitting patterns, etc. For more rules about J-couplings, we refer interested readers to (Keeler, 2005).

To give an example, in the NMR spectrum for ethanol described in Fig 4, the CH<sub>3</sub> group is split into a triplet with an intensity ratio of 1:2:1 by the two protons in neighboring CH<sub>2</sub> group. Similarly, the CH<sub>2</sub> is split into a quartet with an intensity ratio of 1:3:3:1 by the three protons in CH<sub>3</sub> group. In addition, the two CH<sub>2</sub> protons are also neighbored to the proton in OH group, and are split again into a doublet to form a doublet of quartets (bq) (see Fig 8 for the multiplet patterns in the NMR spectrum of ethanol). Note: it often happens that intermolecular exchange of the acidic hydroxyl proton (e.g. protons in OH) results in a loss of this coupling information.

<sup>13</sup> The image in Fig 8 is sourced from Wikipedia at [http://en.wikipedia.org/wiki/Proton\\_NMR](http://en.wikipedia.org/wiki/Proton_NMR).

## 2.1.2.4 Coupling Constant and Connectivity

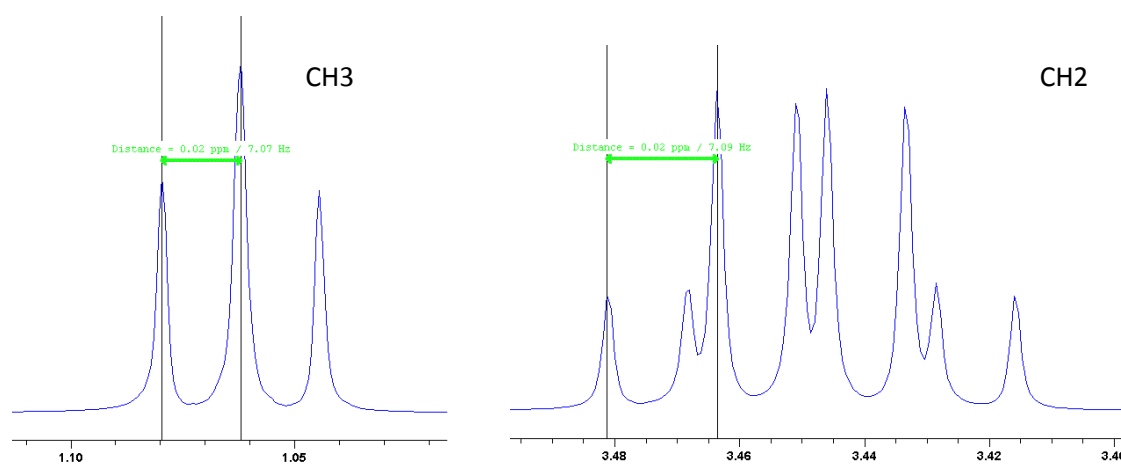


Fig 8 Coupling Constant and Connectivity of Ethanol

Structure	J(Hz)
	6-8
	11-18
	6-15
	4-10
	6-10
	8-11
	a,a: 8 - 14 a,e: 0 - 7 e,e: 0 - 5
	cis: 6 - 12 trans: 4 - 8
	5-7

Table 2 Coupling Constant Ranges for Various Functional Groups in Common Use

The distance between peaks in a multiplet is termed coupling constant, identified as J. The magnitude level of the coupling constant is determined by structures of two functional groups, which interact to produce the J-coupling, and can be predicted (see (Keeler, 2005) for detail). Table 2<sup>14</sup> lists the expected coupling constant ranges for some given structural conformations. Note:

<sup>14</sup>Table 2 is sourced from the NMR tutorial of Department of Chemistry in Central Connecticut State University at <http://www.chemistry.ccsu.edu/glagovich/teaching/316/nmr/coupling.html>.

coupling constants are measured in Hz. This is calculated in the following way: take the distance (in ppm) between two adjacent split peaks in a multiplet, then convert the distance in ppm to Hz by multiplying the distance (in ppm) with the external magnetic field intensity (in MHz). For example, both the multiplet of CH<sub>3</sub> group and the multiplet of CH<sub>2</sub> group in the NMR spectrum of ethanol are measured to have the coupling constant of 7.09 Hz (see Fig 8). In addition, multiplets of protons that split each other will always have the same coupling constant, e.g. the coupling constant of the CH<sub>3</sub> multiplet and one of the coupling constants of CH<sub>2</sub> multiplet in Ethanol are equivalent. This is useful information in determining which multiplets are related to each other in terms of adjacency. In the example of Ethanol, CH<sub>3</sub> group and CH<sub>2</sub> group is determined to be adjacent to each other in the structure by utilizing the equivalency of their coupling constants in the NMR spectrum. Formally this rule about the equivalent coupling constants is named connectivity.

### 2.1.2.5 Second-order Coupling

The description of J-coupling assumes that the coupling constant is small in comparison to the difference in NMR frequencies between different functional groups. If the shift separation decreases (or the coupling strength increases), the multiplet intensity patterns are distorted, and become more complex and less easily analyzed (especially if more than two functional groups are involved). Intensification of some peaks in a multiplet is achieved at the expense of the remainder, which sometimes almost disappear in the background noise, although the integrated area under the peaks remains constant. In most high-field NMR, however, the distortions are usually modest and the characteristic distortions (*roof-top effect*) can in fact help to identify related peaks. For example, the 1D <sup>1</sup>H NMR spectrum in Fig 9<sup>15</sup> illustrates an example of second order couplings among three multiplets. The peak intensities across multiplets A and B are different, that is, the peak on the right side of the multiplet is higher in intensity than the peak on the left side. The purple arrow illustrates a tilt towards the right side for both multiplets. Multiplet C shows an opposite tilt, i.e. to the left side. Multiplets that tilt to form a roof are most likely related protons, and thus are in proximity of each other. Therefore, one can say that multiplets A and B are coupled to multiplet C.

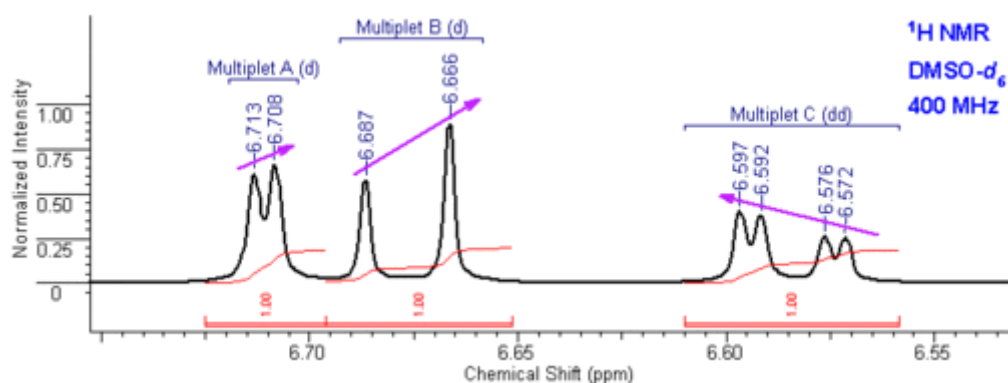


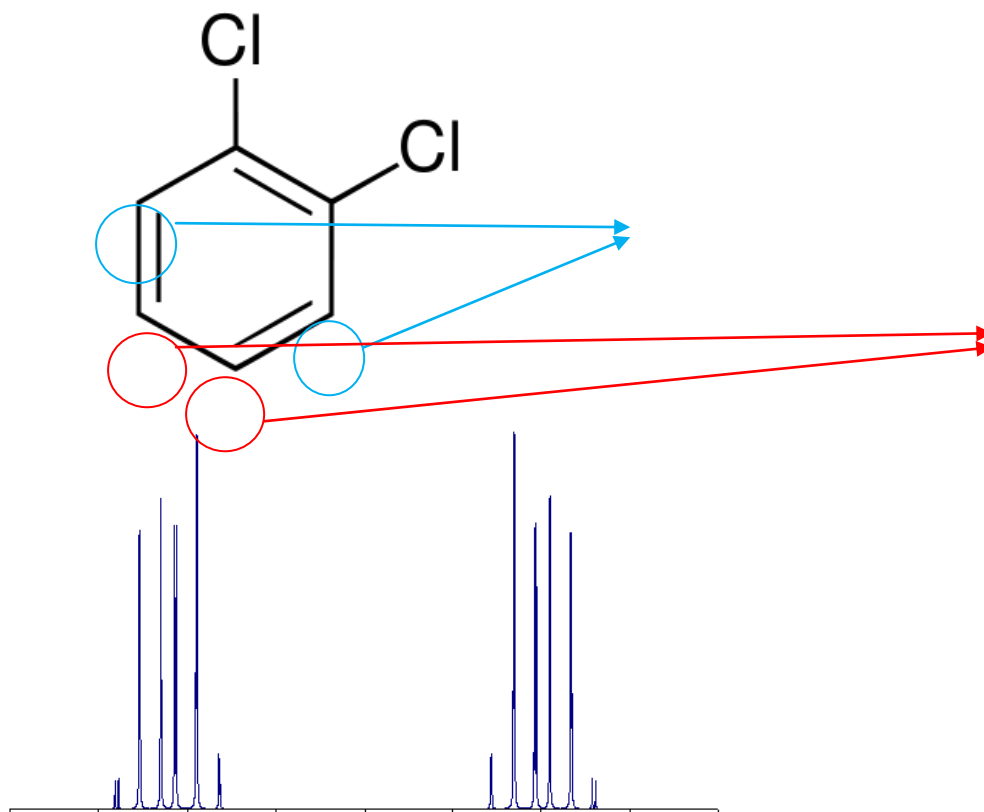
Fig 9 Roof Top Effect

<sup>1515</sup> The image in Fig 9 is sourced from the official website of Advanced Chemistry Development (ACD) Labs at [http://acdlabs.typepad.com/photos/uncategorized/2008/03/24/strongcoupling\\_nmr.gif](http://acdlabs.typepad.com/photos/uncategorized/2008/03/24/strongcoupling_nmr.gif).

Second-order effects decrease as the frequency difference between multiplets increases, so that high-field (high-frequency) NMR spectra display less distortion than lower frequency spectra. Low field spectra at 300 MHz are more prone to distortion than spectra from high field machines, typically operating at frequencies at 500 MHz or above.

#### 2.1.2.6 Magnetic Inequivalence

More subtle effects can occur if chemically equivalent protons (i.e. protons related by geometric symmetry or belonging to the same functional group) have different coupling relationships to external protons. Protons that are chemically equivalent but are not indistinguishable (based on their coupling relationships) are termed magnetically inequivalent. For example, in Fig 10 the 4 protons of 1,2-dichlorobenzene are divided into two chemically equivalent pairs by symmetry (while one group is marked as blue and the other group is marked as red,) and this should produce two triplets in the spectrum. However, magnetic inequivalence causes an individual member of one of the pairs having different couplings to the protons of the other pair, which cause an additional splitting of their signals and so as to produce more complex patterns (see Fig 10). Magnetic inequivalence often leads to highly complex spectra which cannot be analyzed effectively by human spectroscopists. Such effects are more common in 1D  $^1\text{H}$  NMR spectra of aromatic and other non-flexible molecules, while conformational averaging about C-C bonds in flexible molecules tends to equalize the couplings between protons on adjacent carbons, which reduce problems with magnetic inequivalence (see (Keeler, 2005) for more information).



**Fig 10** Molecule Structure and Magnetic Inequivalent Multiplet Pattern in the 1D  $^1\text{H}$  NMR spectrum of 1,2-dichlorobenzene

### 2.1.3 Human Process of Molecular Structure 1D $^1\text{H}$ NMR Verification- an Example

In 2.1.2 we briefly introduce the NMR knowledge that spectroscopists utilize for 1D  $^1\text{H}$  NMR structure verification. To make it easier to understand the human logic behind their structure verification procedure, in this section, we use a real compound and its 1D  $^1\text{H}$  NMR spectrum as an example to describe the spectroscopist's interpretation process. Specifically, we use +-Pseudoephedrin as our example. See Fig 11 for its 2D molecule structure and 1D  $^1\text{H}$  NMR spectrum.

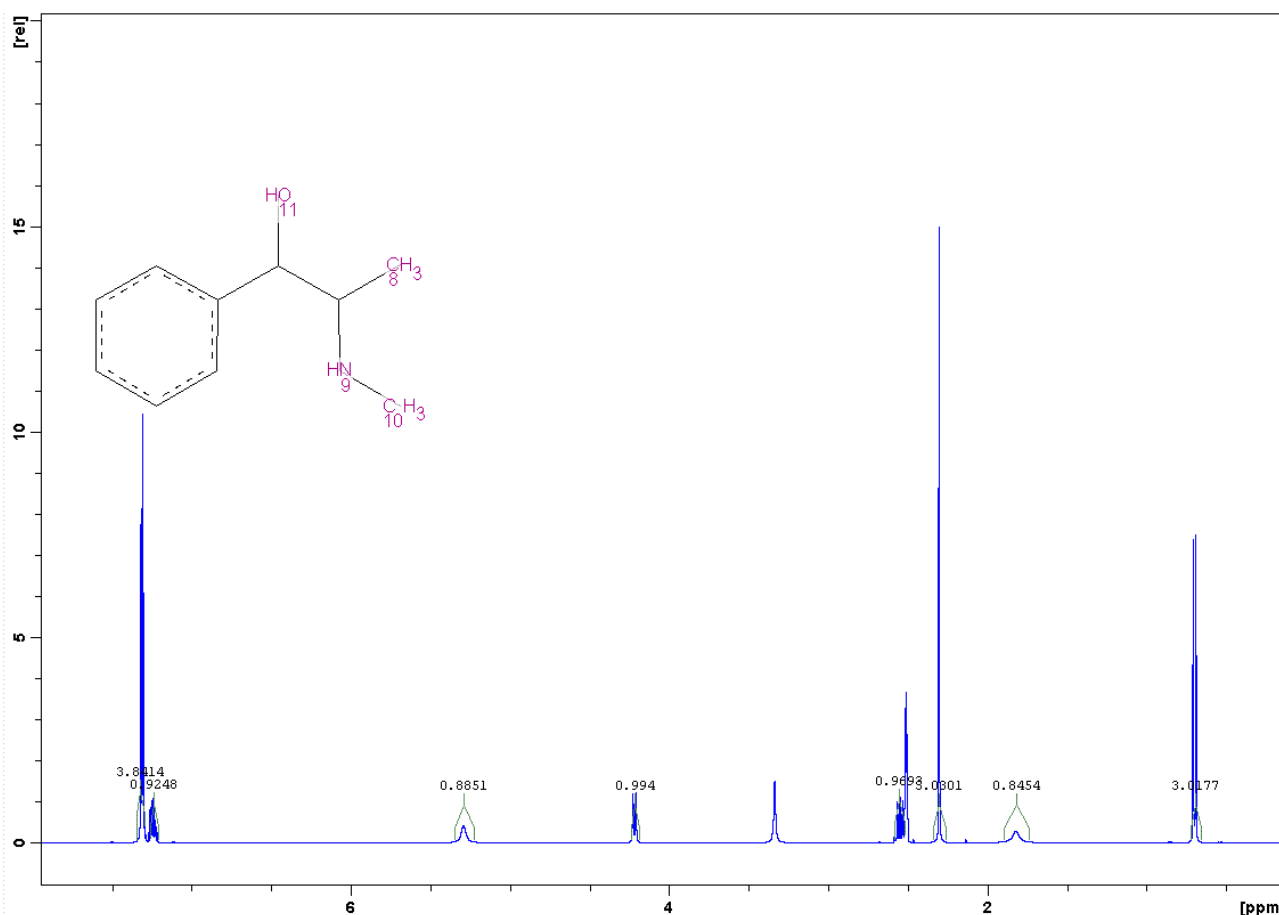


Fig 11 Molecule Structure and 1D  $^1\text{H}$  NMR Spectrum of +-Pseudoephedrin

Generally speaking, human structure verification processes is cursory and empirical. Different spectroscopists may adopt slightly different approaches to check consistency between the molecular structure and the 1D  $^1\text{H}$  NMR spectrum. However, the core methodology of the NMR structure verification among different spectroscopists is the same. We can roughly divide the process into 5 steps: (a) Identification of peak clusters from the spectrum. (b) Identification of solvents from the peak clusters. (c) Computing proton numbers for the peak clusters. (d) Verification of consistency

between the molecular structure and the peak clusters with proton number. (e) Further checking for consistency between the molecular structure and the peak clusters by coupling analysis. In the following, we explain each step in detail with our example.

### 2.1.3.1 Identifying Peak Clusters

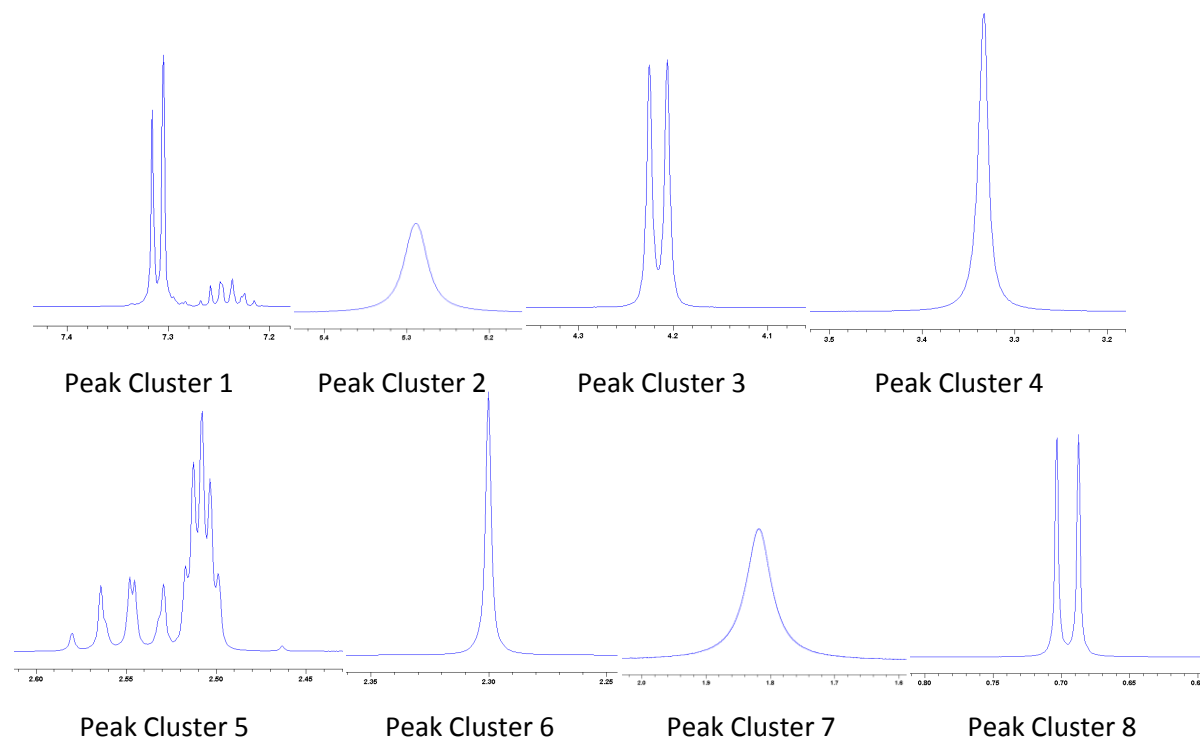
Spectroscopists use the following approach to identify peak clusters:

Starting from left to right (high field to low field) of the spectrum:

1. Selection of a point on the x-axis which has an amplitude around the spectrum baseline, and using this point as the left boundary of a new peak cluster.
2. Starting from the point we move continually to the right, so that the region covers as many peaks as possible.
3. Stopping when the movement touches another point on the x-axis which has an amplitude around the spectrum baseline. This new point is used as the right boundary of the peak cluster. As a result, a new peak cluster is identified.

Repeating this process until all peak clusters are identified.

Specifically, in our example 8 peak clusters are identified from the spectrum of +-Pseudoephedrin with the above approach. We list their peak patterns in Fig 12.



**Fig 12 Peak Clusters Identified from 1D  $^1\text{H}$  NMR Spectrum of +-Pseudoephedrin**

### 2.1.3.2 Identifying Solvent

As we explained in 2.1.1, several solvents could be used in  $^1\text{H}$  NMR experiments, wherein different solvents have different chemical /physical characteristics. Correspondingly, there are different multiplet patterns appearing in 1D  $^1\text{H}$  NMR spectra, which require different pattern recognition techniques. Considering the similarity of techniques among different solvent pattern identifications, and to simplify the problem setup, we decide to limit the solvent detection problem to the identification of one popular solvent used in 1D  $^1\text{H}$  NMR experiments under practical industrial environment. Through the survey among the customer of our industrial partner – a NMR manufacture, we understand that DMSO is the most widely used solvent for 1D  $^1\text{H}$  NMR experiments in practice. Therefore, we limit the problem of solvent identification to the problem of identifying DMSO in the scope of the thesis.

DMSO is the compound to use deuterium to replace hydrogen in dimethyl sulfoxide. Since deuterium does not produce a signal in 1D  $^1\text{H}$  NMR experiment, principally DMSO would not produce signals in 1D  $^1\text{H}$  NMR spectra, and therefore would not interfere with the signals generated from the target compound. However, in practice deuteration is never complete (“100%”). Therefore the signals from the residual protons of DMSO are still observable in the 1D  $^1\text{H}$  NMR spectra. To further clarify above explanations, we list the deuteration degree of DMSO, which are commonly commercially available in Table 3<sup>16</sup>.

Degree of deuteration %	99	99.5	99.8	99.95
Remaining concentration of protons [mol/l]	0.1-0.06	0.05-0.03	0.02-0.01	0.006-0.003
Advisable concentration of substance [mol/l]	0.1	0.05	0.02	0.005

**Table 3 Deuteration Degree of DMSO**

Another practical issue of DMSO (or any other solvents) is that it is never absolutely dry. There are always some amounts of  $\text{H}_2\text{O}$  existent in the DMSO samples. The protons of  $\text{H}_2\text{O}$  produce a NMR signal as well in the 1D  $^1\text{H}$  NMR spectra. As a result, identifying a DMSO signal in the 1D  $^1\text{H}$  NMR spectra is defined as identifying NMR signals of both the residual protons of DMSO and the protons of  $\text{H}_2\text{O}$  in the DMSO sample.

The signal of the residual protons of DMSO is easily identified in the 1D  $^1\text{H}$  NMR spectra. It often shows up at the fixed chemical shift location - around 2.5ppm. And it often appears as a fixed multiplet pattern – a Pentet or a Doublet of Triplet. The size of the signal depends on the deuterated degree of the DMSO sample. For a highly deuterated DMSO, the signal could be dramatically small so that the observable multiplet pattern could be degenerated to a Triplet. This is due to the

<sup>16</sup> Table 3 is sourced from the NMR tutorial from Department of Chemistry and Biochemistry in New Mexico State University at [http://www.chemistry.nmsu.edu/Instrumentation/NMR\\_Solv.html](http://www.chemistry.nmsu.edu/Instrumentation/NMR_Solv.html).



absence of enough residual protons in DMSO so that the signal is too small to be visible (especially at the edges of the multiplet pattern).

Comparatively, the signal of H<sub>2</sub>O in DMSO is relatively difficult to be identified. This is due to the diversity of the H<sub>2</sub>O signal in both its chemical shift location and multiplet shape pattern. Despite these complexities, human spectroscopists could still reliably identify a H<sub>2</sub>O signal relying on some primitive empirical rules. Some common rules are<sup>17</sup>:

- (1) The signal of H<sub>2</sub>O could appear in a large chemical shift range between 3.0ppm and 4.9ppm.
- (2) The signal of H<sub>2</sub>O often shows as a broad shape.
- (3) The signal of H<sub>2</sub>O often shows as a single peak without splitting. But it could also show occasionally as 2 or more peaks. When more than a single peak are generated from H<sub>2</sub>O, all these peaks are overlapped together and are not well separated. In addition, these peaks are asymmetric in both peak amplitudes and peak positions.
- (4) The signal of H<sub>2</sub>O is not accompanied with satellite peaks.

To utilize above rules for solvent detection in our example, we return back to Fig 12 to scan all peak clusters we have identified from the spectrum. It is crystal-clear that Peak Cluster 5 is the signal from residual protons of DMSO. This is because Peak Cluster 5 is the only peak cluster which uniquely contains a Pentet and is located at 2.5ppm. However, from the picture it clearly shows that there is another multiplet, possibly from the target compound, which is overlapped with the residual proton signal of DMSO. This overlapping of different NMR signals increases the complexity of DMSO identification. Despite this distortion, there are still enough evidences for spectroscopists to discriminate DMSO signals from others. This owes to the stability of residual proton signal of DMSO as we mentioned before.

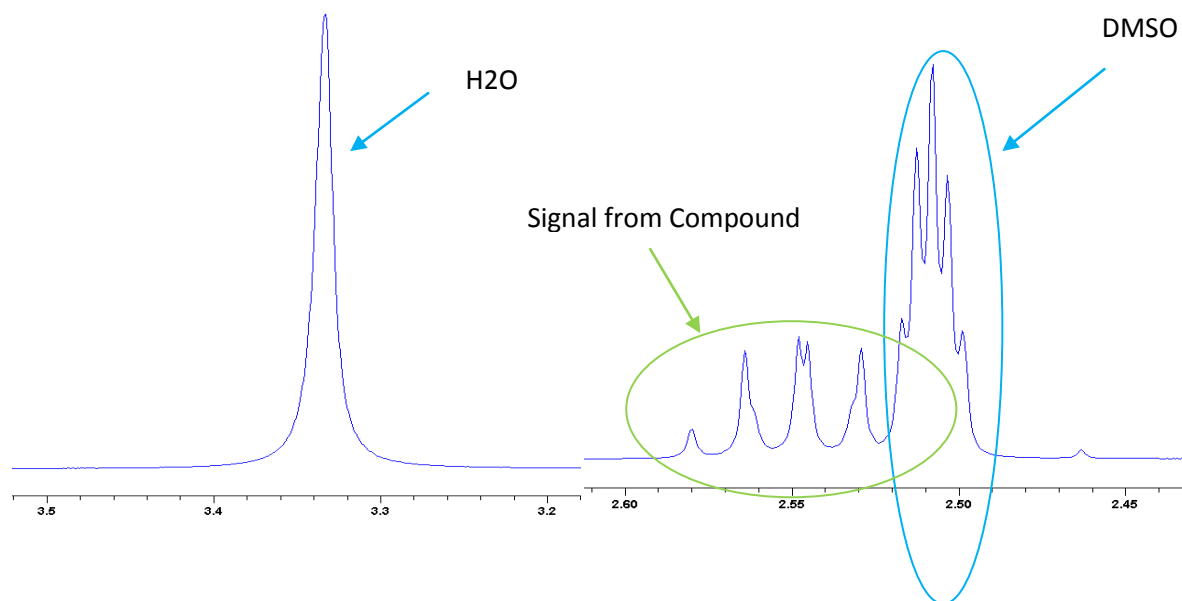
The identification of a H<sub>2</sub>O signal is tricky, since there is no fixed shape pattern for the H<sub>2</sub>O signal, and its chemical shift could be in a wide range. Based on the rough rules described above, spectroscopists would “guess” that Peak Cluster 4 is a good candidate for a H<sub>2</sub>O signal. First, Peak Cluster 4 has chemical shift 3.33ppm, which is a likely chemical shift position for H<sub>2</sub>O signal (note, though H<sub>2</sub>O signal’s chemical shift range is 3.0-4.9ppm, spectroscopists’ subjective probabilistic density distribution over the H<sub>2</sub>O chemical shift range is non-uniform. Their subjective probability density of H<sub>2</sub>O signal to be shown in 3.33ppm is higher than that of other positions in the H<sub>2</sub>O signal’s chemical shift range). Second, Peak Cluster 4 appears as a single wide peak, which gives another evidence to show its aptness of the signal from H<sub>2</sub>O. There is another peak cluster- Peak Cluster 2 in the H<sub>2</sub>O signal’s chemical shift range. However, Peak Cluster 2 has two well-separated peaks to form a nice doublet, which violates the third rule mentioned above, and makes it unlikely to be a signal of H<sub>2</sub>O. With this further evidence, spectroscopists confirm that Peak Cluster 4 is the signal from H<sub>2</sub>O. With both DMSO and H<sub>2</sub>O signals uniquely identified, the task of identification of solvent signals in the 1D <sup>1</sup>H NMR spectrum of +-Pseudoephedrin is finished. For clarity we show the identified patterns in Fig 13.

Note, although the signals of both H<sub>2</sub>O and residual protons in DMSO are uniquely identified in our example, generally it is not the case, especially in the task of the H<sub>2</sub>O signal identification. In case there are more than one peak clusters “suitable” to be the signal from the solvent, spectroscopists often adopt a hypotheses-driven problem solving strategy to loop through all possible solvent candidates. Specifically, they would assume the most “likely” peak cluster as the signal from solvent, and jump to the next step of structure verification. If it is proven in a later step that the assumption

---

<sup>17</sup> Note, there are additional rules which can be used to identify H<sub>2</sub>O signals from NMR spectra. However, their usage is not unified among NMR spectroscopists, and is relied on the experience of the spectroscopists. Therefore, we skip their introduction in the thesis.

is wrong, spectroscopists would return back and choose the second most “likely” peak cluster as the solvent signal. The process would be repeated many times until the correct solvent signal is identified or spectroscopists decide to give up.



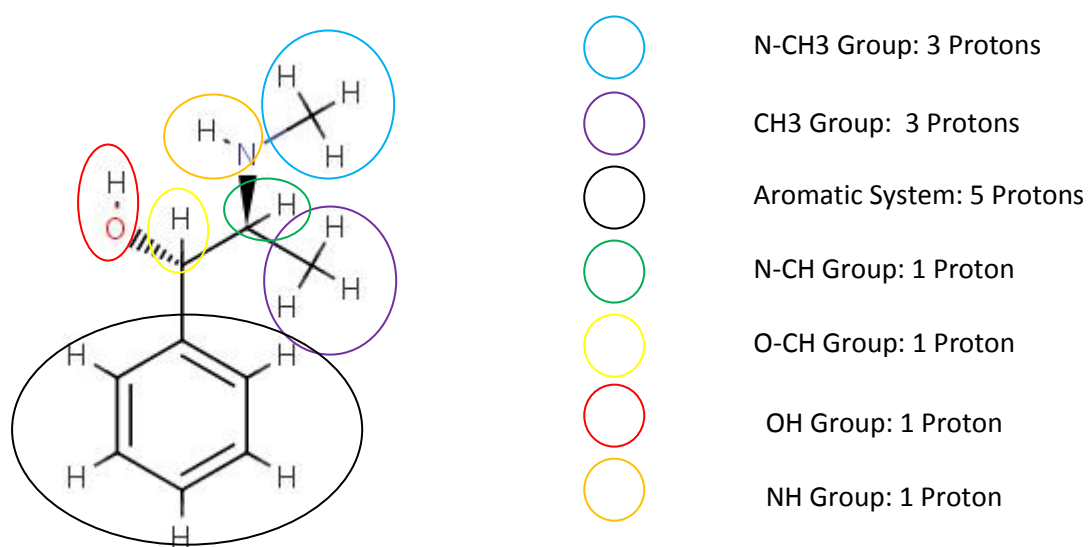
**Fig 13 H2O and DMSO Patterns in 1D <sup>1</sup>H NMR Spectrum of +-Pseudoephedrin**

### 2.1.3.3 Computing Proton Numbers of Peak Clusters

To compute the proton number of the peak cluster (excluding peak cluster representing the signals from solvent), spectroscopists first integrate peak clusters. In digital NMR spectra, this is done by adding amplitudes of all signal points, which all belong to a peak cluster. Note, the spectrum integration functionality is supported by all commercial NMR acquisition and application software.

As we explained in 2.1.2.2, spectroscopists do integrations for peak clusters in order to match them in proton numbers with the functional groups extracted from the molecular structure. To do so, first spectroscopists need to normalize the integrations of peak clusters to the unit of proton numbers so that they become comparable to the proton numbers acquired from molecular functional groups. To do the normalization in a succinct way, spectroscopists recur to the information from the molecular structure to seek a peak cluster, which could be uniquely assigned to a functional group from the molecule, as the normalization reference. Here, it requires spectroscopists to identify all functional groups from the given molecular structure and compute the proton numbers under the functional

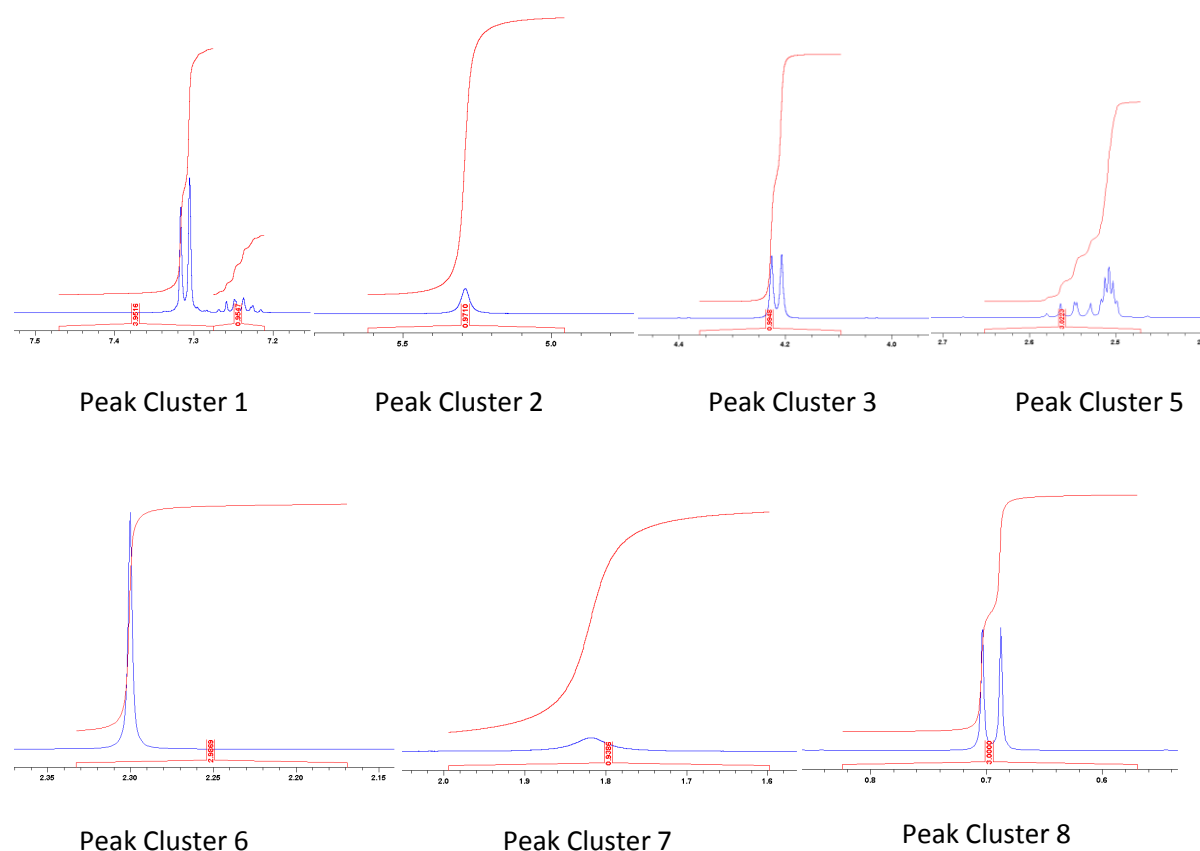
group. Identification of functional groups from molecular structure is a straightforward task for an organic chemist. However, the explanation of how this is done requires knowledge in organic chemistry, which is beyond the scope of this thesis, therefore we leave interested readers to classical organic chemistry textbooks e.g. (Solomons, et al., 2003) for a detailed explanation. With a chemical functional group identified, then the proton number of the functional group could be computed by counting the number of protons in the functional group. For clarity, we illustrate the identified functional groups and their proton numbers of +-Pseudoephedrin – our example in Fig 14.



**Fig 14 Identify Functional Group and Proton Numbers from +-Pseudoephedrin**

After all functional groups are identified, in the next stage spectroscopists seek a peak cluster which can uniquely match to one of the functional groups. This task requires utilizing the knowledge in chemical shift match, as well as coupling analysis. Specifically, in the example of +-Pseudoephedrin, spectroscopists would judge that CH<sub>3</sub> group and Peak Cluster 8 are uniquely matched to each other. The reason of this assignment is that CH<sub>3</sub> group has a typical chemical shift range of 0.8-2ppm, where only Peak Cluster 7 and Peak Cluster 8 can match. Further analysis shows that Peak Cluster 7 is a wide single peak, which is unlikely to be the signal from CH<sub>3</sub> group. This is due to spectroscopists' experience that CH<sub>3</sub> group usually produces sharp (narrow) peaks. This supplies the first evidence why Peak Cluster 8 and CH<sub>3</sub> group are uniquely matched. In addition, the CH group which is three bonds away from CH<sub>3</sub> group would cause the splitting of signal of CH<sub>3</sub> group to a doublet, which is matched to the multiplet pattern of Peak Cluster 8. This supplies the second evidence about the unique matching between Peak Cluster 8 and CH<sub>3</sub> group. With both evidences in hand, spectroscopists can confirm this unique matching – CH<sub>3</sub> group and Peak Cluster 8.

Next, the previously selected uniquely matched peak cluster-functional group pair is used to compute the normalization factor. This is done by dividing the integration of the peak cluster by the proton number of the functional group. Note, the unit of the normalization factor is integration per proton. With the normalization factor computed, the proton number of each peak cluster can be computed by dividing its integration by the normalization factor. As the result, all peak clusters are assigned an estimated proton number. Returning to our example, referring to the above mentioned methodology, the normalization factor is computed by dividing the integration of Peak Cluster 8 by 3 proton number of CH<sub>3</sub> group. This is followed by dividing the integration of each peak cluster by the normalization factor to get the proton number of the peak cluster. Note, all commercially available NMR acquisition software support the normalization of integrations of peak clusters to proton numbers, but require human intervention to select the reference pair (between a peak cluster and a functional group). In Fig 15 we show the calculated proton numbers for 7 peak clusters extracted from the NMR spectrum of +-Pseudoephedrine with the Topspin software, a NMR application software from Bruker Biospin AG. In addition, the signal in Peak Cluster 5 is the overlap of the signal from a functional group and the signal from DMSO residual protons, which makes its integration and therefore its computed proton numbers unreliable. Therefore, spectroscopists choose not to compute its integration and proton number, and to further use them for structure verification.



**Fig 15 Proton Numbers of Peak Clusters of +-Pseudoephedrin**

From the above analysis, we know that the correctness of computed proton numbers of peak clusters relies on the successful discovery of a unique matching pair between a peak cluster and a functional group. However, this unique matching does not generally exist, which is unfortunately not shown in our example. If this happens (the unique matching is not existent), spectroscopists adopt their hypothesis-driven problem solving strategy to iteratively compute the normalization factor by sequentially selecting the most “likely” peak cluster and functional group pairs. Specifically, spectroscopists would choose the most possible pair to compute the normalization factor, and this is followed by computing proton numbers for peak clusters. Obviously, the resulting computed proton numbers are possibly wrong. However, spectroscopists assume that the computed proton numbers are correct, and insist to go to the next step to further verify the consistency between the spectrum and the molecular structure with the computed proton numbers. With the correctness of the computed proton number unwarranted, contradiction could happen during this next step. As a result, if it happened, it would motivate spectroscopists to return back to re-compute the normalization factor with the next most likely pair. Spectroscopists often go through this iteration many times until there is no contradiction found in the next step, while reasonable assignments between peak clusters and functional groups are implemented, or after enough iterations without finding an uncontradictable explanation they decide to give up. Here, “give up” means that spectroscopists can not deduce the consistency between the molecule structure and the NMR spectrum, but they can also not deduce the inconsistency between the molecule structure and the NMR spectrum. Therefore, they arrive at the conclusion that they don’t know whether the molecule structure is consistent with the NMR spectrum or not.

#### 2.1.3.4 Verifying Consistency between the Molecular Structure and the Peak Clusters with Proton Number

With the proton number assigned to each peak cluster, spectroscopists are ready to verify the consistency between the spectrum and the structure with them. However, the spectroscopists’ approach for verification with proton number is rough and cursory. Nevertheless, it’s being proved to be reliable in practice.

Generally speaking, we can divide structure verification with proton number into two steps. In the first step, spectroscopists verify the consistency between the structure and the spectrum by comparing the total proton numbers counted from the spectrum to that from the structure. In the second step, they start detail verification by assigning peak clusters to functional groups with proton numbers and chemical shifts. To make it easier to be understand, we keep using our example to illustrate how these two steps work starting from next paragraph.

In the first step, we need to compare the total proton number in the spectrum to that in the structure. Clearly, by straightly counting on the structure, we summarize that there are 15 protons in +-Pseudoephedrin. On the other side, we add proton numbers of peak clusters, excluding Peak Cluster 4 (since it is the signal purely generated from the solvent), altogether to get 17 protons (see Fig 15). By comparing the total proton numbers – 15 vs. 17, it seems that the total proton numbers are inconsistent. However, through further investigation, we realize that Peak Cluster 5 is the

overlap between the signal of DMSO residual protons and a signal from the molecule, which makes using the integration of whole Peak Cluster 5 to compute the proton number overestimate the real proton number from the molecule. By temporarily ignoring the 3 protons counted from Peak Cluster 5, we get a total of 14 protons from the spectrum. Since there are 15 protons from the molecular structure, we can deduce that only one proton is probably produced by the molecule in Peak Cluster 5. With this perspective, we find a consistent explanation for the total proton numbers. To further confirm our perspective, we split Peak Cluster 5 into two parts, where signals in the left part are mainly from the molecule and signals in the right part are mainly from DMSO residual protons (see Fig 13). By integrating the two parts separately, we see that the integration of the left part is roughly one third of the integration of Peak Cluster 5. This new finding further confirms our assumption that the signal from the molecule in Peak Cluster 5 contains one proton. With this new evidence, spectroscopists are confirmed that the spectrum and the structure are consistent in total proton numbers, and both spectrum and structure contain 15 protons.

In the second step, we need to build assignments among peak clusters and functional groups with their proton numbers and chemical shifts. The reasonable assignments would supply strong evidence to prove the consistency between the molecule and the spectrum, and to prevent possible false-positive alarm. Note, we are going to explain false positive rate (the second type of error) in detail in Chapter 6. Here, we only emphasize the conclusion: **The principle is universally true in any decision problems that more evidences are shown, less risk to produce false positive cases.** Therefore it is necessary and important to do the detail assignments between peak clusters and functional groups.

To clarify the human approach to do assignments, we still utilize our example to illustrate the methodology. Specifically, first we know that there is an aromatic system which contains 5 protons in +-Pseudoephedrin (see Fig 14 and Fig 17). By checking the  $^1\text{H}$  NMR chemical shift (see Table 1), we understand that the aromatic system typically has a chemical shift range of 6.5-9ppm. With this clue, we scan for peak clusters in the chemical shift range of 6.5-9ppm, and check if they contain 5 protons. From Fig 15, we see that only Peak Cluster 1 is in the range, and it happens to contain 5 protons. Therefore, we confirm our first assignment: Peak Cluster1 versus the aromatic system, and they are consistent on both proton numbers and chemical shifts. Similarly, the CH group has a typical chemical shift range of 3-6.5ppm. By checking the peak cluster list, there are Peak Cluster 2 and Peak Cluster 3 in the range, and both of them contain one proton, which makes them consistent to 2 CH groups on both proton numbers and chemical shifts. This seems to give enough evidence of assigning Peak Cluster 2, Peak Cluster 3 to 2 CH groups. However, an empirical rule could be applied here to deny the assignment. Specifically, Peak Cluster 2 is a wide single peak, which makes it impossible to be a signal from one of two CH groups in +-Pseudoephedrin. This is because the signal of CH group appears as sharp peaks instead of broad peaks. This rule will “kick out” the qualification of Peak Cluster 2 to be a candidate for matching CH groups. Another rule could also be applied to deny Peak Cluster 2 for matching CH groups, which relies on coupling analysis. From the structure of +-Pseudoephedrin, both CH groups are coupled to other protons, which cause the peaks from the CH groups splitting and producing a more complex multiplet pattern than a singleton (see 2.1.3.5 for detail coupling analysis). As a result, only Peak Cluster 3 has possible signals from the CH groups. This seems to cause inconsistency, since there is only one proton in Peak Cluster 3, but 2 CH groups, which contain 2 protons. However, by carefully examining the peak cluster list again, we find that

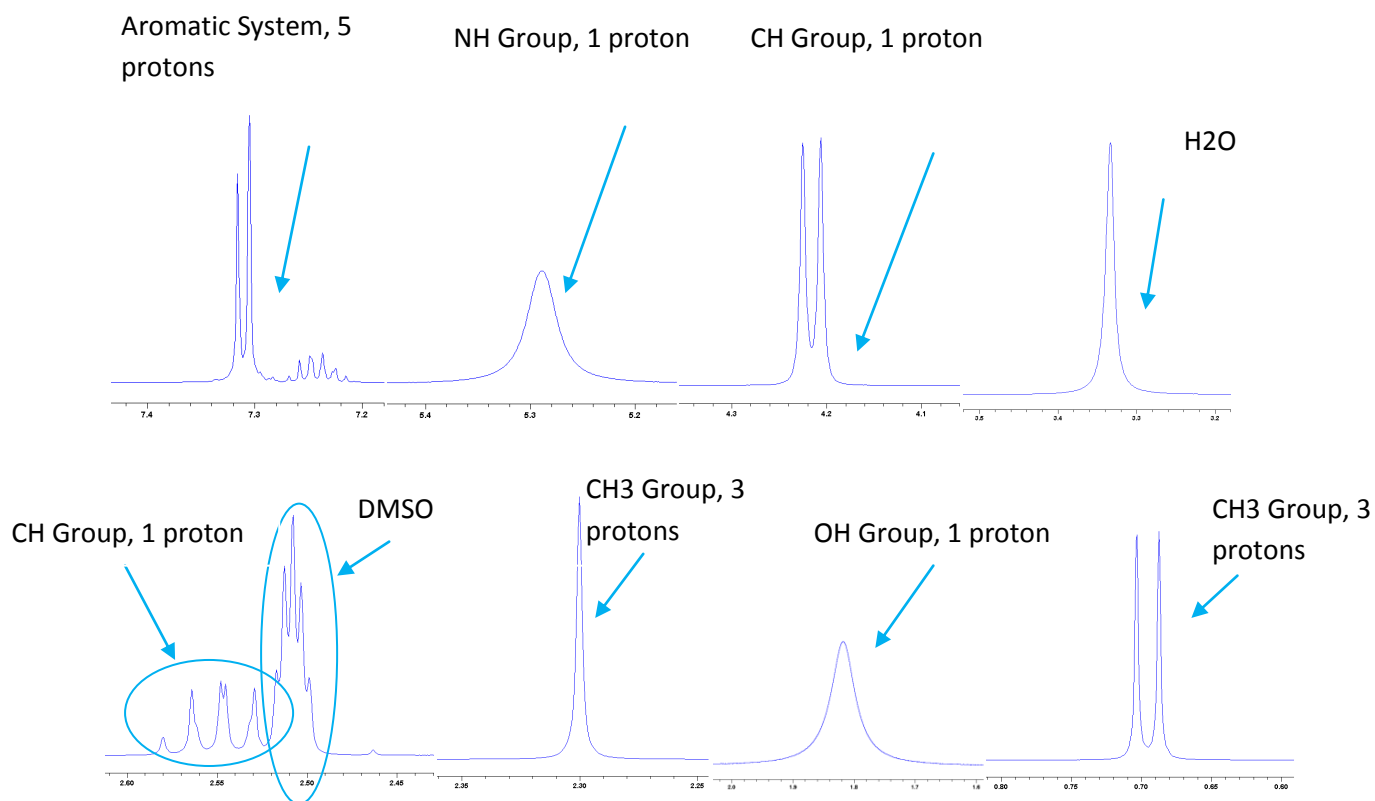
the multiplet in Peak Cluster 5 has a chemical shift of around 2.8ppm, which is close to the CH group chemical shift range and has a complex multiplicity. With continually accumulated NMR experience, we know that chemical shift ranges of functional groups recorded in the shift table are not absolutely accurate, and could slightly shift depending on practical NMR experimental conditions e.g. solvent types, measuring temperature, PH value, etc. Therefore, we flexibly adapt the chemical shift range of the CH group to 2.8-6.5ppm so that we can assign the multiplet in Peak Cluster 5 to CH groups as well. As we show in the previous paragraph, the multiplet in Peak Cluster 5 contains one proton, and it could be the signal from one of the CH group. Thus, totally we have two protons in Peak Cluster 3 and Peak Cluster 5, which is consistent with the 2 protons from 2 CH groups. Though there are still no clear one-to-one mappings between Peak Cluster 3, 5 and two CH groups, it already gives us much more evidence in our structural verification task. Keeping the same strategy, we further do the assignments on CH<sub>3</sub> (methyl) groups. There are two CH<sub>3</sub> groups from +-Pseudoephedrin, which totally count 6 protons. The CH<sub>3</sub> group often has chemical shift range of 0.8-3ppm. Again, from Fig 15, we find that Peak Cluster 6, 7 and 8 is in the range, where only Peak Cluster 6 and Peak Cluster 8 contain 3 protons. Obviously, they are signals from CH<sub>3</sub> groups. In this way, we also create assignments between Peak Cluster 6, 8 and two CH<sub>3</sub> groups, and they are consistent in proton numbers and chemical shifts.

To summarize the assignments we built so far, all peak clusters are assigned to some functional groups except two. They are Peak Cluster 2 and 6. On the other side, there is an OH group and a NH group which are not being assigned yet. With our NMR knowledge, we know that both OH and NH group could exchange proton with the solvent, and therefore we expect that the signals of them could disappear from the 1D <sup>1</sup>H NMR spectrum, or their signal sizes could shrink so that they are disproportional to the proton numbers of their functional groups. Keeping these variables in mind, we would not check consistency of the proton numbers between peak clusters and OH and NH groups precisely. Instead we only check if the proton numbers of the peak cluster are equal or smaller to the proton numbers of NH group or OH group for consistency. In our example, both Peak Cluster 2 and Peak Cluster 6 contain 1 proton, which is equal or smaller than the proton number from OH and NH group, and therefore consistent with OH and NH group in proton numbers. Since the signals of both OH and NH groups could be shown in a very wide chemical shift range, Peak Cluster 2 and 6 are consistent with the OH and NH groups in chemical shift either. Therefore, we assign Peak Cluster 2 and 5 to OH group and NH group. So far, we have built a complete assignment between peak clusters and functional groups. With this the second step of proton number verification on +-Pseudoephedrin is completed, and the conclusion is consistent between the spectrum and the structure of +-Pseudoephedrin.

To supplement the above mentioned analysis, we introduce several (crude) rules for NH and OH group identification so that we can precisely determine one to one assignments between Peak Cluster 2, 6 and NH, OH groups. First, signals of both NH and OH group could appear in wide chemical shift ranges, and in fact their chemical shift ranges are overlapped by those of CH, CH<sub>2</sub>, CH<sub>3</sub> and Aromatic System functional group. Second, signals of NH or OH groups often show up as wide single peaks. Third, the signal of NH group often appears in relatively high field ppm position compared to that of OH group. To apply these rules in our structure verification task on +-Pseudoephedrin, both patterns of the Peak Cluster 2 and 6 are compatible to the characteristics represented in the first and second rules. This supplies us with more evidence to confirm our previous assignments. Furthermore, the third rule gives us “magical baton” to precisely pinpoint two one-to-one

assignments between the two peak clusters and the two function groups. That is: Peak Cluster 2 is matched to NH group, and Peak Cluster 6 is matched to OH group.

To summarize the result, utilizing the two-step proton number verification process, we roughly build assignments between peak clusters and functional groups, and reach the conclusion that the spectrum and structure of +-Pseudoephedrin are consistent on proton number. To clarify the result, we illuminate the assignments we built so far in Fig 16.



**Fig 16 Assignments between Peak Clusters and Functional Groups with Proton Number on +-Pseudoephedrin**

#### 2.1.3.5 Further Verifying the Consistency between Peak Clusters and Function Groups by Coupling Analysis

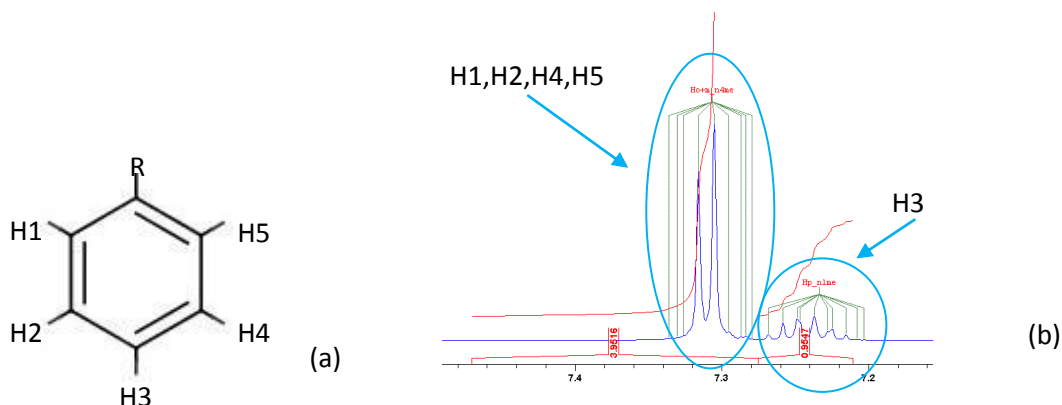
Through consistency analysis with proton numbers in 2.1.3.4, specoscopists could often build rough (non-one-to-one) assignments between peak clusters and functional groups. For example, in our structural verification procedure of +-Pseudoephedrin in 2.1.3.4, it is doubtless that Peak Cluster 1 is uniquely mapped to the aromatic system, Peak Cluster 2 is uniquely mapped to the NH group, Peak Cluster 7 is uniquely mapped to the OH group. In addition, we have already identified that Peak



Cluster 8 is uniquely mapped to the CH<sub>3</sub> group, and the pair was used as the reference to compute the normalization factor in 2.1.3.3. With this additional information, we could directly deduce that another peak cluster – Peak Cluster 6, which is assigned to CH<sub>3</sub> group, is uniquely mapped to the other CH<sub>3</sub> group. For clarity, we name the second CH<sub>3</sub> group as N-CH<sub>3</sub> group (to indicate that it is neighbored to the NH group).

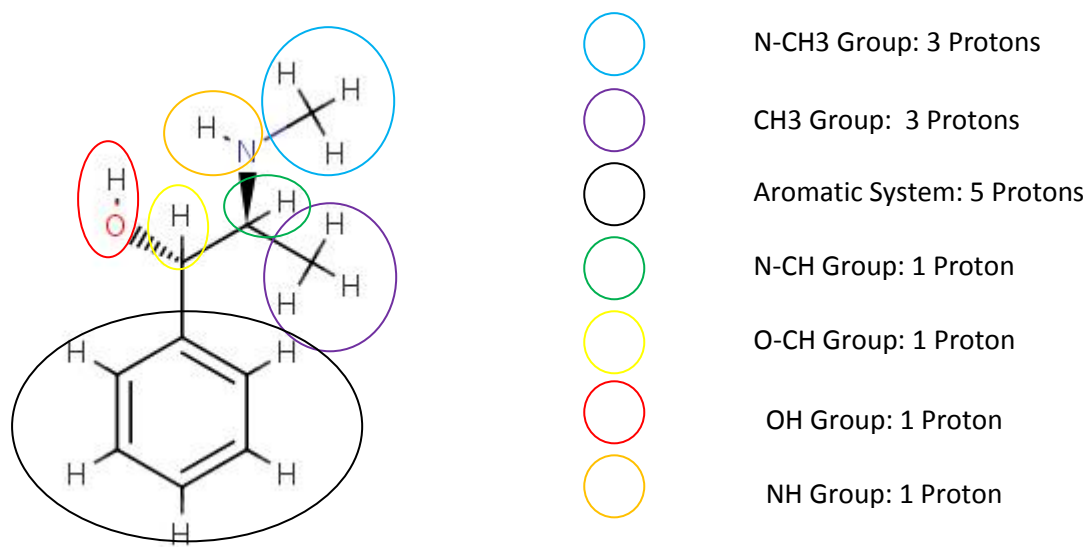
The assignments with proton numbers and chemical shifts are often rough, and not exactly one-to-one. In our example, from 2.1.3.4 we assign Peak Cluster 3 and the multiplet in Peak Cluster 5 to the two CH groups of +-Pseudoephedrin. But, we don't know which peak cluster is assigned to which CH group. To solve the uncertainty there, we rely on coupling analysis. Ultimately, coupling analysis could supply additional information which we could use to further check previous unique assignments we did in 2.1.3.4. By passing through this additional coupling analysis, spectroscopists accumulate more evidence to support their previous decision, and this will further reduce false positive rate. Though there are only two CH groups which are not uniquely assigned in our example, the number of non-one-to-one assigned peak cluster-functional group pairs could be dramatically bigger for structural verifications of other molecules. This situation would particularly happen where the molecular structures are complex. If it happens, spectroscopists turn to mainly rely on coupling analysis to determine delicate assignments between peak clusters and functional groups. To keep it easy to understand, we still utilize the example to illustrate how J-coupling, coupling constant and connectivity are used to do the assignments, with details starting from the next paragraph.

In the first step, we start the coupling analysis from the aromatic system. In the aromatic system of +-Pseudoephedrin, there are 5 protons, which are marked as H1, H2, H3, H4 and H5 (see Fig 17 (a)). By geometrical symmetry, we know that H1 and H5 are chemically equivalent, and H2 and H4 are chemically equivalent. By relying on the knowledge we explained in 2.1.2, we realize that in aromatic systems chemically equivalent protons are magnetically inequivalent. This gives us a hint that the multiplet (NMR signals) generated from H1, H5 and the multiplet generated from H2, H4 could overlap together to produce a complex signal pattern, which is uninterpretable on its multiplicity. Another proton H3, which is three chemical bounds away from H2 and H4, is strongly coupled to H2 and H4, and could produce a Doublet of Doublet (dd). However, in the environment of aromatic system, H3 is also weakly coupled to H1 and H5, and this causes further splitting of the signal of H3. Adding these effects together, we predict that the signal of H3 also shows a complex peak pattern. At the spectrum side, we observe that Peak Cluster 1 has a complex pattern. In addition, Peak Cluster 1 can be further divided into two sub-clusters (see Fig 17 (b)). The left cluster roughly contains 4 protons, and the right cluster contains one proton. Both sub-clusters appear as complex patterns. Through the above deduction, we could conclude that signals from H1, H2, H4 and H5 are all overlapping and produce a complex pattern, which is matched with the left peak cluster in Peak Cluster1. The signal from H3 produces complex peak patterns as well, which happen to match the right peak cluster (see Fig 17).



**Fig 17 Protons in Aromatic Ring and Their Complex Peak Cluster Patterns**

With the assignments on the aromatic system clear, consecutively we apply the coupling analysis on other functional groups. For readability, we redraw Fig 14 in Fig 18. First, we look at O-CH group. By examining the molecular structure (in Fig 18), we know that the only proton which is 3 bound away from the proton in O-CH group is the proton of N-CH group, and any other protons are equal or more than 4 bound away from O-CH group. With our NMR knowledge, we know that only the proton in the N-CH group causes the splitting of the signal of O-CH, and as the result the signal of the O-CH group is shown as a Doublet. Clearly, Peak Cluster 3 now becomes the unique match to the O-CH group.



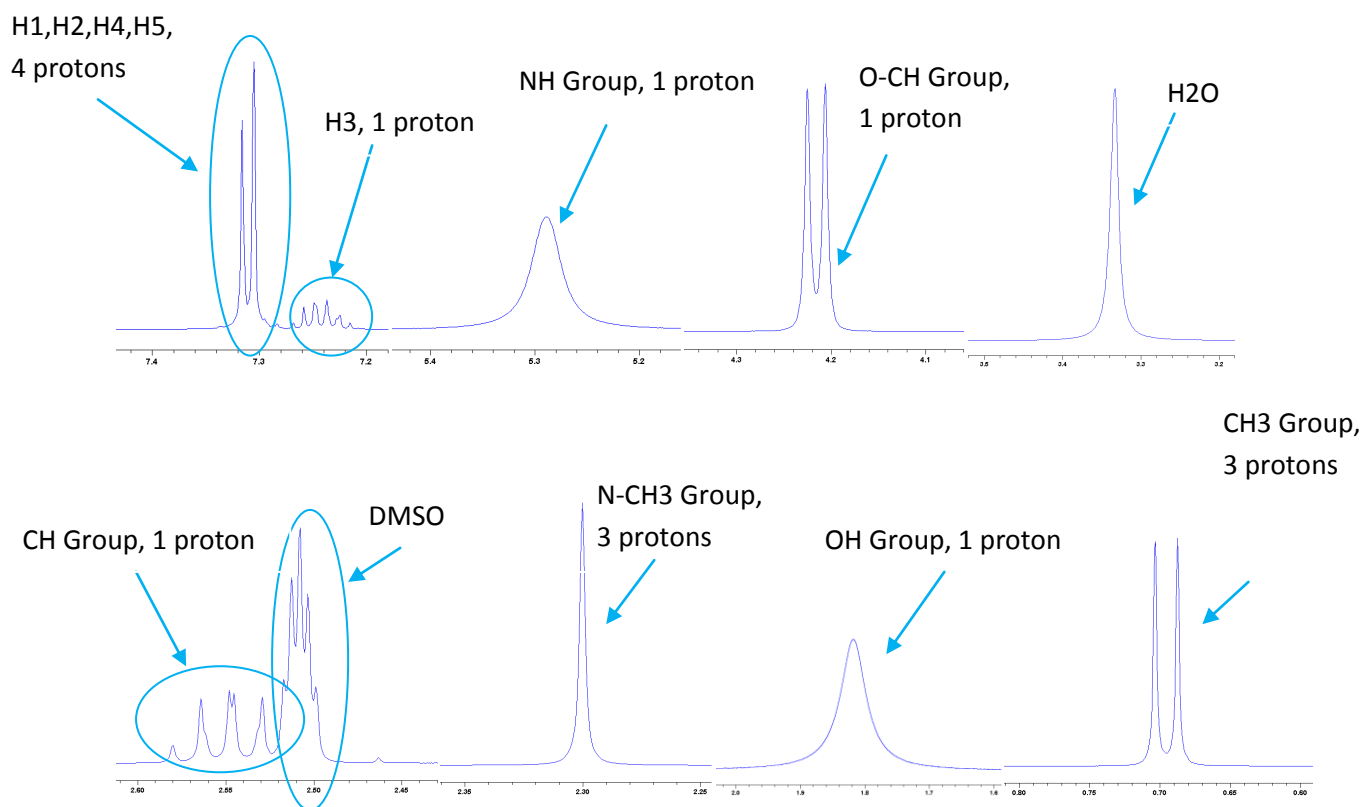
**Fig 18 Identify Functional Group and Proton Numbers from Pseudoephedrin**

To complete the analysis, we next look at the N-CH group. The proton in the N-CH group is three bound away from both the proton in O-CH group and three protons in CH<sub>3</sub> group, and this causes

the signal of the N-CH splits to a Doublet of Quartet. This is exactly matched to the pattern of the multiplet in Peak Cluster 5, since that signal seems to be a Doublet of Quartet with the most right peak overlapping with the signals from DMSO in the spectrum. This evidence helps us to establish a unique map between the multiplet in Peak Cluster 5 and the N-CH group. In addition, it also confirms our previous assumption in 2.1.3.2 that the multiplet in Peak Cluster 5 only contains one proton.

To further confirm the assignments between two CH<sub>3</sub> groups and Peak Cluster 6 and 8, we analyze the multiplicity of CH<sub>3</sub> groups. For the N-CH<sub>3</sub> group, only the proton in the NH group is three bound away from protons in the N-CH<sub>3</sub> group. Due to the fact that the proton in the NH group could exchange with the protons in the solvent, it could not possibly cause the splitting of the signal of the N-CH<sub>3</sub> group. As a result, the signal of the NH-CH<sub>3</sub> could either split to a Doublet or be a Singleton, which is consistent with the multiplet pattern of Peak Cluster 6, which is a Singleton. This confirms our previous assignment during proton number verification in 2.1.3.4. Similarly, the protons of the CH<sub>3</sub> group are only three bound away from the proton of the CH group. Thus, the CH group causes the signal of the CH<sub>3</sub> group to split into a Doublet, which perfectly matches the multiplet pattern of Peak Cluster 8. This evidence further confirms the correctness of previous assignments about the CH<sub>3</sub> groups in 2.1.3.4.

To summarize the above analysis, with coupling analysis, we have created one to one assignments between peak clusters and functional groups. For clarity, we list peak clusters and their assigned functional groups in Fig 19.



**Fig 19 One-to-one assignments between peak clusters and functional groups of +Pseudoephedrin**

Furthermore, connectivity could be used to give additional evidences about the correctness of the current assignments. In the example, the CH group couples with both the O-CH group and the CH<sub>3</sub> group. Correspondingly there are two coupling constants for the Doublet of Quartet in Peak Cluster 5, which is assigned to the CH group. They are experimentally measured as 6.53Hz and 7.58Hz. Correspondingly, the Doublet in Peak Cluster 8, which is assigned to the CH<sub>3</sub>, has the experimentally measured coupling constant of 6.53Hz, which equal to one coupling constant of the Doublet of Quartet in Peak Cluster 5. The Doublet in Peak Cluster 3 which is assigned to the O-CH group, has the coupling constant of 7.58Hz, which is equal to another coupling constant of the Doublet of Quartet. These additional evidence on coupling constants and connectivity give us more confidence about the correctness of our one- to-one assignments between peak clusters and functional groups.

Now we reach our ending point of our 1D 1H NMR spectrum structural verification task on +-Pseudoephedrin. To summarize our findings, we confirm that we have not found any inconsistency between the given spectrum and the molecular structure of +-Pseudoephedrin. Instead we find a reasonable explanation about all functional groups in the structure with peak clusters extracted from the spectrum. In addition, there are no extra peak clusters which cannot be explained either as the signals of the functional group or as the signals from the solvent. Therefore, we derive our conclusion from the given 1D 1H NMR spectrum being consistent with the molecular structure of +-Pseudoephedrin. The structural verification investigation is closed.

#### 2.1.4 Summary of the Human Logic for 1D 1H NMR Molecular Structure Verification

In 2.1.3, we used an example to illustrate the human procedure for 1D 1H NMR spectrum molecular structure verification. Unfortunately, a large part of the human structure verification logic is still not represented in our example. It is partially due to the limited representativeness of our example, and essentially reveals the nature of the flexibility of the human decision logic. Specifically, in the structure verification procedures, there are multiple points where spectroscopists need to make a choice with incomplete knowledge, which could finally lead to the wrong decision. It is amazing that spectroscopists can often avoid the wrong decision by showing the flexibility to return to previous decision points and choose the alternative choice to the goal. This flexibility is shown, for example, in the solvent detection at 2.1.3.2, where we explained that the signal of H<sub>2</sub>O in DMSO is highly dynamic and could show as diversified shape patterns. We also explained there that it is likely to happen that spectroscopists pick up the wrong peak cluster as the signal of H<sub>2</sub>O, and later find that the initial choice of the signal of H<sub>2</sub>O was wrong, and return to reassigning the signal of H<sub>2</sub>O. The same flexibility is also demonstrated in 2.1.3.3 where spectroscopists adopt the hypothesis-driven problem solving strategy to select the peak cluster to compute the normalization factor, and later when a contradiction is found, return to reselect the peak cluster to compute the normalization factor again. The same situation also happens in assignments between peak clusters and functional groups. Since the signal of a typical function group could show in a range of chemical shift positions, it often happens that several experimental peak clusters appearing in a chemical shift range, which

are shared by several functional groups. In this situation, it is hard to do assignments between the peak clusters and the functional groups with chemical shifts. To solve the ambiguity here, spectroscopists still rely on the same hypothesis-driven problem-solving strategy to iteratively assume some assignments as the premise, evaluate all other assignments under the assumption, find contradiction, return back to reassume some other assignments as a new premise. Through enough iteration, spectroscopists can often find the correct assignments. Straying from the point, we believe this hypothesis-driven problem solving strategy, which spectroscopists adopt in structure verification tasks, is the common logic (intelligence) what human beings universally use to solve problems. The strategy spectroscopists use to seek the consistent explanation in our problem is essentially no different to what people use to explore a maze. It is this same strategy (intelligence) which is very well researched in the Artificial Intelligence domain, and as the result is presented as a group of heuristic searching/ optimization algorithms, whose applicability to our problem we will discuss in later chapters. In order to give readers a complete picture about the human structure verification procedure, we summarize it in a flowchart and show it in Fig 20.

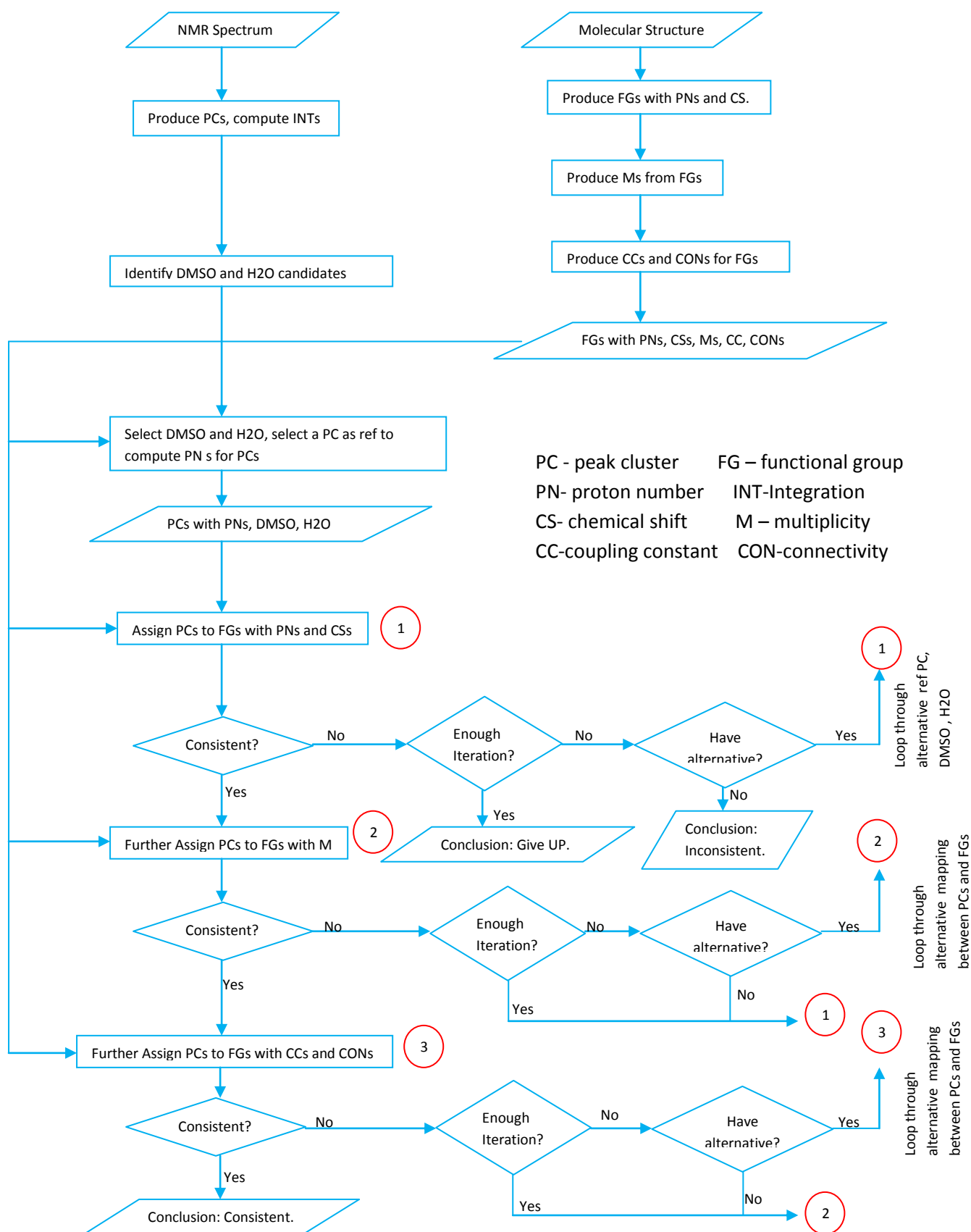


Fig 20 Human Logic for 1D <sup>1</sup>H NMR Structure Verification

## 2.2 Current Automatic NMR Spectrum Molecule Structure Consistency Analysis System

Though the concept of automatic NMR spectrum molecular structure verification is a new topic to NMR community, the efforts to automate the NMR spectrum analysis have a long history. For example, Hamper et al. (Hamper, et al., 1999) demonstrate a qualitative manual inspection approach, in which a set of NMR spectra are performed using stacked plots for each plate row (A–H) in a 96-well plate, and paying attention to the presence of peaks expected in the desired products. Although the results were shown to be very consistent with that of HPLC conversion data, the amount of labor involved significantly hindered the analysis of a large amount of NMR data. A similar approach to aid the interpretation of the NMR spectra from 96-well plates involves a pseudo-2D map, in which spectra are glued by row or columns (Keifer, et al., 2000). Such a graphical presentation of the data is capable of highlighting violations in the expected systematic patterns of NMR signals, but it still requires a lot of attention from spectroscopists and the accuracy of the approach is hard to be quantified either.

Another track of automation relies on the spectrum pattern recognition of R-groups<sup>18</sup> that have been introduced during the reaction for structure validation. For example, unsupervised neural networks have been used to cluster NMR spectra which contain common patterns of R-group, and outliers within such cluster are used to identify inconsistency (Kalelkar, et al., 2002). This approach has been validated for selecting NMR spectra that do not fit the pattern common to a given substituent. However, the structure is not necessarily incorrect, and it remains a challenge for spectroscopists to identify why these spectra are not consistent with the expected pattern. In addition, the technique does not appear to be reliable when significant contributions to the spectral signals, derived from impurities having similar R-patterns (e.g. starting materials or by-products), are present.

Another method based on R-group recognition, named Autodrop, considers that the structure is a combination of R-groups. Correspondingly, 2D HSQC NMR spectrum of the structure is measured and treated as a sum of the spectral patterns from the individual R-groups. The proposed structure is confirmed if the spectral patterns of all R-groups are present in the spectrum (Schröder, et al., 2000). While this method may offer a good visual aid to the interpretation of results, it is restricted to 2D NMR spectral data, which, as we discussed in 1.3.2, has a lower throughput than 1D <sup>1</sup>H NMR spectra. Another source of error comes from the assumption that the spectral patterns are stable. This can sometimes become misleading because magnetically active nuclei in the vicinity of the reaction site may cause changes in the spectral patterns.

The approaches based on R-group recognition are limited in principal since they require the knowledge of reaction, which is often unavailable. A better strategy would start directly from molecular structures and their NMR spectra. For example (Griffiths, 2000) directly verifies the consistency between the structure and spectrum pairs by comparing predicted and experimental chemical shifts. Specifically the method identifies both a list of experimental multiplets from the

---

<sup>18</sup> R-group: in a chemical structural formula, a generic substituent can be written as R. This is a generic placeholder which may replace any portion of the formula as the author finds convenient. Here a substituent means an atom or group of atoms substituted in place of a hydrogen atom on the parent chain of a hydrocarbon in organic chemistry and biochemistry.

spectrum and extracts their chemical shifts, and a list of multiplet from the molecular structure and predicts their chemical shifts<sup>19</sup>. Then it creates a mismatch matrix of predicted and experimental chemical shifts, and this is followed by manipulating the matrix to minimize the sum of the diagonal. The resulting sum of the diagonal measures the mismatch among predicted and experimental chemical shifts, and as a consistent result it should not exceed a predefined threshold. Though the approach only relies on the information of the chemical shift, in the paper (Griffiths, 2000) it has been shown to produce very good result in the given test set. This approach gives us a good starting point since it only requires the NMR spectrum and molecular structure as its input, which is closer to the approach spectroscopists are familiar with. However, it only supplies a mismatch value, but does not supply information regarding which predicted resonance is paired with which experimental signal. This limitation denies spectroscopists the great value that is contained in an assignment produced during structure verification. The assignments, as we explained in 2.1.3, would allow spectroscopists to directly compare the properties of predicted and experimental signals and as the arbitral to further control the accuracy of the automatically generated structural verification conclusions. In addition, the approach completely relies on chemical shift information for structure verification. But, as we see in 2.1.3, information of chemical shifts alone is not able to discriminate functional groups which appear in close chemical shift positions. Therefore, the accuracy of the approach heavily relies on the accuracy of the pinpoint prediction of the chemical shift positions of the protons in the given molecule. As we will discuss later in 2.2, accurate prediction of chemical shift positions of protons are difficult tasks since the chemical shift position of the given proton is not only determined by the proton's local environment, but is also being influenced by many other external factors e.g. experimental conditions, etc. This denies the applicability of the approach in practical 1D <sup>1</sup>H NMR structure verification tasks.

To address the problem and improve the accuracy of the approach, (Golotvin, et al., 2006) proposes to further introduce proton number and multiplicity into the structure validation. Specifically, a mismatch function is created to linearly combine the dissimilarities between experimental multiplets and predicted multiplets along chemical shift, proton number and multiplicity, and then a similar mismatch matrix is built among predicted and experimental multiplets, where the computed dissimilarity values between experimental multiplets and predicted multiplets are recorded (ibid.). To seek the minimal sum of the diagonal values in the matrix, a Mont Carlo based optimization approach (Press, et al., 1992) is adopted to assist the primitive searching approach used in the previous strategy (Griffiths, 2000). With the introduction of multiplicity (J-coupling) into the system, the accurate assignments between predicted and experimental signals become possible. This is an important improvement, which makes it possible to directly compare the performance of the automated process to that of spectroscopists. In fact, the approach (Golotvin, et al., 2006) has been commercialized as a product, and has been proven to be the best automatic structure verification system developed so far. Therefore, in this chapter we focus on introducing the technology and the software architecture of this system, and discuss its advantages and disadvantages in detail. Specifically, in section 2.2.1, we explain its system architecture and methodologies, and in section 2.3, we discuss its advantages and disadvantages.

---

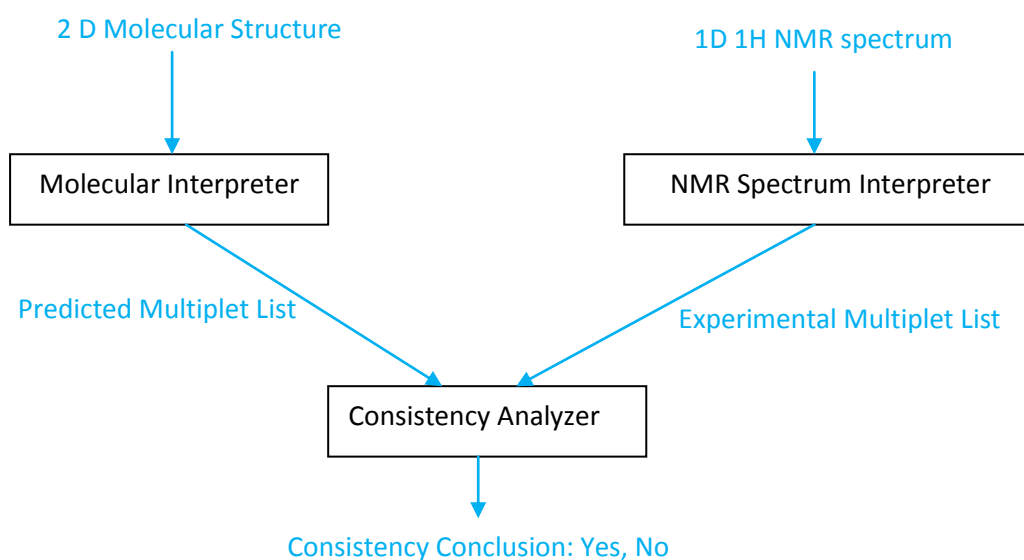
<sup>19</sup> Note, the methodologies of both identifying experimental multiplets and predicting chemical shifts from the spectrum are introduced later in 2.2.1.



### 2.2.1 General System Architecture

The automatic 1D <sup>1</sup>H NMR spectrum molecular structure verification system contains three components (see Fig 21):

- (1) Molecular Interpreter: a module to automatically calculate a list of predicted multiplets from the 2D molecular structure.
- (2) NMR Spectrum Interpreter: a module to automatically interpret a list of experimental multiplets from the 1D <sup>1</sup>H NMR spectrum.
- (3) Consistency Analyzer: a module to analyze the consistency between the predicted multiplet list and the experimental multiplet list.



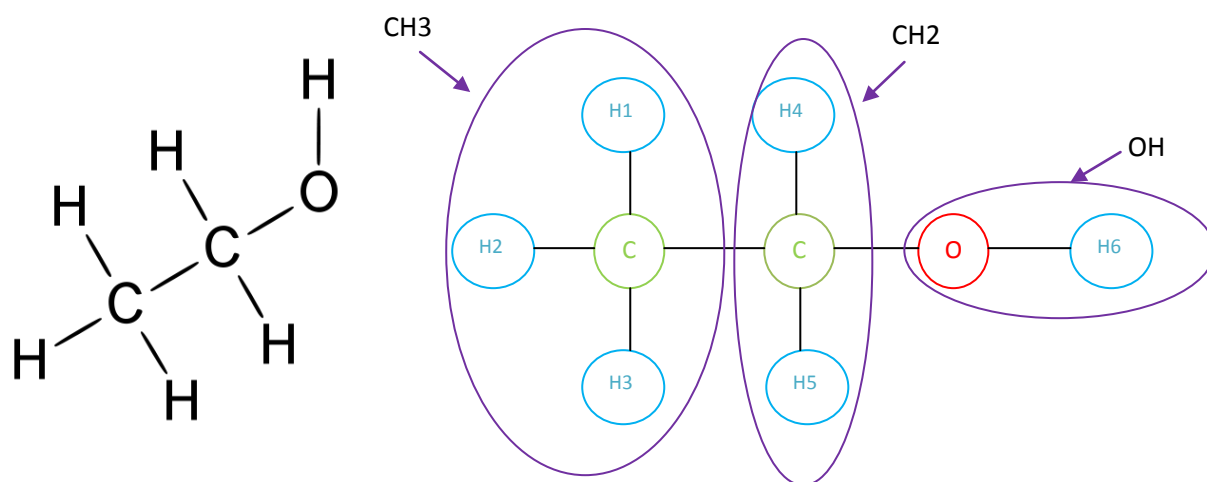
**Fig 21 Structure of NMR Structure Verification System**

#### 2.2.1.1 Molecular Interpreter

Molecular Interpreter is a module to calculate/predict multiplets from the two dimensional molecular structure. Calculating/predicting multiplets is a procedure to identify all chemically equivalent functional groups from the molecule, and extract NMR properties such as chemical shift, J-coupling, coupling constants, connectivity, proton numbers for each chemically equivalent functional group.

### 2.2.1.1.a Identifying Chemical Equivalent Functional Groups

Chemical equivalent protons are defined as protons which are geometrically symmetric to each other in the 2D molecular structure. For example, in Ethanol (see Fig 22), we see that H1, H2 and H3 in the CH<sub>3</sub> group are symmetric, and therefore are chemically equivalent. H4 and H5 in the CH<sub>2</sub> are chemically equivalent, and H6 in the OH is an individual proton which is not chemically equivalent to other protons.



**Fig 22 Chemical Equivalent Protons in Ethanol**

Chemically equivalent protons can be automatically identified by building a graph upon the 2D molecular structure (see Fig 22.) and traveling the graph. Here each atom represents a vertex in the graph, and each edge represents a chemical bound between two atoms. Specifically, traveling starts from every proton node, and identification of equivalent protons is done by comparing traces of expansion trees starting from each proton node.

### 2.2.1.1.b Predicting Chemical Shift

Dominating approaches for proton chemical shift prediction include database approaches (Williams, 2000) (KnowItAll Informatics, 2009) (Chemical Concepts, 1998) (ACD, 1996 - 2009), additivity rules approaches (Williams, 2000) (Schaller, et al., 1995) (Schaller, et al., 1994) (Schaller, et al., 1996) (Pretsch, et al., 1991) (Fürst, et al., 1990) (Pretsch, et al., 2004) (Fürst, et al., 1990) (Steinbeck, et al., 2003), and quantum chemical approaches (ABRAHAM, 1999).

Database approaches utilize the availability of large NMR spectral databases containing chemical structure with assigned chemical shifts to predict the chemical shifts of target molecular structure. In such databases, the surrounding environments of atoms in a molecular structure are encoded as

'spherical' codes, e.g. HOSE (Hierarchical Organization of Spherical Environments) codes (Bremser, 1978), and NMR spectral signals e.g. chemical shift, coupling constants are assigned to the corresponding atoms. During the prediction, the algorithm searches for matches between the 'spherical' codes for each atom in the target molecule and the 'spherical' codes in the database to fetch the suitable shift for prediction. Note, the 'spherical' code based database approach could be applied to the prediction of coupling constants, as well.

With these approaches, prediction accuracy can reach the level of less than  $\pm 0.3$  ppm on average (Golotvin, et al., 2006) (Williams, 2000). However, the prediction accuracy is sensitive to the structural diversity and therefore is proportional to the size of the molecular structure database. Collecting a large and reliable molecular structure database along with the corresponding NMR spectra is an expensive and time consuming task. In addition, chemical shifts of protons are easily fluctuating depending on measurement condition. Despite above disadvantages, the approaches are commercialized into several NMR shift prediction software packages, which include Sadtler's Know-It-All package (KnowItAll Informatics, 2009), Chemical Concepts' SpecInfo (Chemical Concepts, 1998), and Advanced Chemistry Developments (ACD Labs)'s ACD/NMR Predictors for  $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^{19}\text{F}$  and  $^{31}\text{P}$  nuclei (ACD, 1996 - 2009).

Alternatively, a set of chemical shift prediction rules, which are summarized based on empirical rules (so called additivity rules), are used to result in a less elaborate and therefore cheaper but cruder chemical shift prediction (Schaller, et al., 1995) (Schaller, et al., 1994) (Schaller, et al., 1996) (Pretsch, et al., 1991) (Fürst, et al., 1990) (Pretsch, et al., 2004) (Fürst, et al., 1990) (Steinbeck, et al., 2003). Briefly, first a number of substructures with applicable additivity rules are automatically identified. The rest of the molecule is treated as substituents associated with each of the substructures. Next it assigns a rough chemical shift to each proton in the substructure, and then relies on the protons' local structure properties e.g. connected bond types, bond angles, etc. to adjust chemical shift prediction for the proton. Estimates have been given by (Schaller, et al., 1996) that the NMR shift locations can be predicted up to within 0.3 ppm accuracy. However, for structures where no or few additivity rules are available, the technique suffers (Williams, 2000). Nonetheless, the commercial implementation of this approach is available in packages such as CambridgeSoft's ChemDraw Ultra (CambridgeSoft, 2009) and Upstream Solutions' SpecTool (Heller, 1994).

Beyond the above mentioned empirical approaches, quantum chemical theory can be used to theoretically calculate proton chemical shifts based on electronic and steric interactions. A report in (ABRAHAM, 1999) shows high prediction accuracies (typically  $\pm 0.1$  ppm) with this approach. To reach this accuracy, a precise three-dimensional molecular structure is needed, which again has to be determined by multi dimensional NMR or X-ray crystallography. It is also possible to calculate the molecule's three-dimensional structure theoretically. However, since accurate quantum mechanical geometry optimization routines only yield gas phase structures, substantial differences to the solution structure observed by NMR spectroscopy are common. This makes the approach inferior to database approaches.

#### *2.2.1.1.c Predicting Number of Couplings and Coupling Constant*

A set of empirical rules can be used to automatically predict the existence of couplings between protons. In practice, this approach often yields a reliable number of couplings and thus results in a multiplicity prediction with the exception of long range couplings (Karplus, 1963) (Karplus, 1960) (Barfield, et al., 1969). To obtain an accurate prediction of long range couplings and to precisely

estimate coupling constants, an accurate three dimensional structure of the molecule is needed. As mentioned above, the theoretical calculation of the three dimensional molecular structure is not reliable enough, so that neither the long range coupling prediction, nor the coupling constant prediction of it is reliable.

#### *2.2.1.1.d Count Total Number of Protons within a Molecule*

This simple task is easily reached by adopting a graph traveling algorithm (Sedgewick, 2001).

To summarize, reliable approaches exist for the prediction of the number of couplings and counting the number of protons,. For chemical shift and coupling constant prediction, current approaches still cannot reach the prediction accuracy and reliability required by a fully automated molecular structure verification system.

#### *2.2.1.2 NMR Spectrum Interpreter*

NMR spectrum interpretation is split into three subtasks (Griffiths, 2000) (Golotvin, et al., 2002) (Hoye, et al., 2002):

- (1) Automatically identify peaks in spectrum.
- (2) Group symmetric peaks into peak clusters.
- (3) Estimate multiplicities and coupling constants for each peak cluster.

##### *2.2.1.2.a Automatically Identifying Peaks in Spectrum*

The automatic identification of peaks in a spectrum is a fundamental problem widely spreading over different domains e.g. electronics, communication, spectrum interpretation, etc. Dominant technologies used in peak picking include derivative-based approaches, and deconvolution-based approaches. However, both techniques suffer severe drawbacks. Derivative-based approaches are notorious for introducing noise peaks in noisy spectra. On the other hand, deconvolution-based approaches tend to create artifact peaks. Here, noise peaks denote tiny local minima, local maxima, and inflexion points in the spectrum resulting from the NMR system noise. Artifact peaks denote pseudo local minima, local maxima, and inflexion points created during the deconvolution procedure. In the field of NMR spectroscopy, some NMR specific knowledge is used, e.g. Lorentzian or Gaussian peak shapes have been introduced into the deconvolution-based peak picking approaches to reduce peak picking errors. However, peaks in experimental NMR spectra are often different from the theoretical assumption (e.g. Gaussian/ Lorentz shape assumption), which limits their efficacy to increase peak picking accuracy. As a result, NMR peak picking programs still rely on the immemorial approaches e.g. setting high noise threshold or reducing the number of peak fitting iterations to reduce noisy peaks and artifact peaks. However, these approaches are well known of missing real peaks. This limits their applicability in structure verification tasks.

### 2.2.1.2.b Grouping Symmetric Peaks into Peak Clusters

Griffiths (Griffiths, 2000) proposes an approach, which is used as a standard technique in current structure verification software packages e.g. (Griffiths, et al., 2002) (Golotvin, et al., 2006). First, the NMR peak list is split into zones depending on the distances between individual peaks. Then, in each zone, peaks are grouped into a set of disjoint peak clusters with peaks' positional symmetry. The problem of this approach is that it only builds the most likely peak clusters within a given zone instead of building all possible peak clusters, and this may cause missing multiplet interpretations (see Fig 24 on page 59 for an example).

### 2.2.1.2.c Estimating Multiplicities and Coupling Constants for Each Peak Cluster

The current technique to estimate the multiplicity and coupling constant is fairly reliable in non-overlapped spectra. (Golotvin, et al., 2002) and (Hoye, et al., 2002) report that a complex multiplicity up to seven coupling constants could be determined automatically (with a given, error-free peak list). In principal, the task of deducing the multiplicity and the coupling constants from a peak cluster can be considered as a reverse process of generating a conventional splitting tree from a single peak through first order multiplet analysis. This makes it easy to implement it with typical divide and conquer algorithms (Sedgewick, 1997).

### 2.2.1.3 Consistency Analyzer

The matching problem is generally solved with the following framework:

- (i) Build a matching matrix, with one dimension representing the experimental multiplet, and another dimension representing the calculated multiplet.
- (ii) For each pair of experimental and calculated multiplets, compute the similarity between them, and store the similarity value into the matching matrix.
- (iii) Search for a matched list (of experimental and calculated multiplet pairs) in the matching matrix to maximize a given criterion.

The published <sup>1</sup>H structure verification approaches ( (Griffiths, et al., 2002), (Golotvin, et al., 2006)) use the framework described above to match the experimental and calculated multiplets, they mainly differ in the similarity measurements they apply.

E.g., (Griffiths, et al., 2002) use a chemical shift rule –

```

if ((|experimental chemical shift – calculated chemical shift|) < Chemical Shift Error Tolerance)
    then similarity = 1
    else similarity = 0
  
```

and then a multiplicity rule –

if (experimental multiplicity = calculated multiplicity) then similarity = 1  
 else similarity = 0;

to assign a similarity value to each experimental and computed multiplet pair.

Iterative permutations are repeated through either matrix columns or matrix rows to find the maximum number of non-zero diagonal elements. As a result, these max diagonal non-zero elements give a matching list between experimental and calculated multiplets. The ratio of the non-zero diagonal element and the asymptote maximum non-zero diagonal element are used as structure verification score.

In contrast, (Golotvin, et al., 2006) use chemical shift, multiplicity, and proton number (normalized integration) to measure the similarity. Instead of using a 0/1 decision boundary, they use a penalty function to assign the chemical shift similarity score. Additionally, they also consider the consistency between the normalized integration of the experimental multiplet and the proton number of the calculated multiplet. Here, the normalized integration is determined as the ratio of the multiplets' integration and the normalization factor, which is computed by dividing total integration outside the "dark region"<sup>20</sup> (ACD, 2005) by total proton numbers in the molecular structure. Finally, the similarity value between an experimental multiplet and a calculated multiplet is defined as:

$$S = W\_shift \times S\_shift + W\_proton\_num \times S\_proton\_num + W\_mult \times S\_mult \quad (1)$$

Where  $S\_shift$ ,  $S\_proton\_num$ , and  $S\_mult$  denote the similarity between the experimental and the calculated multiplets' chemical shift, proton number and multiplicity, while  $W\_shift$ ,  $W\_proton\_num$  and  $W\_mult$  denote the weighting factors which are used to control the relative importance of the three NMR properties: chemical shift, proton number and multiplicity, and which are designed to be manually changeable by spectroscopists. With formula (1), the best possible matching list is searched by maximizing  $\sum S$  with matrix permutations, which is followed by a Mont Carlo optimization. (Note, the exact format of the penalty function and the approach of optimization are unpublished.)

Both approaches heavily rely on accurate chemical shift predictions. Furthermore, simple coupling scalars are used for number of coupling matching, but more detailed analysis such as coupling connectivity and coupling constants matching are still missing during coupling analysis. This is both due to the unreliable experimental coupling constants estimation and the computational complexity of checking the connectivity. In addition, using a normalized integration to approximately check the consistency of the proton count is an approach sensitive to the noise in the spectrum, e.g. depending on the accuracy of selecting a "dark region", etc.

---

<sup>20</sup> dark region: any extraneous peaks from an 1D <sup>1</sup>H NMR Spectrum, which do not overlap significantly with signal peaks of the Molecule,

## 2.2.2 Difference between Human Structure Verification Logic and Techniques used in the Structure Verification System

Above, we introduced the system architecture and the techniques used in current automatic molecular structure NMR verification systems. Such systems are commercialized and have been supplied to the pharmaceutical industry. However, market research from the NMR manufacturer shows that the systems are not being used to replace human spectroscopists for structure verification tasks in compound library management. Instead it is used as an assistant tool to aid spectroscopists in structural verification tasks. Unfortunately, this utilization of the system has deviated from the original goal – **to build a fully automatic structure verification system to replace human spectroscopists in compound library management**. The underuse of the system is explained by the practical participants as its inability to supply the decision accuracy comparable to human spectroscopists. We believe that the inferiority of the system originates from its difference to human spectroscopists' logic to solve the problem. Therefore, in this section, we start to analyze the difference between the techniques used in the system and the human logic adopted in NMR structure verification tasks.

### 2.2.2.1 Differences in Molecular Interpretation

Both the system and human spectroscopists attempt to build a predicted/calculated multiplet list from the given 2D molecular structure. As we explained in 2.2.1.1, the system builds a multiplet for a functional group by predicting its pinpoint chemical shift position, predicting its number of coupling, and counting the proton numbers of the functional group. As we have commented multiple times in 2.2.1, the system pursues the absolute chemical shift prediction, which is significantly different from the approach of spectroscopists. As we explained, chemical shift prediction in proton NMR is a difficult task, and this is due to the fact that the experimental chemical shift position of the proton is sensitive to the experimental environment (e.g. measurement temperature, solvent, PH value, etc). To avoid this problem, spectroscopists turn to give a chemical shift interval to a functional group, in which it insures that the signal of the functional group will appear in the given interval. Obviously, defining an interval to cover the signal is a much easier task compared to the task of predicting the location of the signal. Therefore, the human approach produce the less error-prone predicted/calculated multiplet list compared to the system.

Besides chemical shift, the system also predicts the number of coupling for each predicted multiplet. This additional information helps to describe the shape of the predicted multiplet. However, the description of the shape of the multiplet by the system is approximate and incomplete. Comparatively, spectroscopists produce two additional NMR properties to refine the description of the multipet shape. Specifically, spectroscopists give a coupling constant interval for each predicted coupling. (Note, analogy to human strategy of processing chemical shift, spectroscopists prefer producing coupling constant intervals instead of directly predicting coupling constant.) In addition, spectroscopists also build a connectivity network upon the predicted multiplet list to describe coupling correlations among predicted multiplets. By supplying this additional information, human

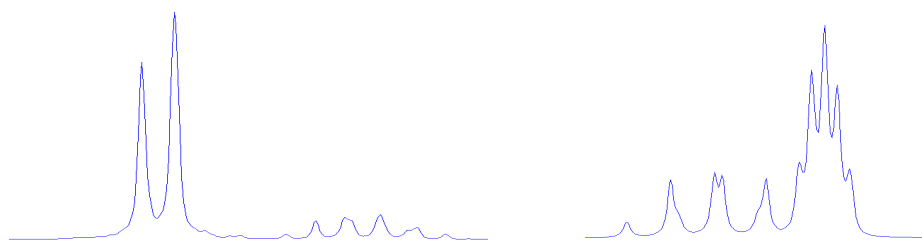
spectroscopists produce finer multiplet shapes of predicted/calculated multiplet list compared to the system. Note, spectroscopists could even supply more information e.g. magnetic equivalency, high-order multiplet distortion, etc about the predicted multiplet shape if necessary.

To summarize the above mentioned differences, through introducing chemical shift interval and coupling constant interval, spectroscopists essentially define a predicted multiplet *hypotheses space* to cover all possible variance in shape and position of the multiplet, which could be generated from a given functional group. On the contrary, the system attempts to predict the *exact* position of the multiplet despite its possible variance, and ignore the multiplet shape to a large extent. This approach oversimplifies the problem in nature, and could produce the predicted multiplet which is significantly deviated from the experimentally observable multiplet, and finally deteriorates the performance of consistency analysis at a later stage.

#### 2.2.2.2 Differences in NMR Spectrum Interpreter

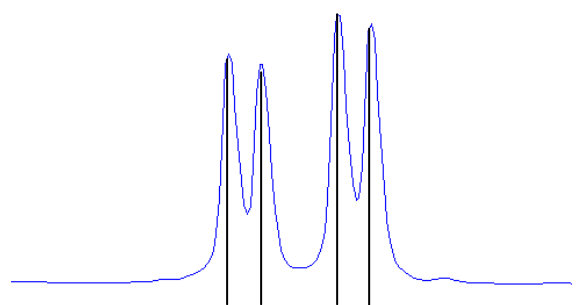
The approach that the system adopts to interpret the spectrum is significantly different from that of human spectroscopists. As we explained in 2.2.1.2, the system attempts to extract an experimental multiplet list from the 1D <sup>1</sup>H NMR spectrum. Briefly, this is achieved by automatic peak picking, followed by grouping peaks into first order multiplets through position and amplitude symmetry analysis (first order multiplet analysis) (Prost, et al., 2006). Comparatively, spectroscopists attempt to build a peak cluster list instead of an experimental multiplet list (see 2.1.3.1 for detail). In spectroscopists' logic, a peak cluster is a unit of NMR signals whose integration could be reliably estimated. It is not necessary that a peak cluster is mapped to a first-order multiplet. In fact, it could happen that a peak cluster maps to multiple first-order multiplets which happen to overlap altogether. It could also happen that a peak cluster represents a high (second)-order multiplet, or NMR signals from protons which are chemically equivalent but magnetically inequivalent. As we introduced in 2.1, both high (second)-order multiplets or magnetic inequivalency often appear as complex signal patterns, while first-order multiplet analysis makes no sense in principal. Similarly, first order multiplets generated from similar functional groups often overlap together in the spectrum, which creates complex signal patterns as well. These complex signal patterns are not experimentally following the Pascal triangle rules anymore, which make first order analysis impossible (see Fig 23 for an example). Therefore, the approach to build peak clusters gives spectroscopists flexibility to avoid applying the first-order multiplet analysis upon complex signal patterns. Instead only peak clusters, which show clear first-order multiplet patterns, are assigned multiplicities. Whereupon complex signal patterns, only integration are recorded for later consistency analysis, and the multiplicity of the patterns are ignored. Therefore, the peak cluster has to provide a reliable integration. On the contrary, the system insists applying the first-order multiplet analysis everywhere on the spectrum no matter if it is possible or reasonable to do that. As the result, it often happens that the system creates nonsensical first-order multiplets on the complex signal patterns e.g. signals in the aromatic region of the spectrum. These noise/ artificial multiplets produce misleading interpretation, and will deteriorate the consistency analysis at the later stage.





**Fig 23 (a) High order multiplet and (b) overlap of first order multiplets**

In addition, for a peak cluster, which appears as the simple signal pattern (clear first-order multiplet pattern), it often happens that multiple first-order multiplet interpretations exist. For example in Fig 24, the signal shows a clear 4-peak pattern. With the first-order multiplet analysis, we could reasonably assign it as a Doublet of Doublet. Spectroscopists agree that the pattern is most likely to be a Doublet of Doublet. But they also give other (possibly unlikely) interpretations e.g. the pattern represents two independent Doublets or even four independent Singletons. With the above strategy, spectroscopists essentially build the *experimental multiplet hypotheses space*, which covers all possible first-order multiplet interpretations of the given peak cluster, and leave the precise identification of the given peak cluster's multiplicity to the consistency analysis stage, where the information of the predicted multiplets is supplied. Comparatively, for the pattern in Fig 24, the system only produces one interpretation – a Doublet of Doublet, while ignoring other interpretations. We could understand this strategy as the system attempts to only produce the most “likely” multiplet interpretation. In other words, there is the possibility that the multiplet interpretation of the system is wrong. This would finally deteriorate the consistency analysis in later stage.



Doublet of Doublet or Doublet + Doublet?

**Fig 24 Example of missing experimental multiplet interpretations**

### 2.2.2.3 Differences in Consistency Analysis

With more abundant and more reliable NMR information extracted from both the structure and the spectrum, spectroscopists exhibit more cognitive consistency analysis procedure, compared to the system. Specifically, spectroscopists start consistency analysis by comparing the peak cluster list and the predicted multiplet list, while comparatively the system does consistency analysis by matching the experimental multiplet lists and the predicted multiplet list. As we explained in 2.2.2.2, matching between peak clusters and predicted multiplets gives spectroscopists flexibility to deal with complex signal patterns, which are normally produced by first-order multiplet overlapping, high-order multiplet, magnetic inequivalency, etc. For a peak pattern which contains complex signal patterns, it is only compared with predicted multiplets by chemical shift and proton numbers, while the comparison upon multiplicity is ignored. For other peak clusters which only contain simple signal patterns, complete comparisons are conducted which include chemical shift matching, proton number matching and complete coupling analysis. On the contrary, the system intentionally matches each experimental multiplet to the predicted multiplet despite its authenticity. As a result, both mendacious experimental multiplets and imprecise experimental multiplets could conduce wrong assignments between an experimental multiplet and a predicted multiplet, and therewith lead to the wrong consistent analysis conclusion.

Moreover, for the comparison of multiplicity between a peak cluster with the simple signal pattern and a predicted multiplet, spectroscopists utilize more NMR information, compared with what the system applies. First, relying on the experimental multiplet hypotheses space built upon the peak cluster, spectroscopists have a chance to compare multiple multiplet interpretations of the peak cluster to the predicted multiplet. On the contrast, there is only one multiplicity comparison between an experimental multiplet and a predicted multiplet in the system. Obviously, with the wrong given experimental multiplet interpretation, an error-prone comparison conclusion could be produced, which will finally deteriorate the consistency analysis. Second, with a more precise multiplet shape pattern prediction in 2.2.2.1, during multiplicity comparison between a peak cluster and a predicted multiplet, spectroscopists match the consistency upon coupling constants in addition to the matching upon number of couplings. This additional comparison dimension extremely increases the accuracy of spectroscopists' multiplicity matching decision. Comparatively, the system only utilizes the number of coupling to match the multiplicity consistency between the experimental multiplet and the predicted multiplet, which make it uncompetitive to that of spectroscopists. Third, to further increase the accuracy of the assignments between peak clusters with simple signal patterns and predicted multiplets, spectroscopists utilize the connectivity network defined among the predicted multiplets to validate the correctness of the assignments (see 2.1.3.5), which the system never touches upon.

In addition, spectroscopists deduce a consistent conclusion by confirming that all individual comparisons along chemical shift, proton number and coupling analysis between the peak cluster list and the predicted multiplet list are consistent. A single inconsistency along any of these comparison dimensions causes the inconsistent conclusion. We could abstractly understand this strategy as setting an equal weight among different comparison dimensions. The benefit of this strategy is that

it helps spectroscopists to maintain a low false positive rate, which finally guarantee the accuracy of the conclusion. Comparatively, the system utilizes the linear function to synthetically consider the three comparison dimension – chemical shift, proton number and number of coupling altogether (see 2.2.1.3). By intentionally introducing weighting factors for different comparison dimensions, the system attempts to gain the flexibility of farther relying on certain comparison dimensions than others. In fact, the system is often set to give majority weights to the chemical shift comparison so as to the comparison results along proton numbers and multiplicity becomes negligible to the final consistent conclusion (Golotvin, et al., 2006). Therefore, it could happen that even if inconsistency is found in the proton number comparison or in the multiplicity comparison, the system still produces a consistent conclusion. This will dramatically increase the false positive rate, and finally deteriorates the accuracy of the system.

## 2.3 NMR Structure Verification Technique beyond 1D <sup>1</sup>H NMR Spectra

Other spectra such as 2D <sup>1</sup>H NMR spectra and 1D <sup>13</sup>C NMR spectra have been used to provide additional information for molecular structure verification tasks (Griffiths, et al., 2005) (Golotvin, et al., 2007). However, time expenses of acquiring these types of spectra are dramatically higher than that of 1D <sup>1</sup>H NMR spectra, which makes them impractical for large batch structural verification tasks in practical compound library management. Additionally, the costly NMR instrument time intrinsically makes the whole process more expensive. Therefore, to avoid the deviation from our main topic, we leave the interested readers to (Griffiths, et al., 2005) and (Golotvin, et al., 2007).

## 2.4 Conclusion

The molecular structure 1D <sup>1</sup>H NMR verification system currently available in the academic and industrial world is still not robust enough to be used without human supervision in practice. Human interaction and supervision is still necessary, and so far these tools are only used to assistant human spectroscopists, while the human expert still has to interpret each spectrum individually. Today, the traditional NMR spectroscopist based human interpretation is still the core methodology for structural verification tasks.

## **Chapter 3 The Proposal**

In 2.2, we introduced the system architecture and the technologies used in the automatic 1D 1H NMR molecular structural verification system, and compared the difference between it and that adopted by human spectroscopists. As a conclusion, the comparison shows the superiority of the human logic over the techniques applied in the system. This superiority of the human logic gives us new hints how to approach structure verification systems. As we introduced in 1.3.5, artificial intelligence, as a branch in computer science, has researched the human problem solving logic for more than half a century. And it is backed by multiple successful deployments of expert systems, which are built by mimicking human experts in the domain and having successfully demonstrated that human labor can be replaced in the domain. These previous success stories encourage us to propose utilizing the methodologies developed in modern artificial intelligence to mimic spectroscopists' structure verification procedure. With this strategy, we hope the new system based on mimicking spectroscopists can address the problems of the current structure verification system, and reach the consistency analysis accuracy comparable to that of human spectroscopists. As the ultimate goal, the system should be qualified to completely replace the human spectroscopists for molecular structure verification tasks in compound library management.

Through wide and deep negotiation with spectroscopists and compound library management practitioners from both NMR manufactures and pharmaceutical companies, a suitable goal has been set to guarantee that the system is accurate and reliable enough to be used to replace human experts in the practical compound library management environments. Specifically,

- a. The new system is required to produce consistency decisions which are correct in above 90% cases (see 6.1).
- b. The system should produce less than 5% false positive alarm rate (see 6.1).
- c. The system should be able to select a NMR signal (a multiplet in most cases) of main substance from 1D 1H NMR spectrum to be used for quantification.
- d. The consistency decision of the system has to be expatiated by a human understandable consistency analysis report, which is supposed to explain how the system reaches its decisions. The reports will be used by NMR spectroscopists to confirm the correctness of the structure verification conclusion generated by the system. The practical participants emphasize that this human intervention should act as the arbitral approach to further control the structure verification quality.

### **3.1 Implementation Plans**

With the goal defined above, the system architecture needs to be modified to include both newly designed mechanisms and to delete old mechanisms which are out of date. As we discussed in 2.2.2, the

major advantages of the human structure verification process versus the structure verification system is that :

- (1) Human spectroscopists build hypotheses space from both 2D molecular structure and 1D <sup>1</sup>H NMR spectrum to cover all possible NMR property interpretations,
- (2) Human spectroscopists find a reasonable matching (explanation) by efficiently searching through both hypotheses spaces.

To mimic these human logics, a list of new mechanisms needs to be implemented. Specifically,

- a. A new mechanism is needed to be designed, implemented and embedded into the Molecular Interpreter to generate the predicted multiplet hypotheses space for each chemically equivalent functional group from the input 2D molecular structure.
- b. A new mechanism is needed to be designed, implemented and embedded into the Spectrum Interpreter to automatically identify the NMR peaks from the spectrum.
- c. A new mechanism is needed to be designed, implemented and embedded into the Spectrum Interpreter to automatically identify the NMR signals from solvent.
- d. A new mechanism is needed to be designed, implemented and embedded into the Spectrum Interpreter to generate the peak cluster hypotheses from the spectrum. In addition, a sub-routine is needed to be designed and implemented to build an experimental first-order multiplet hypothesis space for each peak cluster to represent all possible interpretation. Correspondingly, the original first-order multiplet analysis mechanism in the system is abandoned.
- e. A new human-mimicking optimization mechanism is needed to be designed, implemented and added into the Consistency Analyzer to efficiently search a reasonable explanation between the peak cluster hypotheses and the predicted multiplet hypotheses. Correspondingly, the original matrix-manipulation based searching routine and Monte Carlo based optimization is replaced (see 6.2.3 for discussion on Monte Carlo optimization).

## 3.2 Possible Challenges

With the requirements to design human-mimicking mechanism mentioned in 3.1, we expect new technique challenges will emerge. Specifically, to list some, we expect

- a. To insure not missing experimental multiplet interpretations and cover all possible experimental multiplet hypotheses for a peak cluster, the peak picking routine is required to pick up all possible peaks from the given 1D <sup>1</sup>H NMR spectrum. This requirement is high and obviously beyond the ability of the current peak picking routines, since it is well known that current peak picking techniques miss peaks or produce artifact peaks. To meet the requirement, new peak picking approach is need to be designed and implemented.

- b. Assuming all peaks are successfully picked from the spectrum, a second technical challenge comes from the requirement to group the peaks into all possible multiplet interpretations. Note, in the current system, the first-order multiplet analyzer only need to extract the most likely experimental multiplets from peaks. There are no requirements upon interpretation completeness over the current routine.
- c. Similar to b, assuming we have the peak list, how to generate all possible peak cluster hypotheses, which could be given reliable integration estimation, is also a challenging subject.
- d. NMR signals of H<sub>2</sub>O in 1D <sup>1</sup>H NMR spectra could change drastically in both shape and position. To design a routine to catch this flexibility is another challenge.
- e. With the introduction of hypotheses space both from the structure and the spectrum, the search space for a solution is dramatically (geometrically) increased. This could bring in new computational challenge, which makes previous simple heuristic searching based optimization routines computationally infeasible. Additional comparisons introduced (e.g. connectivity analysis) generate new requirements for searching methodology. Spectroscopists can find the consistency explanation quickly in most cases despite facing the same searching space. We attribute this to the complex and flexible heuristic nature of the human logic. Therefore, building an efficient consistency analyzer relies on successfully designing an optimization routine to mimic this human logic. This forms the most difficult challenge in this thesis.

In summary the thesis context represented so far consists of an introduction of the motivation, background information, and our proposal of building a new automatic molecular structure 1D <sup>1</sup>H NMR spectrum verification system. In the following chapter 4, we are going to explain our approach in detail. Specifically, we provide a detailed design of our system architecture, which we use to mimic the human spectroscopists' structure verification process. Especially, we focus on introducing technology that we use to solve challenges we mentioned in 3.2. In addition, in chapter 5, we attempt to utilize the mathematic language to strictly describe our human-mimicking optimization routine, and to unify it under the maximal likelihood framework. In chapter 6, we will explain the experimental setup, and present our evaluation results in detail.

## **Part II**

### **Automatic 1D 1H NMR Molecule Structure Verification Software Architecture, Methods and Evaluation**





## **Chapter 4 Automatic 1D 1H NMR Molecule Structure Verification Architecture and Methods**

To build a fully automatic 1D 1H NMR structure verification system to achieve the performance comparable to that of human spectroscopists, we need to:

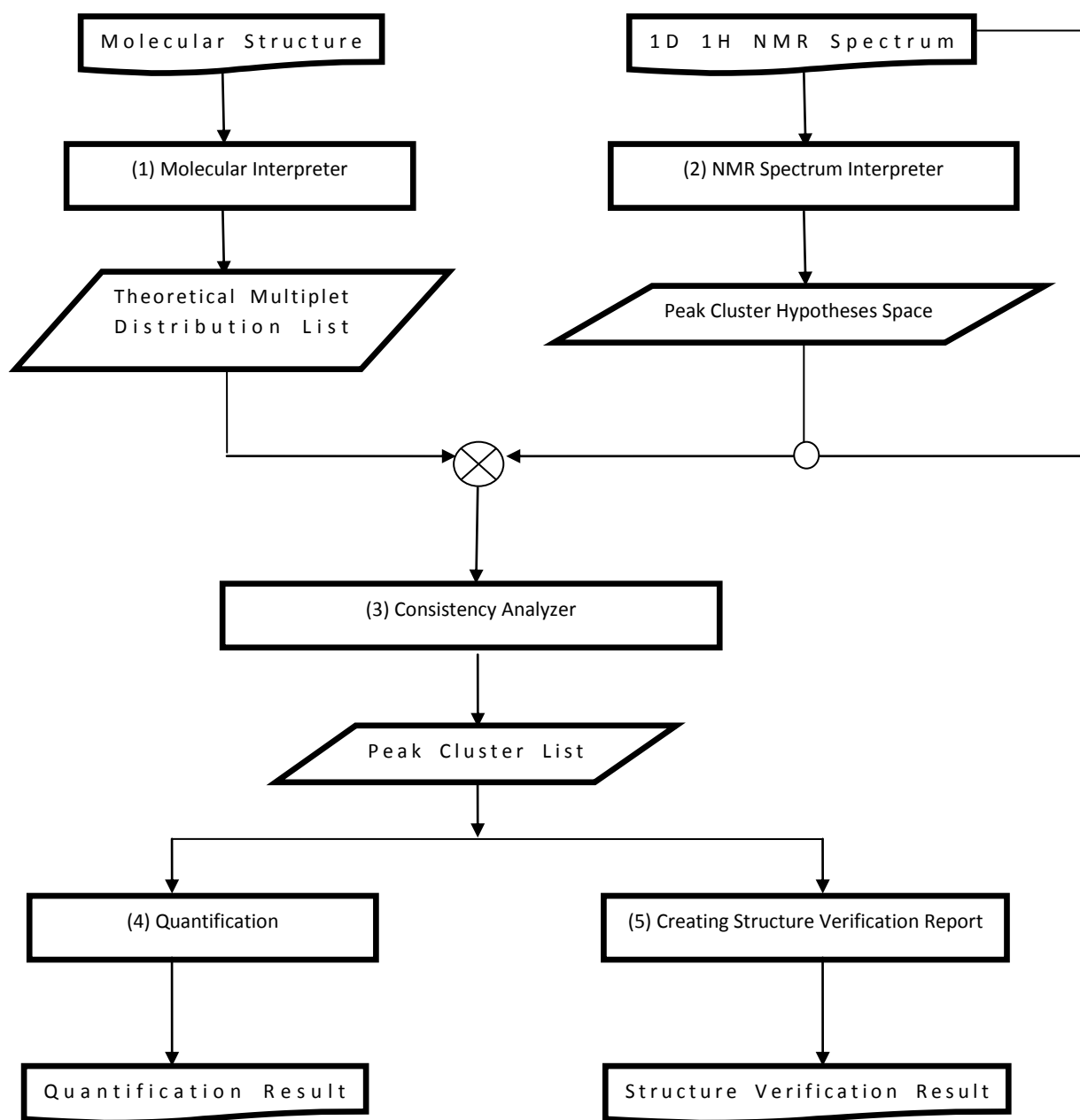
- (1) predict all possible multiplets concerning their distribution in multiplet position and multiplet shape, which could be produced from the molecular structure. Formally, we name the set of predicted multiplets the theoretical multiplet distribution list;
- (2) build all possible peak clusters from the spectrum, and formally name the set of all possible peak clusters as the peak cluster hypothesis space, in which each peak cluster is further assigned an experimental multiplet hypotheses space;
- (3) search for an explanatory peak cluster list from the peak cluster hypothesis space, which could reasonably match all theoretical multiplet distributions. To avoid the inefficient brute-force searching, which has exponential computational complexity, a new heuristic searching approach is designed, which mimics the spectroscopist's interpretation procedure, to find a consistent interpretation in an efficient manner.

Through above three steps, a list of peak cluster hypotheses are selected from the peak cluster hypotheses space, which has a one-to-one or one-to-multi mapping to the theoretical multiplet distribution list.

### **4.1 System Architecture**

The system contains of five modules:

- (1) a molecular interpreter to generate a theoretical multiplet distribution list,
- (2) a NMR spectrum interpreter to generate an experimental peak cluster hypotheses space,
- (3) a consistency analyzer - a searching routine to find an explanatory peak cluster list,
- (4) a quantification module,
- (5) and a structure verification report generator.



Module (1): Molecule Structure Generator

Module (2): Experimental Peak Cluster Hypotheses Generator

Module (3): Searching Routine to Find an Experimental Peak Cluster List

Module (4): Quantification Module

Module (5): Structure Verification Report Generator

**Fig 25 System Flow Chart**

The data processing flow chart of the system is shown in Fig 25. From top to bottom, first, the 2D molecular structure is fed into Module (1) to compute a list of theoretical multiplet distributions. Specifically, it includes identifying functional groups from the molecule, predicting their chemical shifts and their fluctuant ranges, predicting their multiplicities, predicting their coupling constants and their fluctuant ranges, predicting their average signal line widths, predicting the existence of their satellite peaks, etc. Abreast with the molecule interpretation, an experimental 1D <sup>1</sup>H NMR spectrum is fed into Module (2) to build an experimental peak cluster hypotheses space. The experimental peak cluster hypotheses space contains all possible independent peak clusters interpretable from the spectrum. In addition, the experimental multiplet hypotheses space is built for each peak cluster hypothesis to describe all possible experimental first-order multiplets interpretable from the peak cluster hypothesis. Then, the experimental peak cluster hypotheses space, the theoretical multiplet distribution list and the input NMR spectrum are fed into Module (3) to find an experimental peak cluster hypotheses list, which is consistent with the theoretical multiplet distribution list. As the output of Module (3), the peak cluster hypotheses list is produced, and is fed both into Module (4) and Module (5). Finally, the Module (4) selects a peak cluster hypothesis from the peak cluster hypotheses list for quantification, and Module (5) uses the peak cluster hypotheses list to create a structure verification report.

## 4.2 Molecular Interpreter

The Molecular Interpreter contains two modules: (1) a module to identify the theoretical multiplets (chemically equivalent functional groups), (2) a module to assign the distributions through the theoretical multiplets by estimating the theoretical multiplet's chemical shift and its fluctuant range, the theoretical multiplet's multiplicity, the theoretical multiplet's coupling constants and their fluctuant ranges, the theoretical multiplet's average signal line width, and the existence of the theoretical multiplet's satellites.

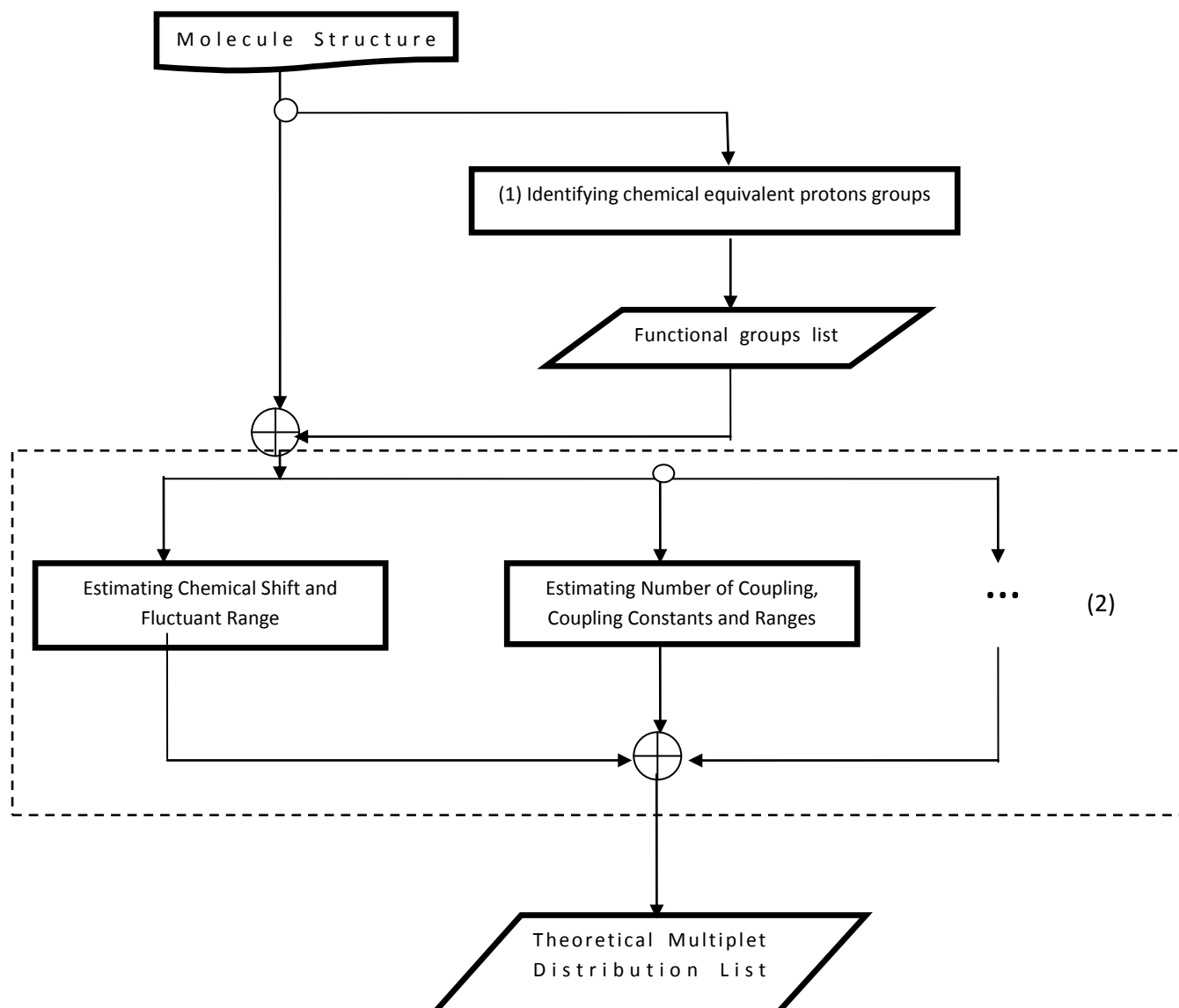
There are three major differences between our Molecular Interpreter and the previous molecular structure interpretation routines, which makes our approach consistent with that of human spectroscopists (See 2.2.1.1 and 2.2.2.1). Specifically, they are:

- a. In addition to the prediction of the chemical shift of the theoretical multiplet, a chemical shift range is also estimated for the theoretical multiplet. In addition to the prediction of the coupling constants of the theoretical multiplet, a coupling constant range is also estimated for each coupling constant of the theoretical multiplet. Known from 2.2.1.1 and 2.2.2.1, the accurate chemical shift prediction in proton NMR is difficult. Therefore, the spectroscopist only relies on loose chemical shift ranges for structure verification tasks. This implies that the prediction of the accurate chemical shift is unnecessary. As a result, in our system a loose chemical shift range is estimated for each theoretical multiplet in addition to the prediction of its chemical shift, to relief the requirement on chemical shift prediction accuracy. The estimation of chemical shift ranges is easily acquired by programming the chemical shift tables given in the NMR/ organic chemistry text books (Keeler, 2005) (Solomons, et al., 2003). Similarly, the accurate prediction of the coupling constant requires

the precise 3D structure of the molecule (See 2.2.1.1). To solve the problem, both the coupling constant and its coupling constant fluctuant range are estimated instead. This is achieved by querying the coupling constant chart (e.g. Karplus correlation chart) in the NMR/organic chemistry text books (Solomons, et al., 2003) (Keeler, 2005). With a chemical shift range and coupling constant ranges assigned, the theoretical multiplet is formally termed by its theoretical multiplet distribution.

- b. Beyond the NMR knowledge used in current structure verification system, additional NMR knowledge is calculated and assigned to the theoretical multiplet distribution. Specifically, the peak's line width and the satellite properties are estimated for each theoretical multiplet distribution. The value of the line width mainly depends on the protons' local structure properties e.g. connected bond types, bond angles, etc, which can be looked up in the organic chemistry book (Solomons, et al., 2003). Satellite peaks are observed in 1D  $^1\text{H}$  NMR spectra if a proton is directly bonded to a nuclear spin  $\frac{1}{2}$  particles e.g.  $^{13}\text{C}$ ,  $^{15}\text{N}$ , etc. The amplitudes and the positions of satellite peaks depend on the abundance and type of the involved nuclear spin  $\frac{1}{2}$  particles. Details on the subject can be found in NMR text book e.g. (Keeler, 2005).
- c. The proton's geometric symmetry in the 2D structure of the molecule is used to identify theoretical multiplet distribution (chemically equivalent functional groups). In contrast, in current structure verification systems, chemically equivalent proton groups are identified by grouping the protons with the same predicted chemical shift together. This requires using the database approach to predict chemical shifts (see 2.2.1.1 for detail). To be consistent with human approach, in our system, we identify the functional group by checking protons' geometric symmetry. Specifically, the 2D structure's geometric symmetry is converted to a graph searching problem (Sedgewick, 2001), where the molecular 2D structure is represented by a graph. This is followed by building a search tree starting from each proton. Then the geometric symmetry is identified by seeking the searching trees which contain the same tree structure.

The data processing flow chart of Molecular Interpreter Module is shown in Fig 26. From top to bottom, first the 2D structure of the molecule is fed into Module (1) to group the chemically equivalent protons (theoretical multiplets). Next, both the theoretical multiplet list and the input molecule's 2D structure are fed into Module (2) to estimate the theoretical multiplet's NMR properties such as chemical shift and fluctuant range, number of couplings (multiplicity), coupling constant and its fluctuant range, line width, satellite peaks in parallel. As an output, a list of theoretical multiplet distributions is built.



Module (1): grouping the chemically equivalent protons (theoretical multiplets)

Module (2): estimating the theoretical multiplet's NMR properties

**Fig 26 Molecular Interpreter Module Flow Chart**

Many commercial programs have been developed to extract NMR knowledge from 2D molecular structures. To simplify the problem, our industry partner, who cooperates with us on the project, supplies us a commercial program named Perch (PERCH, 2005) to help our work. As we explained above, Perch, similar to other programs, supplies most NMR information we need e.g. chemical shift value, coupling constant value, etc, except the estimation of chemical shift ranges and coupling constant ranges. Therefore, our industry partner implements a program to supplement the functionality of Perch by reading in the predicted chemical shift values and coupling constant values

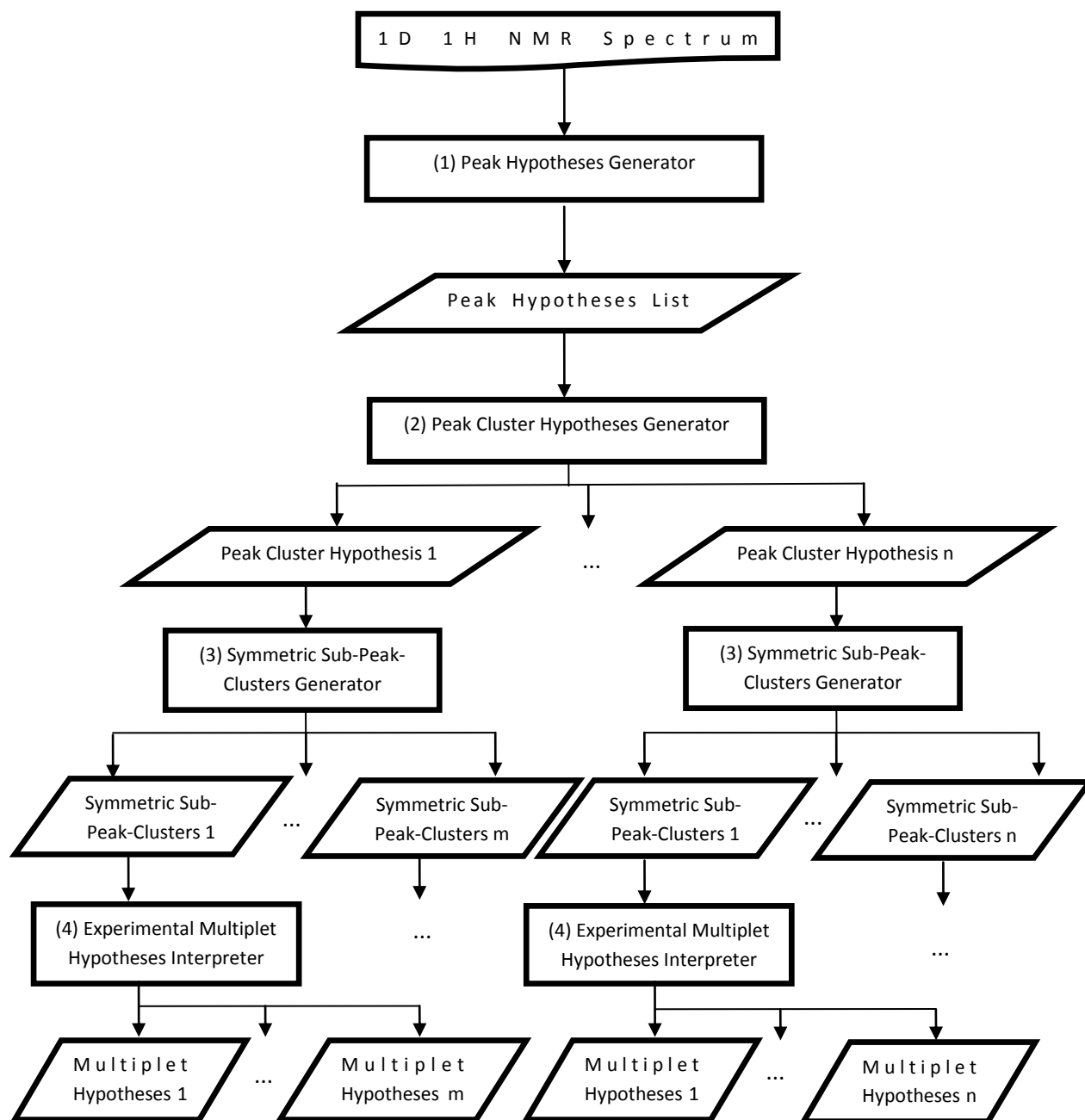
and extending them to ranges by defining sufficiently large intervals to be assured to cover the NMR signals acquired in experiments. The evaluation result shows that this practical shortcut works well in practice (see 6.2).

### 4.3 NMR Spectrum Interpreter

NMR Spectrum Interpreter contains four modules (see Fig 27): (1) a peak hypotheses generator, (2) a module to group peaks into all possible well-separated peak clusters, (3) a module to further group peaks in a peak cluster into all possible symmetric sub-peak-clusters with peak positional symmetry, and (4) a module to interpret a symmetric sub-peak-cluster into all possible first order multiplets by Pascal triangle analysis. The entirety of all these peak clusters produced in (2) is denoted as peak cluster hypotheses space. The entirety of all first-order multiplets for a peak cluster hypothesis is denoted as the experimental multiplet hypothesis space of the peak cluster hypothesis.

The data processing flow chart of NMR Spectrum Interpreter is shown in Fig 27.

From top to bottom, first, the input 1D  $^1\text{H}$  NMR spectrum is fed into Module (1) to detect all possible peak hypotheses, and to assign a confidence score to each possible peak hypothesis. Next, the peak hypotheses are fed into Module (2) to build all possible well-separated peak cluster hypotheses. After that, peak cluster hypotheses are in turn fed into Module (3) to build all possible symmetric sub-peak-clusters of the peak cluster with positional symmetry. After that, the list of all symmetric sub-peak-clusters is fed into Module (4) one by one to cut them into all possible first-order multiplets. Correspondingly, their multiplicities and coupling constants are estimated through the first order multiplet analysis. As a result, both the experimental peak cluster hypotheses space, and the experimental first-order multiplet hypothesis spaces of each peak cluster hypothesis are built.



Module (1): Peak Hypotheses Generator

Module (2): Peak Cluster Hypotheses Generator

Module (3): Symmetric Sub-Cluster-hypothesis Generator

Module (4): Experimental Multiplet Hypotheses Interpreter

**Fig 27 NMR Spectrum Interpreter Module Flow Chart**

### 4.3.1 Peak Hypothesis Generator (Deconvolution Method + Derivative Method)

A new peak picking approach is designed to avoid missing peaks. As mentioned in 2.2.1, the current peak picking routines suffer from a tradeoff between missing peaks and introducing noise and artifact peaks. To avoid missing real peaks, the traditional derivative-based and deconvolution-based peak picking techniques are combined to detect all possible peak positions from the input 1D  $^1\text{H}$  NMR spectrum. Here, the deconvolution approach means the techniques to continually fit and subtract peaks from the spectrum using predefined peak shape. Note, in proton NMR domain, the predefined peak shape often chooses to have Gaussian shape, Lorentz shape, or mixture of both. The derivative approach means the techniques to identify peaks by calculating local maximums/minimums in the first derivative and second derivative transforms of the spectrum. Specifically, the deconvolution routine is used two times. In the first deconvolution iteration, the prominent peak shapes are extracted from the spectrum, and the residual spectrum is used to automatically determine the spectrum baseline. In the second deconvolution iteration, the prominent peaks, which are significantly bigger than the baseline, are captured by the spectrum. After that, the derivation routine is used to capture the small indistinctive peaks near prominent peaks. However, this approach introduces vast noise peaks and artifact peaks. This introduces unnecessary computational complexity in the later structure verification process. Note, to distinct real peaks from the picked peaks, we uniformly name picked peaks as peak hypotheses. To reduce the disturbance of the noise and artifact peaks, a human-mimicking mechanism is designed to rank the peak hypotheses. Specifically, a confidence score is assigned to a peak hypothesis based on the peak hypothesis's NMR properties e.g. the peak hypothesis's amplitude, baseline level, overlap level, line width, symmetry, etc. With this approach, "the high confidence peak hypothesis first" principle could be used to evaluate peak hypotheses efficiently in the later stage. To give readers a clearer picture, we present peak picking flow chart in Fig 28.



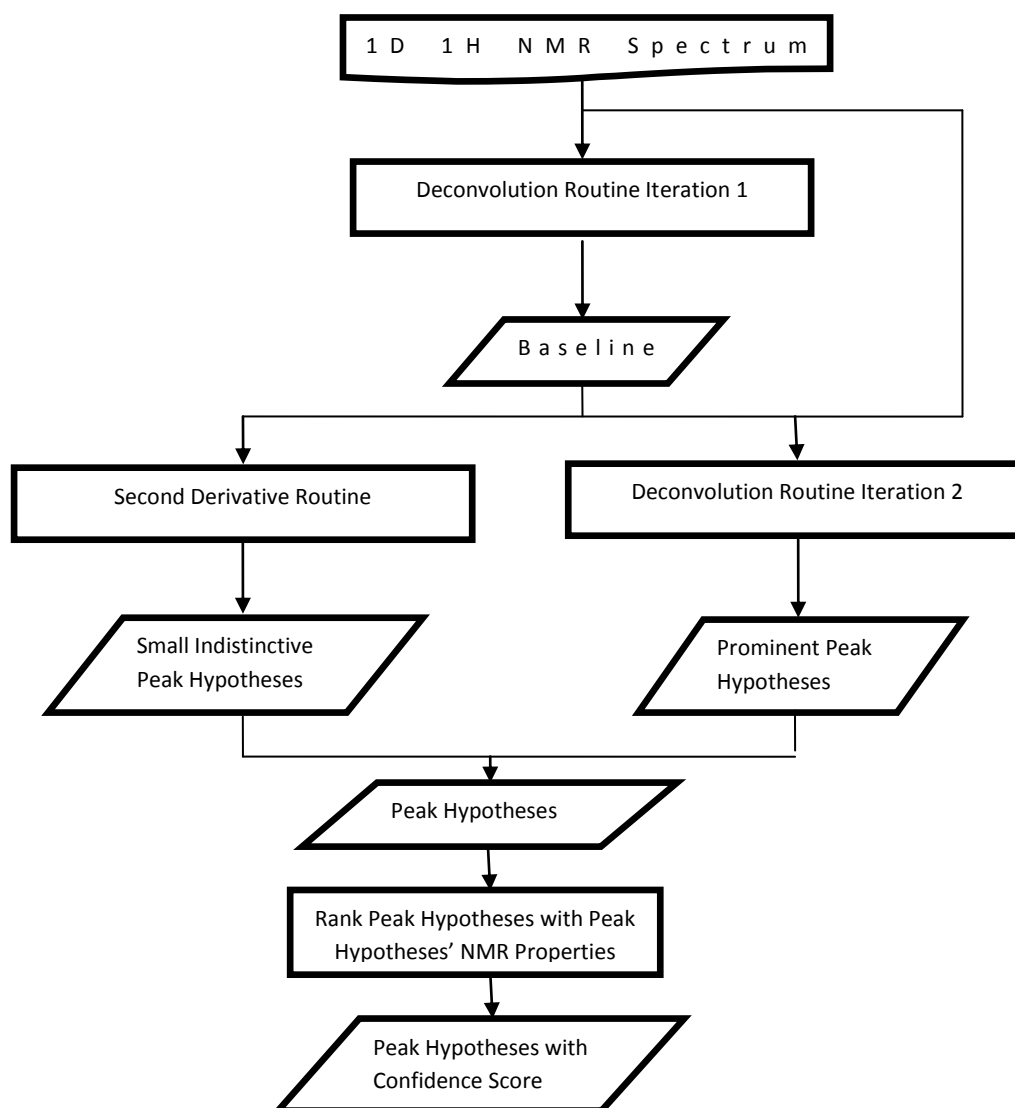


Fig 28 Peak Pick Routine Flow Chart

### 4.3.2 Peak Cluster Hypothesis Generation

The approach to compute peak cluster hypotheses space is to mimic the spectroscopists' approach described in 2.1.3. Specifically, the approach is represented as below:

Starting from left to right (high field to low field) of the spectrum:

- (1) Select a point on the x-axis which has an amplitude around the spectrum baseline, and use this point as the left boundary of a new basis peak cluster.
- (2) Start from the point to continually move right to cover as many peaks as possible.

- (3) Stop when the movement touches another point on the x-axis which has an amplitude around the spectrum baseline. This new point is used as the right boundary of the basis peak cluster. As a result, a new basis peak cluster is identified.
- (4) Repeat (1)-(3) until all basis peak clusters are identified.
- (5) Enumerate all subsets of the set of basis peak clusters to construct the peak cluster hypotheses space, wherein each subset defines a new peak cluster hypothesis.

### 4.3.3 Experimental Multiplet Hypothesis Interpreter

Peaks in a given peak cluster hypothesis are grouped into all possible positional-symmetric peak groups. This is followed by the first order multiplet (Pascal Triangle) interpretation on each of these peak groups. During the first-order multiplet analysis, *all* possible multiplet interpretations are extracted from the given peak group. (Note, this is directly implementable by applying the divide and conquer strategy (Sedgewick, 1997), and this is due to the recursive nature of the first-order multiplet analysis (Golotvin, et al., 2002)). Formally, we call the ensemble of all possible multiplet interpretations extractable from a peak cluster hypothesis the experimental multiplet hypotheses space of the peak cluster hypothesis. Note, although the same criterion of the positional symmetry is utilized to group peak clusters for later first-order multiplet analysis in the previously designed automatic structure verification systems, only the most “likely” positional-symmetric peak cluster were generated. In contrast, in our system, *all* possible positional-symmetric peak groups are attempted to be extracted from the spectrum. Therefore, the number of all possible positional-symmetric peak groups is normally two or even more orders bigger than that of the most “likely” symmetric peak clusters generated by the previous structure verification system. This introduces additional computational complexity to the consistency analysis module later. Also, the attempt to extract all possible multiplet interpretations from the positional-symmetric peak group further expands the search space, and therefore increases the difficulty. To address this problem, another spectroscopist-mimic mechanism is introduced to rank both the peak cluster hypotheses space and the experimental multiplet hypotheses space so that the human-mimicking heuristic search could be based on to increase the search efficiency (see 4.4 for detail). Specifically, both the peak cluster hypotheses and the multiplet hypotheses are further scored by their signal intensity (e.g. integration of the peak cluster hypothesis, total peak amplitudes in the multiplet hypothesis), signal complexity (e.g. clearness of first order multiplet patterns in the peak cluster hypothesis, number of peaks in the multiplet hypothesis), signal symmetry (e.g. symmetry upon peaks’ amplitude in both the peak cluster hypothesis and the multiplet hypothesis).

## 4.4 Consistency Analyzer - Searching Consistent Peak Cluster List

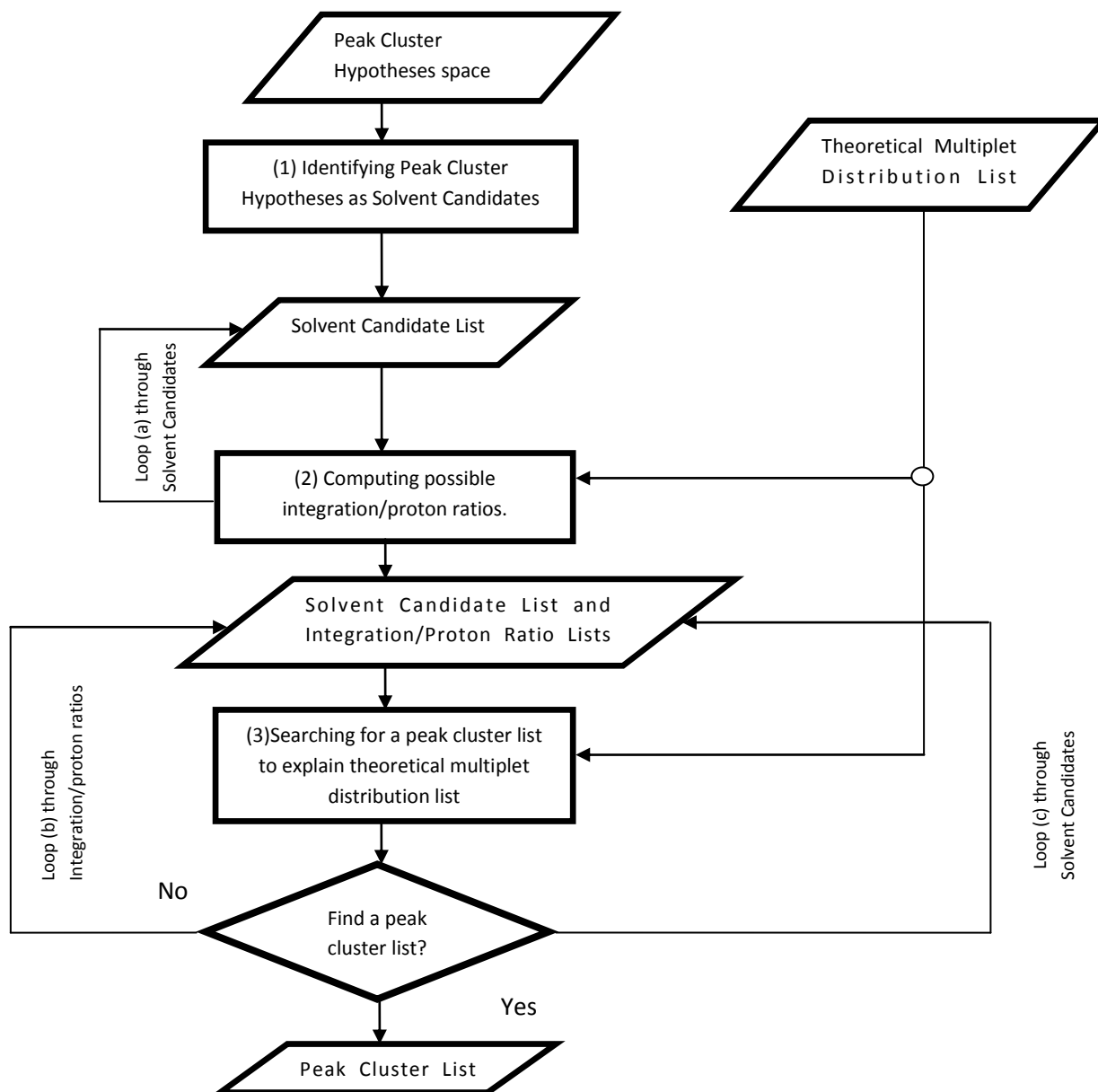
With both peak cluster hypotheses space and theoretical multiplet distribution lists built, another module is implemented to identify a reasonable match between them. Specifically, a peak cluster list needs to be selected from the peak cluster hypotheses space, which can reasonably explain the

theoretical multiplet distribution list with their chemical shift, multiplicity, proton number, and connectivity. Following spectroscopists' structural verification logic, two sub-tasks need to be done before the searching can start. They are: (1) to identify peak cluster hypotheses which are the signal of the solvent, and (2) to compute the integration proton ratio (the normalization factor). The signals of the solvent are not produced from the molecule, and therefore should be deleted from the peak cluster hypotheses space to avoid disturbing the matching procedure. The integration proton ratio is used to estimate the proton numbers of peak cluster hypotheses, which will be further applied in the matching procedure. Actually, they act as one of the most important information in the matching process (see 2.1.3.2 and 2.1.3.3 for detail). Then, with peak clusters representing the signal of solvent excluded and the proton numbers of the peak cluster hypotheses computed, the peak cluster hypotheses with their associated chemical shifts, their proton numbers, and their associated multiplet hypotheses spaces are fed into the searching routine to match them with theoretical multiplet distributions.

Both solvent detection and proton number computation techniques are empirical and hence inaccurate. It is often impossible to uniquely identify the solvent signals, and to uniquely identify the integration proton ratios (see 4.4.1 and 4.4.2 for discussion). Therefore, a list of all possible solvent signals and their lists of all possible integration proton ratios are computed instead. In fact, the calculation of the list of possible integration proton ratios depends on the identification of the solvent signals. Different choices of the solvent signals will induce the different possible integration ratio list. Reversely, the computation of the integration ratio list in turn helps to identify the correctness of the solvent signals (see 4.4.2 later). Specifically, to compute an integration proton ratio list, the solvent signals from the solvent candidate list are sequentially fed into the integration proton ratio computing routine, in which reasonable integration proton ratios are computed and stored together with the solvent signal. In case of no reasonable integration proton ratio computable from the solvent candidate list, it gives a hint to the system that the solvent signal is wrongly identified, and therefore should be deleted from the solvent candidate list. In this way, by going through all solvent signal candidates, all reasonable solvent signals and their corresponding integration proton ratio lists are recorded for further use. In the worst case, after going through all possible solvent signal candidates, there is still no reasonable integration proton ratio. Then the system gathers enough evidence to show the inconsistency between the peak cluster hypotheses space and the theoretical multiplet distribution list, and therefore calls the verification report generation module to produce an inconsistent conclusion.

When a reasonable solvent signal and the corresponding integration proton ratio list exist, the integration proton ratio in the list is sequentially fed into the searching module together with the solvent signal. Together with the peak cluster hypotheses space and the theoretical multiplet distribution list, they are used to search for a consistent peak cluster list. While the solution is found, the system concludes that the structure verification result is consistent, and the system will call the reporting module to produce a consistency report. In case a consistent peak cluster list cannot be found, it gives the system another hint that either solvent signal or integration proton number ratio could be wrongly calculated, and therefore another restored solvent signal and/or integration proton ratio are selected and fed into the searching module. As the result, the new iteration starts. If after iterating all these solvent signal candidates and integration proton ratio candidates, the system still cannot find a consistent peak cluster list, an inconsistent conclusion is produced. In case that iterating through all possible solvent signal candidates and integration proton ratio candidates takes

too much time, the system stops in the middle and produces a conclusion as “give-up”. Note, we explain the solvent detection technology in detail in 4.4.1, and the approach to compute integration proton ratio in 4.4.2. The detailed explanation of searching process itself is in 4.4.3. To illustrate the relationship among solvent detection, integration proton ratio computation and searching modules, and to describe the back tracking mechanisms, we show the flow chart of Searching Consistent Peak Cluster List Module in Fig 29.



Module (1): Identify peak cluster hypotheses as solvent candidates

Module (2): Compute possible integration proton ratios

Module (3): Search module: with a given solvent signal and an integration proton ratio, search a peak cluster list which matches with theoretical multiplet distributions.

**Fig 29 Searching Peak Cluster List Module**

#### 4.4.1 Solvent Detection

In this section we introduce the technology to identify solvent signals from the spectrum. Several solvents could be used in  $^1\text{H}$  NMR experiment. As we have explained in 2.1.1, we focus on the techniques to detect NMR signals from the proton residual of DMSO and  $\text{H}_2\text{O}$  in DMSO in the thesis.

##### 4.4.1.1 DMSO detection

Several empirical rules are used by spectroscopists to identify signals from DMSO. They are:

- (1) The DMSO signal is likely to appear at chemical shift position 2.5ppm.
- (2) The DMSO signal often shows the multiplicity of a pentet or a doublet of triplet.
- (3) The DMSO signal often has the proton numbers that are not proportional to the proton numbers of the NMR signals from the molecule.

These rules are imprecise and ambiguous, and therefore difficult to be used to precisely identify the solvent signal. To faithfully reflect the implied uncertainty of these rules, the system utilizes them to calculate a DMSO likelihood score for each peak cluster hypothesis in the peak cluster hypotheses space. This is followed by selecting a subset of the peak cluster hypotheses space as the solvent signal candidate list, in which every peak cluster hypothesis has a high DMSO likelihood score. With this strategy, the system reduces the risk of making wrong DMSO signal identification by relaxing the DMSO identification problem to the problem of the identification of a set of likely DMSO signals. This in turn makes it easy to embed the DMSO signal identification routine into the human mimicking hypothesis-driven problem solving framework, where the matching between peak clusters and theoretical multiplet distributions supplies additional information to further discriminate the DMSO signals in the DMSO candidate list (see Fig 29).

Specifically, three measurement scores are used to estimate the DMSO likelihood score, which are corresponding to three rules defined above. They are the *chemical shift measurement score*, the *multiplicity measurement score* and the *proton number measurement score*. The DMSO likelihood score is defined as the multiplication of the three factors. Formally, we show it in formula (2).

$$\begin{aligned} \text{DMSO likelihood score} = & \\ & \text{Chemical Shift Measurement Score} \times \text{Multiplicity Measurement Score} \times \\ & \text{Proton Number Measurement Score} \end{aligned} \quad (2)$$

To compute the *chemical shift measurement score*, the experimental chemical shift of the peak cluster hypothesis is measured and compared with the expected DMSO chemical shift (2.5ppm). Specifically, a DMSO chemical shift interval of 2.0-3.0ppm, which covers 2.5ppm position, is defined and used as the reference to evaluate the experimental chemical shifts of the peak cluster hypothesis. Precisely, a human mimicking rule is implemented to give the score. It is:

**If (Chemical Shift  $\in$  2.25-2.75ppm), then Chemical Shift Measurement Score = 1,  
 Else If ((Chemical Shift  $\in$  2.0-2.25ppm) || (Chemical Shift  $\in$  2.75-3.0ppm)),  
     **then Chemical Shift Measurement Score = 0.5,**  
**Else Chemical Shift Measurement Score = 0.****

A similar technique is used to compute the *multiplicity measurement score*. A list of possible multiplicities of DMSO residual proton signals are built and utilized as the reference to evaluate the experimental multiplicity of the peak cluster hypothesis. Note, a pentet or a doublet of triplet is the most likely multiplet patterns of the DMSO signal. There are other possible multiplet patterns as well. For example, while the proton residual signal of the DMSO is significant smaller than the signals from the molecule, it could appear as a triplet or a doublet with the peaks at the multiplet boundary submerging into the spectrum noise. To model this flexibility, another human-mimicking rule is implemented to calculate the multiplicity measurement score. It is:

**If (the multiplet hypotheses space of the peak cluster hypothesis contains a pentet or a doublet of triplet),  
     then Multiplicity Measurement Score = 1,  
 Else If (the multiplet hypotheses space of the peak cluster hypothesis contains a triplet),  
     then Multiplicity Measurement Score = 0.5,  
 Else If (the multiplet hypotheses space of the peak cluster hypothesis contains a doublet),  
     then Multiplicity Measurement Score = 0.25,  
 Else Multiplicity Measurement Score = 0.**

The calculation of the *proton number measurement score* requires the proton number of the peak cluster hypothesis computed, which itself is computed by dividing the integration of the peak cluster hypothesis by the integration proton ratio (see loop indicator in Fig 29). Note, this is another evidence to show the integration proton ratio calculation could be used to help identifying the solvent signals. With the proton number of the peak cluster hypothesis known, another human mimicking rule is implemented to calculate the proton number measurement score. Specifically, it is:

**If (proton numbers close to integer value),  
     then Proton Number Measurement Score = 0.5,  
 Else Proton Number Measurement Score = 1.**

Note, in this rule, we do not give 0 as the proton number measurement score. This is due to the fact that even the proton number of the peak cluster hypothesis is close to a “reasonable” integer value, it does not give enough evidence to prove that the peak cluster hypothesis is the signal from the molecule. But it does supply some information to reduce the peak cluster hypothesis’ likelihood to be the signal of DMSO residual protons, and therefore a “softer” score 0.5 is used instead of “severe” score 0.

With the three computable rules, the DMSO likelihood score is computed and assigned for each peak cluster hypothesis. To limit the size of DMSO signal candidates, a subset of peak cluster hypotheses space is selected as the DMSO signal candidate list, where each peak cluster hypothesis has a high DMSO likelihood score. Note, with this list, the DMSO identification process is efficiently embedded into the system's hypothesis-driven human mimicking framework, in which a back-tracking mechanism is utilized to reselect alternative DMSO signal candidates from the list as the DMSO signal to avoid the mistake of DMSO identification. This implementation is highly consistent with the DMSO identification approach that human spectroscopists adopt.

#### 4.4.1.2 H<sub>2</sub>O Detection

The identification of the H<sub>2</sub>O signal is a challenging task. This is due to its wide chemical shift range and its varying signal shape. Therefore, spectroscopists rely on several inexact rules to approximately select some NMR signals as the likely H<sub>2</sub>O signals. To further reduce the ambiguity on the identification of the H<sub>2</sub>O signal, spectroscopists turn to rely on the matching analysis between the spectrum and the molecular structure to validate the eligibility of the H<sub>2</sub>O signal candidates. Closely following this human strategy, the system implements several weak rules to compute a H<sub>2</sub>O likelihood score for each peak cluster hypothesis, and to utilize these scores to select a subset of peak cluster hypotheses space as the H<sub>2</sub>O signal candidate list. The H<sub>2</sub>O signal candidates are sequentially fed into the searching module in Fig 29 to search for a consistent peak cluster list. The existence of the consistent peak cluster list supplies the additional evidence to validate the correctness of the selected H<sub>2</sub>O signal candidate. In contrast, the nonexistence of this list indicates the impropriety of the selected H<sub>2</sub>O signal candidate, and therewith brings on the deletion of the H<sub>2</sub>O signal candidate from the H<sub>2</sub>O candidate list (see Fig 29 for the flow chart).

Some weak rules, spectroscopists use to identify H<sub>2</sub>O signal candidates, are:

- (1) The H<sub>2</sub>O signal often appears in the chemical shift range of 3.0-4.9ppm.
- (2) The H<sub>2</sub>O signal often has broad signal shapes.
- (3) The H<sub>2</sub>O signal often appear as a singleton (single broad peak), but it is also likely that it appear as 2 or 3 heavily overlapped (non-well-separated) peaks.
- (4) Peaks from the H<sub>2</sub>O signal are not well-separated peaks.
- (5) The H<sub>2</sub>O signal often has the proton numbers, which is not proportional to the proton numbers of the signals from the molecule.

Similar to the approach of identifying the DMSO signal, a H<sub>2</sub>O likelihood score is computed for each peak cluster hypothesis. Specifically, five measurable factors are introduced, while each factor is measured along a weak rule mentioned above. They are chemical shift measurement score, signal width measurement score, peak number measurement score, peak separation measurement score, and proton number measurement score. H<sub>2</sub>O likelihood score is defined as the multiplication of the five factors. Formally, we show it in formula (3).

$$\begin{aligned} \text{H}_2\text{O likelihood score} = & \text{Chemical Shift Measurement Score} \times \text{Peak Width Measurement Score} \times \\ & \text{Peak Number Measurement Score} \times \text{Peak Separation Measurement Score} \times \\ & \text{Proton Number Measurement Score} \end{aligned} \quad (3)$$



To compute the *chemical shift measurement score*, the experimental chemical shift of the peak cluster hypothesis is measured and compared with the expected H<sub>2</sub>O chemical shift range (3.0-4.9ppm). Similar to the process of computing the DMSO chemical shift measurement score, a group of non-intercrossed bins are defined in the chemical range of 3.0-4.9ppm, and a chemical shift measurement score is assigned to a peak cluster hypothesis based on which bin its chemical shift falls into. Note, a uniform bin partition is implemented into the system to compute the chemical shift measurement score for H<sub>2</sub>O and DMSO. But in principal, any other non-uniform bin partitions, which could better model human spectroscopists' subjective belief, could be used to replace the uniform partition to improve the system performance. Specifically, the rule we adopt is:

**If (Chemical Shift  $\in$  3.50-4.40ppm), then Chemical Shift Measurement Score = 1,  
 Else If ((Chemical Shift  $\in$  3.0-3.5ppm) || (Chemical Shift  $\in$  4.4-4.9ppm)),  
 then Chemical Shift Measurement Score = 0.5,  
 Else Chemical Shift Measurement Score = 0.**

To compute the *peak width measurement score*, a similar rule is implemented. It is:

**If (half height width > 1.2 Hz), then Peak Width Measurement Score = 1,  
 Else If ((half height width < 1.2Hz) && (half height width > 0.9Hz)),  
 then Peak Width Measurement Score = 0.5,  
 Else If ((half height width < 0.9Hz) && (half height width > 0.5Hz)),  
 then Peak Width Measurement Score = 0.25,  
 Else Peak Width Measurement Score = 0.**

Analogically, the rule to compute the *peak number measurement score* is:

**If (number of peak = 1), then Peak Number Measurement Score = 1,  
 Else If (number of peak = 2), then Peak Number Measurement Score = 0.5,  
 Else If (number of peak = 3 || number of peak = 4 ),  
 then Peak Number Measurement Score = 0.25,  
 Else Peak Number Measurement Score = 0.**

To compute the *peak separation measurement score*, the overlapping level among signal peaks are measured. Note, the peak cluster hypothesis, which only contains a single peak, is given a *peak separation measurement score* as 1 (this is clear since there is no overlap in a single peak pattern). The following rule is used to measure the overlapping level among peaks and give the peak separation measurement score:

**From the most left peak position to the most right peak position of the peak cluster hypothesis, the system scans for the maximum amplitude and the minimum amplitude.  
 Overlapping Indicator = minimum amplitude/ maximum amplitude,**

**If (*Overlapping Indicator* > 0.7), then *Peak Separation Measurement Score* = 1,**  
**Else If (*Overlapping Indicator* > 0.4 && *Overlapping Indicator* < 0.7),**  
**then *Peak Separation Measurement Score* = 0.5,**  
**Else If (*Overlapping Indicator* > 0.1 && *Overlapping Indicator* < 0.4),**  
**then *Peak Separation Measurement Score* = 0.25,**  
**Else *Peak Separation Measurement Score* = 0.**

The calculation of the *proton number measurement score* requires the proton number of the peak cluster hypothesis to be computed, which itself is computed by dividing the integration of the peak cluster hypothesis by the integration proton ratio (see loop indicator in Fig 29). With the proton number of the peak cluster hypothesis known, another human mimic rule is implemented to calculate the *proton number measurement score*. Specifically, it is:

**If (proton numbers close to integer value), then *Proton Number Measurement Score* = 0.5,**  
**Else *Proton Number Measurement Score* = 1.**

Note, in this rule, we do not give 0 as the proton number measurement score. This is due to the fact that even the proton number of the peak cluster hypothesis is close to a “reasonable” integer value, it does not give enough evidence to prove that the peak cluster hypothesis is the signal from the molecule. But it does supply some information to reduce the peak cluster hypothesis’ likelihood to be the signal of H<sub>2</sub>O, and therefore a “softer” score 0.5 is used instead of “severe” score 0.

Relying on the five computable rules, the H<sub>2</sub>O likelihood score is computed and assigned for each peak cluster hypothesis. To limit the size of H<sub>2</sub>O signal candidates, a subset of peak cluster hypotheses space is selected as the H<sub>2</sub>O signal candidate list, where each peak cluster hypothesis has a high H<sub>2</sub>O likelihood score. Note, with this list, the H<sub>2</sub>O identification process is efficiently embedded into the system’s hypothesis-driven human mimicking framework, in which a back-tracking mechanism is utilized to reselect alternative H<sub>2</sub>O signal candidates from the list as the H<sub>2</sub>O signal to avoid the mistake of H<sub>2</sub>O identification. This implementation is highly consistent with the H<sub>2</sub>O identification approach what human spectroscopists adopt.

To illustrate the approach of solvent detection, we represent the solvent detection flowchart in Fig 30.

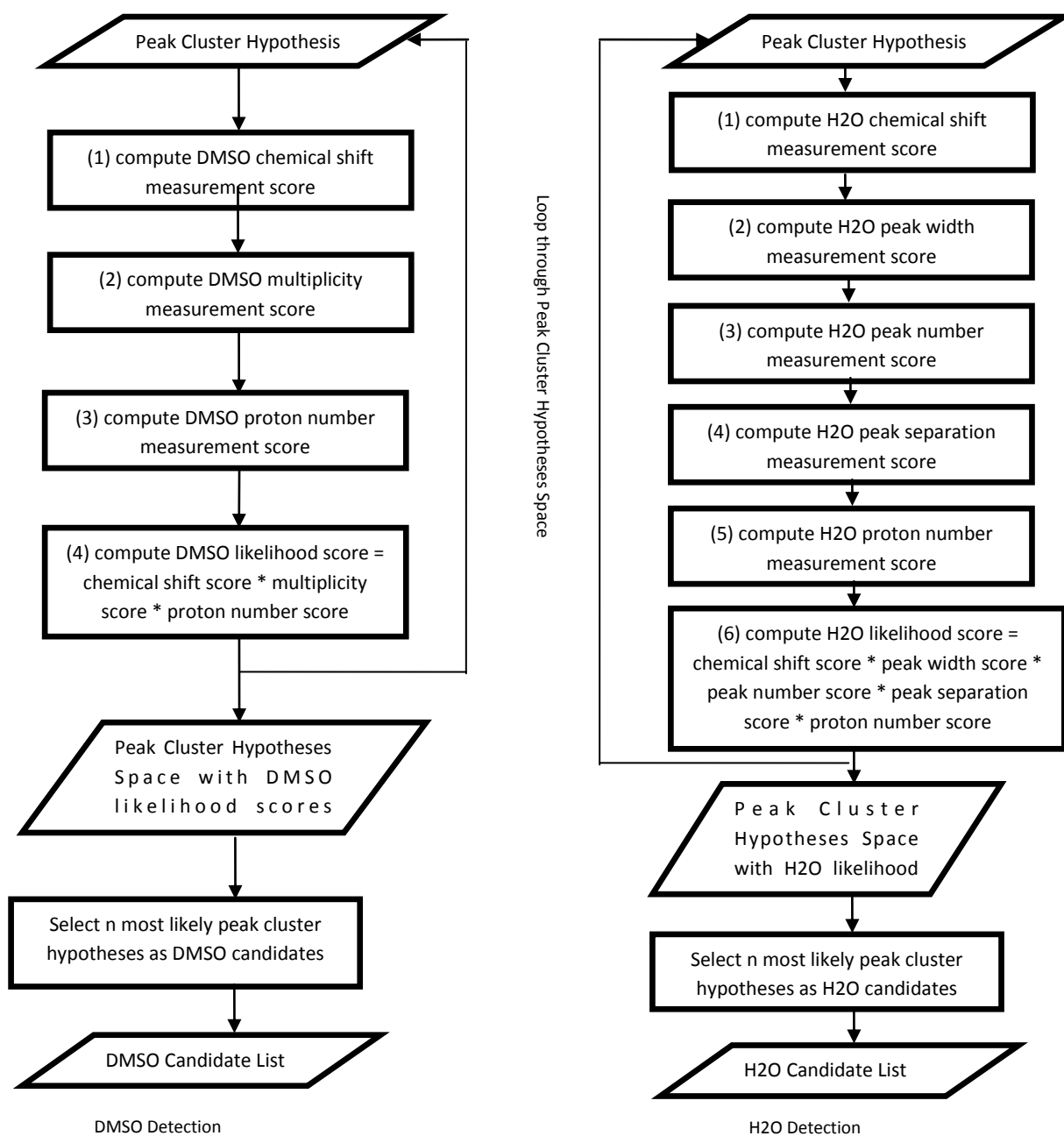


Fig 30 Solvent Detection

#### 4.4.2 Determine Integration Proton Ratio (Integration per Proton)

The Integration proton ratio is a reference for computing the proton number of the peak cluster hypothesis. Its calculation is the prerequisite of the searching module in Fig 29. To compute the

integration proton ratio, the system uses the spectroscopists' integration proton ratio computation process (see 2.1.3.4) as the reference. Specifically, the theoretical multiplet distribution list is utilized to define a set of chemical shift bins (intervals), in which the range of each chemical shift bin is defined by the chemical shift range of a theoretical multiplet distribution. This is followed by classifying peak cluster hypotheses into different chemical shift bins with their experimental chemical shift. When a peak cluster hypothesis is assigned to a chemical shift bin, a chemical shift match between the peak cluster hypothesis and the theoretical multiplet distribution happens. **Formally we denote the match a *chemical shift consistent peak cluster hypothesis theoretical multiplet distribution pair*.** The computation process to identify all *chemical shift consistent peak cluster hypothesis theoretical multiplet distribution pairs* has quadratic computational complexity (assuming that there are  $n$  theoretical multiplet distributions and  $m$  peak cluster hypotheses, totally there are  $m \times n$  peak cluster hypothesis theoretical multiplet distribution pairs needed to be checked for chemical shift consistency). For each *chemical shift consistent peak cluster hypothesis theoretical multiplet distribution pair*, an integration proton ratio is computed. This is done by dividing the proton number of the theoretical multiplet distribution by the integration of the peak cluster hypothesis. Through computing integration proton ratios of all *chemical shift consistent peak cluster hypothesis theoretical multiplet distribution pairs*, the system builds an integration proton ratio list. Note, some *chemical shift consistent peak cluster hypothesis theoretical multiplet distribution pairs* produce the similar integration proton ratio in value, and therefore their integration proton ratios are averaged and only recorded once in the list.

To further discriminate the rational integration proton ratios from the integration proton ratio list, the total proton numbers in the spectrum is computed with the given integration proton ratio and compared to the total proton number in the theoretical multiplet distribution list. This is implemented by integrating all signals in the spectrum except the signals from the solvent candidates, and followed by dividing the integration of the whole spectrum by the given integration proton ratio.

With a rational integration proton ratio, the computed proton number of the spectrum should be comparable to the sum of proton numbers from the theoretical multiplet distribution list. If the computed total proton numbers of the spectrum is significantly deviated from the total proton numbers of the theoretical multiplet distribution list, it gives a strong evidence to deny the correctness of the given integration proton ratio, therefore it causes the deletion of the integration proton ratio from the integration proton ratio list. After iterating the integration proton ratio list, the remaining integration proton ratios, which pass the checking of the total proton numbers, are recorded as the final output integration proton ratio list. To further discriminate upon the output integration ratio list, the system relies on the searching module (in Fig 29) to find a consistent peak cluster hypothesis list, which could reasonably match the theoretical multiplet distribution list. If the list is not existent, it could indicate the wrong integration proton ratio. Therefore a back-tracking mechanism (see the second return back loop in Fig 29) is implemented in the system to select an alternative integration proton ratio to avoid the mistake in integration proton ratio calculation. This implementation supplies a highly reliable integration proton ratio computation procedure, which is highly consistent with the integration proton ratio calculation approach that human spectroscopists adopt.

To illustrate the integration proton ratio computation procedure, we represent its flowchart in Fig 31.

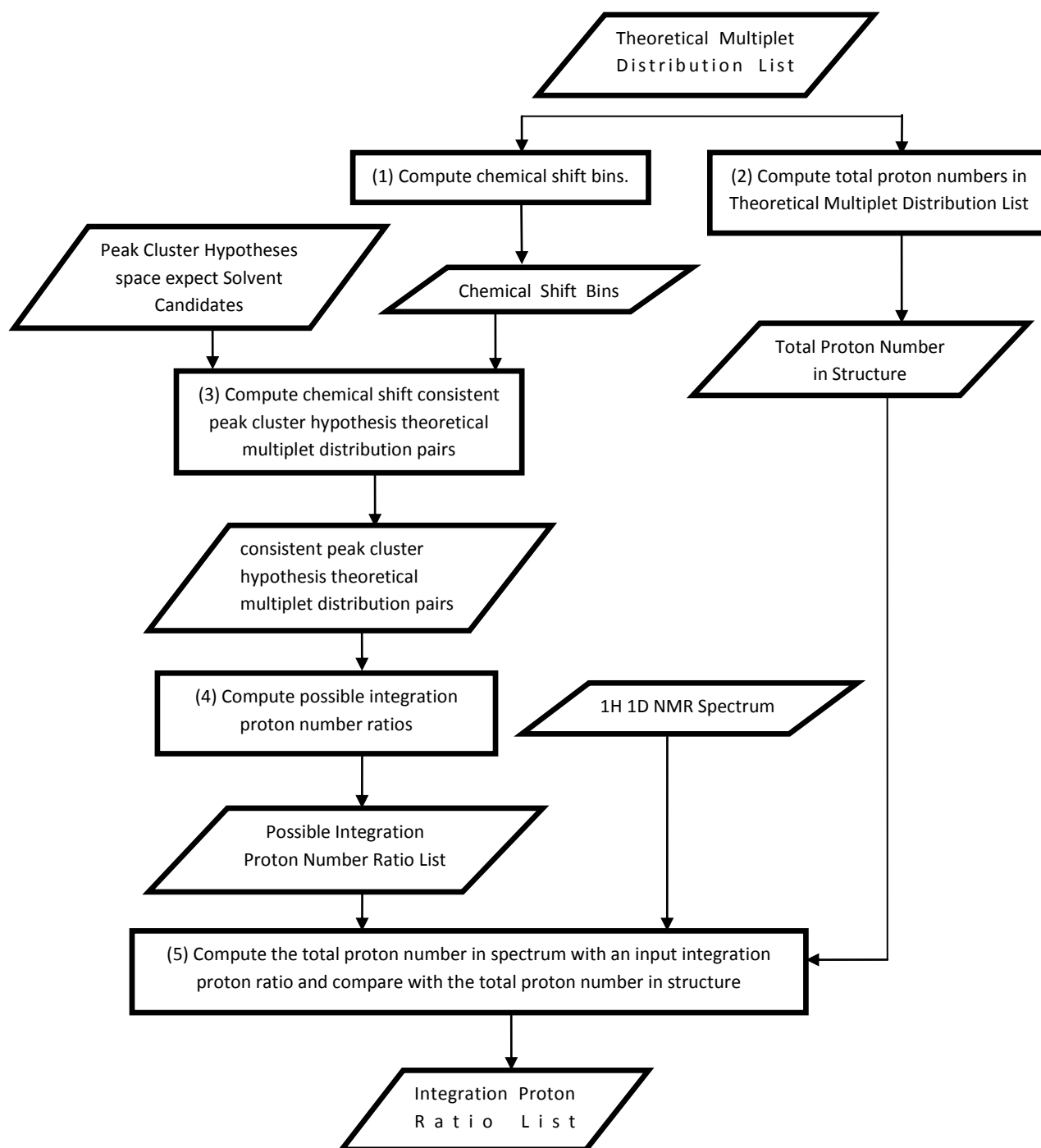


Fig 31 Integration Proton Ratio Computation Flowchart

### 4.4.3 Matching of Experimental Peak Cluster Hypotheses and Structural Multiplet Distributions

In the previous NMR spectrum molecular structural verification system mentioned in 2.2 (Golotvin, et al., 2006), a brute force search has been adopted to match the experimental multiplet list and the theoretical multiplet list with chemical shift. This is followed by a Monte Carlo routine to further optimize the match with proton number and multiplicity. The (primitive) brute force searching strategy is feasible under the previous system's problem setup. This is due to the fact that chemical shift is a NMR property assigned to individual multiplets, and therefore a brute force searching for a match between the experimental multiplet list and the theoretical multiplet list only requires pair-wise comparisons on chemical shift between the experimental multiplet list and the theoretical multiplet list. The computational complexity of complete pair-wise comparisons is quadratic (assuming that there are  $n$  experimental multiplets and  $m$  theoretical multiplets, then there are totally  $n*m$  experimental-theoretical multiplet pairs needed to be compared towards their chemical shifts). However, the feasibility of the brute-force searching is at the expense of ignoring other useful NMR information, and therefore producing a wrong match. The accuracy and the efficiency of stochastic optimization e.g. Monte Carlo optimization rely on the reasonable selection of the starting point of the search. A wrong (non-suboptimal) starting point could conduce the wrong solution so that the performance of the optimization is deteriorated (see 7.2.3.2). Hence, even with the brute-force searching strategy, using chemical shift only, the previous system produces a "bad" match as the optimization starting point, which will deteriorate the utility of the following optimization upon proton number and multiplicity. Hence, it often biases the system to converge to an unreasonable match as the output. In contrast, spectroscopists utilize all NMR information (e.g. chemical shift, proton number, coupling, coupling constants, connectivity) to build a match between experimental multiplets and predicted multiplets. When a reasonable match is built, the consistency analysis is over. There is no stochastic optimization process during spectroscopists' structure verification procedure. However, the human's approach to build a reasonable match with all NMR information is not an easy task. This is due to the essential complexity of some NMR properties. Specifically, distinctive from chemical shift, most NMR properties are not bound with an individual multiplet, instead they are assigned to a number of multiplets together. For example, it is often impossible to accurately estimate the integration and up to the proton numbers of an experimental multiplet if it is heavily overlapping with other experimental multiplets. Alternatively, it is possible to reliably estimate the integration and to compute the proton number of the set of experimental multiplets, which are overlapped altogether. This violates the pair-wise comparison assumption assumed in the previous system (Golotvin, et al., 2006), and essentially increases the computational complexity of the searching routine. Specifically, without a pair-wise comparison assumption, all possible subsets of both the experimental multiplet list and the theoretical multiplet list are needed to be generated and utilized for building a match. To analyze its computational complexity, we assume that there are  $n$  experimental multiplets and  $m$  theoretical multiplets. Therefore, there are totally  $n!$  different experimental multiplet subsets and  $m!$  different theoretical multiplet subsets, which could be generated from the experimental multiplet list and the theoretical multiplet list. To search for a match with NMR properties e.g. proton numbers, etc. with the brute force searching strategy, pair-wise comparison between the experimental multiplet subsets and the theoretical multiplet subsets

need to be implemented. To sum them up, there are totally  $n! \cdot m!$  possible pairs. This shows the factorial complexity (exponential computational complexity), which makes brute force searching infeasible.

Furthermore, in 2.2.2, we discussed the limitation of the approach to build an experimental multiplet list and a theoretical multiplet list – both of them are missing multiplet interpretations. Specifically, we argue that both the experimental multiplet list and the theoretical multiplet list only represent a small subset of all possible multiplet interpretations, and we emphasize that the missed multiplet interpretations could become vital elements for a correct structural verification decision. To avoid the problem, our system extends the experimental multiplet list to the peak cluster hypotheses space, and extends the theoretical multiplet list to the theoretical multiplet distribution list (see 4.1). These extensions dramatically increase sizes of searching spaces built from the spectrum or the structure. Especially, by introducing the concept of theoretical multiplet distribution, the system creates the continuous multiplet space, which represents infinite theoretical multiplets that can continuously change in their chemical shifts and signal shapes. Obviously, with the number of theoretical multiplets going to infinity, the methodologies used in the previous systems, especially brute-force searching strategy adopted, are out of the scope.

With this analysis, we realize that by introducing both experimental peak cluster hypotheses space and theoretical multiplet distribution list, and relying on additional NMR knowledge e.g. proton numbers, connectivity, etc, we dramatically increase the computational complexity of the system, which makes the primitive brute-force searching strategy incompetent. To efficiently search for a match in the new problem setup, an optimization/ heuristic search needs to be designed to replace the brute-force search to build a match. Many optimization approaches have been proposed in the computer science community and been used in various application domains. To list a few of them, simulation annealing, genetic programming, genetic algorithms, Monte Carlo sampling are all well-known optimization approaches. All these approaches are based on the combination of a greedy hill climbing strategy and a random walk strategy. The purpose of introducing random walk into the approach is to avoid the problem of “trapping” in a local maximum instead of the global maximum. The same strategy is also widely adopted in many computing domains such as building artificial neural networks (Duda, et al., 2000) (Mitchell, 1997) (Hastie, et al., 2003), inducing logistic regression (Duda, et al., 2000) (Hastie, et al., 2003). For our problem, it would be convenient to directly apply one of these techniques. However, as we will discuss in 7.2.3, all these optimization approaches work in solution space, and their performance relies on a reasonable searching starting point. Unfortunately, the task of building a reasonable match as the starting point itself is difficult in our problem setup. As we have already explained, a full search with NMR information beyond chemical shift is infeasible. Also, any simplification of using NMR information will conduce a “bad” starting point in the solution space. These facts refute the proposal of directly applying these classical optimization approaches. Instead an optimization /heuristic searching approach needs to be designed to ease the complexity of building a reasonable match. In fact, with the reasonable match built by using all NMR information, a structure verification solution is found, which makes additional optimization in the solution space unnecessary.

To find an effective optimization policy to build a reasonable match, we try to mimic human spectroscopists’ logic. This is motivated by the fact that spectroscopists can quickly find a reasonable match, even though they have a big space of all possible experimental and predicted multiplet interpretations in their brains. We believe that this superiority of human over computer relies on

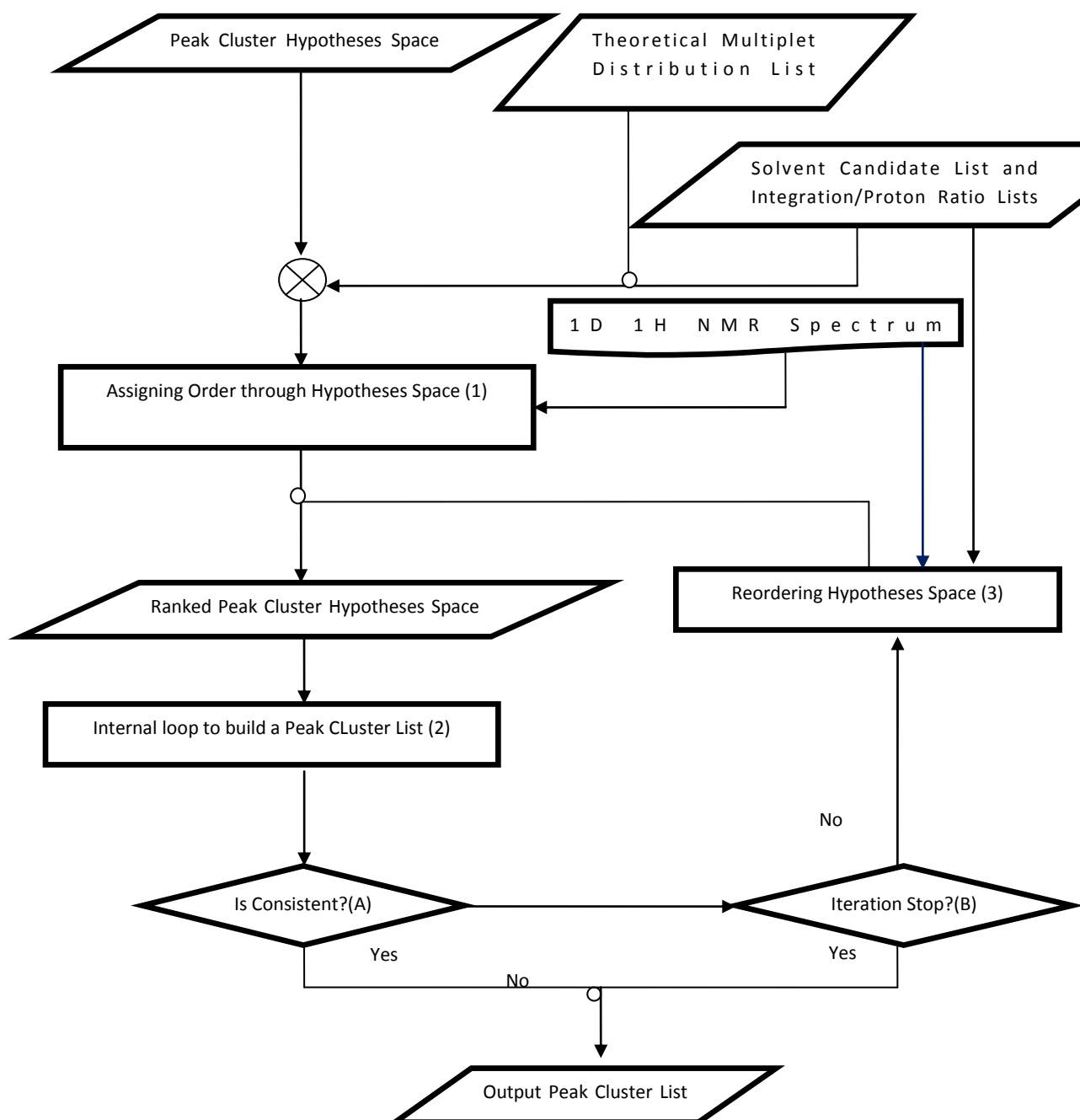
spectroscopists' flexibility to build complex search heuristics and back-tracking mechanisms. Therefore, our design of the optimization policy is focused on building a search heuristic as close as possible to that adopted by spectroscopists, and implementing back-tracking mechanisms to mimic spectroscopists' flexibility to change previously made decisions. Specifically, a mechanism is designed to mimic spectroscopists assigning a rank throughout the peak cluster hypotheses space. This is implemented by pair-wise comparisons between the experimental peak cluster hypotheses space and the theoretical multiplet distribution list along chemical shift, proton number and multiplicity. A mechanism is designed to mimic spectroscopists to sequentially build the peak cluster list by greedy searching the ranked peak cluster hypotheses space. A mechanism is designed to mimic spectroscopists' selection and deselection of the peak clusters from the partially built peak cluster list. This is implemented with the inner consistency analysis among the peak cluster list along connectivity, whenever a newly selected peak cluster hypothesis is added to the peak cluster list. As a result, the inconsistent peak cluster hypotheses are deselected from the peak cluster list, and put back in the peak cluster hypotheses space. Note, the deselected peak cluster hypotheses are not deleted from the peak cluster hypotheses space. Instead they are "punished" by reducing their priorities. In addition, a mechanism is designed to mimic spectroscopists to completely overthrow the current peak cluster list when the reasonable peak cluster hypotheses list is nonexistent, and rebuild a new peak cluster list by adopting a boosting mechanism (Freund, et al., 1996) (Efron, et al., 1994). As we represented before in Fig 29, the mechanism is used for example to select the alternative solvent signals, the alternative integration proton ratio, and this varying input NMR information will perturb the ranks defined in the peak cluster hypotheses space so as to produce the different searching track in the next searching iteration.

#### 4.4.3.1 Searching Module Architecture

The searching module is composed of the two iteration loops: a boosting loop (outer loop) and a loop to build the peak cluster list (internal loop). The data processing flow chart of the outer loop is shown in Fig 32. The outer loop is an iterative procedure (1) to rank the peak cluster hypotheses space, (2) to call the internal loop to build a peak cluster list, (3) to modify the rank of the peak cluster hypotheses space, and (4) to restart searching a new peak cluster list.

The flow chart of the internal loop is shown in Fig 33. The internal loop is an iterative procedure (1) to sequentially add the peak cluster hypothesis into the peak cluster list, (2) to analyze the consistency for the partial built peak cluster list, and to deselect the inconsistent peak cluster hypotheses from the peak cluster list, (3) to reduce the priorities of the deselected peak cluster hypotheses in the peak cluster hypotheses space.





Module (1): to rank the peak cluster hypotheses space

Module (2): to call the internal loop to build a peak cluster list

Module (3): to modify the order of the peak cluster hypotheses space

Decision Module (A): to judge the consistency

Decision Module (B): to judge the end of iteration.

**Fig 32 Searching Module Out Loop Flow Chart**

Specifically, in Fig 32, from top to bottom, first, the experimental peak cluster hypotheses space, the list of the theoretical multiplet distributions, the list of solvent signal candidates, the list of integration proton ratios and the 1D  $^1\text{H}$  NMR spectrum are fed into Module (1) to assign an initial confidence score to each peak cluster hypothesis in the peak cluster hypotheses space. Here, each peak cluster hypothesis is compared with every theoretical multiplet hypothesis and all possible combinations of them through chemical shift, proton numbers, multiplicity, coupling constants to compute a structure matching score of it. (Note, the number of theoretical multiplet distributions is reasonably small compared to the size of peak cluster hypotheses space. Therefore pair-wise comparisons upon all combination of them are computationally feasible in practice.)

In parallel, the peak cluster hypothesis is compared with the corresponding section of the 1D  $^1\text{H}$  NMR spectrum (e.g. the peak cluster hypothesis' integration, baseline level, multiplicity complexity) to give it a spectrum fitting score. This is followed by combining the structure matching score and the spectrum fitting score to give the peak cluster hypothesis a confidence score. In this way, an initial rank is defined within the peak cluster hypotheses space, while the consistent peak cluster hypothesis is associated with a high confidence score, and the inconsistent peak cluster hypothesis is associated with a low confidence score.

Next, the ranked peak cluster hypotheses space is fed into internal loop (Module (2)) to build the peak cluster list (see next paragraph for detail).

After that, a decision mechanism (Decision Module (A)) is applied to the peak cluster list to judge if it is reasonable enough to be used to explain all theoretical multiplet distributions. As a result, if all theoretical multiplet distributions are explained by the peak cluster hypotheses in the list with high confidence, the outer loop iteration terminates, and the peak cluster list is reported. Else, the peak cluster list is fed into Module (3) together with 1D  $^1\text{H}$  NMR spectrum, the solvent signal candidate list and the integration proton ratio list to re-calculate the confidence score for each peak cluster hypothesis in the peak cluster hypotheses space. This causes the change of the rank in the peak cluster hypotheses space. Then, with the peak cluster hypotheses space re-ranked, the current peak cluster list is deleted, and the procedure goes back to Module (2) to restart the iteration. This continues until the number of iteration reaches the maximum number of steps defined.

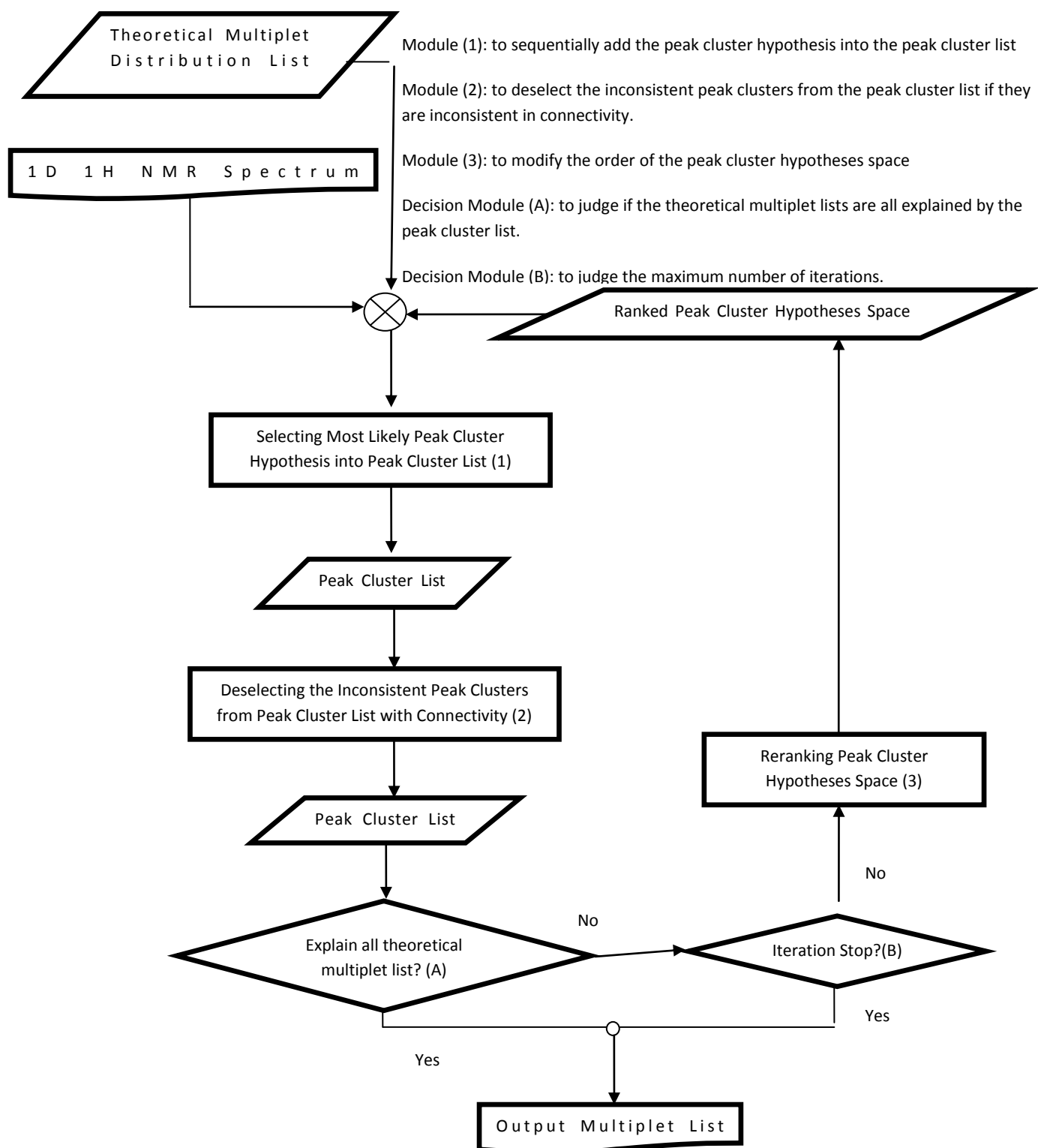


Fig 33 Searching Module Internal Loop Flow Chart

In Fig 33, from top to bottom, first, the ranked peak cluster hypotheses space and partially built peak cluster list are fed into Module (1), where the partially built peak cluster list is used to re-rank the peak cluster hypotheses space. This is followed by selecting the peak cluster hypothesis with highest confidence score from the re-ranked peak cluster hypotheses space to add into the partially built peak cluster list. Specifically, to re-rank the peak cluster hypotheses space, a pseudo spectrum is constructed from the partially built peak cluster list, and then difference between the input 1D 1H NMR spectrum and the pseudo spectrum is calculated. This “difference spectrum” is used to recalculate the spectrum fitting scores of the peak cluster hypotheses in the peak cluster hypotheses space.

Next, the consistency on coupling connectivity is analyzed (in Module (2)) among the peak cluster hypotheses in the partially built peak cluster list to adjust the peak cluster hypotheses’ confidence score in the peak cluster list. The peak cluster hypotheses with low confidence scores are deselected from the partially built peak cluster list.

After that, a decision mechanism (Decision Module (A)) is used to judge if all theoretical multiplet distributions are explained by the peak cluster list with chemical shift, proton number, multiplicity, and connectivity. If it is, the internal iteration loop terminates, and the construction of the peak cluster list is finished. Otherwise, the procedure goes to Module (3), where the confidence scores associated with the deselected peak cluster hypotheses are reduced to decrease their priorities to be used to build the peak cluster list. After that, the procedure returns back to Module (1) to select the next peak cluster hypothesis to add into the peak cluster list. The iteration continues until the maximum number of iterations is reached.

Note, in 4.4.3, the system architecture of the searching modules is described. In Chapter 5, the probabilistic model of the searching module is explained to further describe the searching heuristics and back-tracking mechanisms the system adopts. Specifically, the heuristic searching criteria are introduced in 5.2, and the computational detail of the confidence score is introduced in 5.3.

#### 4.4.4 Quantification Module

With the reasonable peak cluster list identified, quantification becomes simple. Specifically, a relative confidence score is computed for each peak cluster in the peak cluster list. The relative confidence score is computed as the absolute difference between the best structure matching score of the peak cluster hypothesis to the theoretical multiplet distribution subsets and the second best structure matching score of the peak cluster hypothesis to them (see formula 15 and 16 in 5.3 at page 100). As a result, any peak clusters with significantly large relative confidence scores are selected for quantification. In case of all peak clusters in the peak cluster list having low relative confidence scores, a “give up” signal is sent by the module to show the inability to do the quantification.

#### 4.4.5 Creating a Structure Verification Report

With the peak cluster list identified, the generation of a structure verification report is straightforward. Specifically, the confidence score of the peak cluster hypothesis in the peak cluster list is used to measure how well it explains the theoretical multiplet distribution. In case of all peak cluster hypotheses in the peak cluster list having significantly big confidence scores, a conclusion of structure verification consistency is made and reported by the module. In case of some peak cluster hypotheses in the peak cluster list having significantly low confidence scores, a conclusion of structure verification inconsistency is made and reported by the module. Note, the peak cluster hypothesis' confidence score is further decomposed to discover and report which matching factors (e.g. chemical shift, proton number, multiplicity, coupling constants, coupling connectivity, the spectrum fitting level, etc) are the cause to the inconsistency.

## **Chapter 5 A Probabilistic Explanation of the System Architecture**

In this chapter, we describe the heuristic search methods in the Consistency Analyzer (see Fig 29) with the maximum likelihood principal, and give the computational detail of how to estimate the search heuristics. The content in the chapter is an explanation of the system architecture in Chapter 4 from the probabilistic perspective. Therefore, readers who are not interested in the math detail can safely skip the chapter without loss of continuity.

### **5.1 Probabilistic Model of the Search Module**

The structure verification is a procedure to search the peak cluster hypotheses space for a reasonable peak cluster list to explain the theoretical multiplet distributions (computed from the molecular structure). For this target, a series of peak cluster hypothesis evaluations are implemented to assign and reassign a confidence score to the peak cluster hypotheses. The majority of these evaluations use empirical chemical and NMR knowledge e.g. chemical shift range, coupling constant range. To deal with these uncertainties in the evaluation procedure, a probabilistic model is appropriate.

Here, in the scope of this thesis, we denote the input 1H NMR spectrum as  $S$ , peak cluster hypotheses space as  $H$ , the theoretical multiplet distribution list as  $\mathbf{y} = (y_1, y_2, \dots, y_m)$ , where the size of  $\mathbf{y}$  is denoted as  $m$ . Correspondingly, we denote the peak cluster list as  $\mathbf{x} = (x_1, x_2, \dots, x_m)$ . With the notation given above, to simplify the problem, by assuming the peak clusters in peak cluster list are one-to-one mapped to the theoretical multiplet distributions, there are  $m^n$  possible peak cluster lists in total which can be built from  $H$  to explain  $\mathbf{y}$ . Note, that  $n$  is used to denote the size of  $H$ . The ensemble of all possible  $\mathbf{x}$  constructs a peak cluster list hypotheses space, denoted as  $\mathbf{X}$ . Obviously, each  $\mathbf{X}$  is decomposable as  $(X_1, X_2, \dots, X_m)$ , while there is a one-to-one mapping between  $y_i$  and  $X_i$ , where,  $0 < i < m$ .

By considering  $\mathbf{X} = (X_1, X_2, \dots, X_m)$  as  $m$  random variables, the conditional joint probability  $p(\mathbf{X}|\mathbf{y}, S) = p(X_1, X_2, \dots, X_m | y_1, y_2, \dots, y_m, S)$  gives a natural measurement of consistency between the peak cluster list hypotheses space  $\mathbf{X}$  and theoretical multiplet distribution list  $\mathbf{y}$  for a given spectrum  $S$ . With this interpretation, the structure verification problem is transformed into the maximum likelihood estimation framework. Formally, the optimal peak cluster list  $\mathbf{x}_{max}$  is computed as:

$$\mathbf{x}_{max} = \arg \max_{\mathbf{x} \in \mathbf{X}} p(\mathbf{X}|\mathbf{y}, S) \quad (4)$$

In mathematics,  $\mathbf{x}_{max} = \arg \max_x f(x)$  denotes the value of  $x$  for which  $f(x)$  is maximized.

The purpose of this modeling is to show the existence of the optimal peak cluster list and to prove that the optimal peak cluster list is guaranteed to be available in theory. Furthermore, the use of a

probabilistic model makes probabilistic inference theory - a powerful computational tool available for computing  $x_{max}$ .

Obviously, the search through  $X$  has exponential computational complexity. This is due to the fact that the computation of the joint probability distribution  $p(X|y, S)$  requires to compute all  $m^n$  samples' probability in  $X$ . To reduce the computational complexity, the structure of  $p(X|y, S)$  is decomposed into a product of a group of conditional probabilities (chain rule):

$$\begin{aligned} p(X|y, S) &= p(X_1, X_2, \dots, X_m | y_1, y_2, \dots, y_m, S) \\ &= p(X_1 | y_1, y_2, \dots, y_m, S) \times p(X_2 | X_1, y_1, y_2, \dots, y_m, S) \times \dots \times p(X_m | X_1, X_2, \dots, X_{m-1}, y_1, y_2, \dots, y_m, S) \end{aligned} \quad (5)$$

Given the above decomposition, the following series of inequalities are tenable, where the inequations turn into equations if  $X_1, X_2, \dots, X_m$  are conditionally independent.

$$\begin{aligned} \arg \max_{x \in X} p(X|y, S) &= \arg \max_{x \in X} p(X_1, X_2, \dots, X_m | y_1, y_2, \dots, y_m, S) \\ &\geq \arg \max_{x_1, x_2, \dots, x_{m-1} \in X_1, X_2, \dots, X_{m-1}} p(X_1, X_2, \dots, X_{m-1} | y_1, y_2, \dots, y_m, S) \\ &\quad \times \arg \max_{x_m \in X_m} p(X_m | X_1, X_2, \dots, X_{m-1}, y_1, y_2, \dots, y_m, S) \\ &\geq \arg \max_{x_1, x_2, \dots, x_{m-2} \in X_1, X_2, \dots, X_{m-2}} p(X_1, X_2, \dots, X_{m-2} | y_1, y_2, \dots, y_m, S) \\ &\quad \times \arg \max_{x_{m-1} \in X_{m-1}} p(X_{m-1} | X_1, X_2, \dots, X_{m-2}, y_1, y_2, \dots, y_m, S) \\ &\quad \times \arg \max_{x_m \in X_m} p(X_m | X_1, X_2, \dots, X_{m-1}, y_1, y_2, \dots, y_m, S) \\ &\geq \dots \\ &\geq \arg \max_{x_1 \in X_1} p(X_1 | y_1, y_2, \dots, y_m, S) \times \arg \max_{x_2 \in X_2} p(X_2 | X_1, y_1, y_2, \dots, y_m, S) \times \dots \times \\ &\quad \times \arg \max_{x_m \in X_m} p(X_m | X_1, X_2, \dots, X_{m-1}, y_1, y_2, \dots, y_m, S) \end{aligned} \quad (6)$$

As a conclusion from (6), the maximum likelihood estimation of  $X_1, X_2, \dots, X_m$  could be asymptotically approached with the product of the maximum likelihood estimation of the disjunctive subsets of  $X_1, X_2, \dots, X_m$ . In fact, this asymptotical property supplies the theoretical backbone of the heuristic searching criterions (see 4.4.4.3) used in the searching module to approximately find the optimal peak cluster list in an efficient way.

## 5.2 Searching Heuristics

The set of inequations (6) suggest an order to build the optimal peak cluster list  $x_{max}$  in  $H$  instead of directly computing  $x_{max}$  in the peak cluster list hypotheses space  $X$  (see Fig 34). This gives the searching heuristics of the searching module (see 4.4.3).

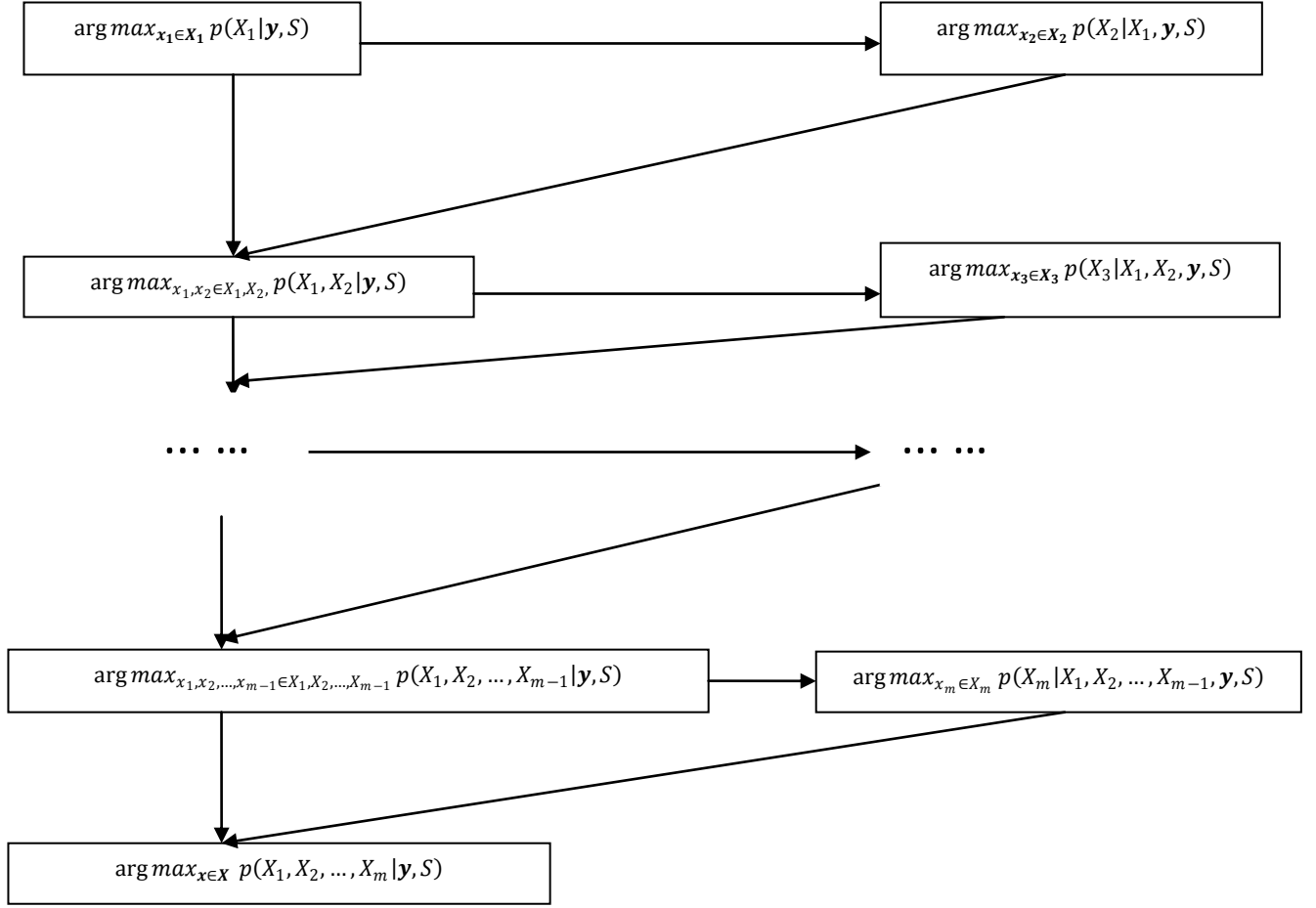


Fig 34 The Order to Build the Peak Cluster List

To asymptotically estimate  $\arg \max_{x \in X} p(X_1, X_2, \dots, X_m | y, S)$ , the maximum likelihood of the partially built peak cluster list  $\arg \max_{x_1, x_2, \dots, x_i \in X_1, X_2, \dots, X_i} p(X_1, X_2, \dots, X_i | y, S), i \in (1, \dots, m)$  is estimated sequentially. However, the estimation of  $\arg \max_{x_1, x_2, \dots, x_i \in X_1, X_2, \dots, X_i} p(X_1, X_2, \dots, X_i | y, S)$  itself has the factorial computational complexity. To further simplify the problem, the products of  $\arg \max_{x_j \in X_j} p(X_j | X_1, X_2, \dots, X_{j-1}, y, S), j \in (1, \dots, i)$  is computed to approximately estimate  $\arg \max_{x_1, x_2, \dots, x_i \in X_1, X_2, \dots, X_i} p(X_1, X_2, \dots, X_i | y, S)$  instead. This simplification makes the greedy search applicable. Specifically, the likelihood  $p(x_i | y, S)$  is estimated as the initial confidence score of the peak cluster hypothesis in the peak cluster hypotheses space (see Fig 32, Module (1)), and  $\arg \max_{x_1 \in X_1} p(X_1 | y, S)$  is used as the heuristic to select the first peak cluster hypothesis  $x_1$  to add into the peak cluster list. Similarly, the likelihood  $p(x_i | x_1, x_2, \dots, x_{i-1}, y, S)$  is estimated to modify the confidence score of the peak cluster hypothesis  $x_i$  in the peak cluster hypotheses space (see Fig 33, Module (1)), and



$\arg \max_{x_i \in X_i} p(X_i | x_1, x_2, \dots, x_{i-1}, \mathbf{y}, S)$  is used as heuristic to select the next peak cluster hypothesis  $x_i$  to be added into the peak cluster list.

However, the above greedy searching criterion is inaccurate, since for all  $i \in (1, \dots, m)$ , we have  $\prod_{j=1}^i \arg \max_{x_j \in X_j} p(X_j | x_1, x_2, \dots, x_{j-1}, \mathbf{y}, S) \leq \arg \max_{x_1, x_2, \dots, x_i \in X_1, X_2, \dots, X_i} p(X_1, X_2, \dots, X_i | \mathbf{y}, S)$ . To remove this inaccuracy, the conditional probability  $p(x_{\max j} | x_{\max 1} \dots, x_{\max j-1}, x_{\max j+1}, \dots, x_{\max i}, \mathbf{y}, S)$  is estimated for every  $x_{\max j} \in (x_{\max 1}, x_{\max 2}, \dots, x_{\max i})$ , here  $(x_{\max 1}, x_{\max 2}, \dots, x_{\max i})$  denotes the partial peak cluster list built by sequentially selecting  $x_j$  to maximize  $p(X_j | x_1, x_2, \dots, x_{j-1}, \mathbf{y}, S)$ ,  $j \in (1, \dots, i)$ . This is followed by deselecting all  $x_{\max j}$  from  $(x_{\max 1}, x_{\max 2}, \dots, x_{\max i})$ , which makes

$$p(x_{\max j} | x_{\max 1} \dots, x_{\max j-1}, x_{\max j+1}, \dots, x_{\max i}, \mathbf{y}, S) < \arg \max_{x_a \in H, x_a \text{ not } \in (x_{\max 1}, \dots, x_{\max i})} p(X_1 | \mathbf{y}, S) \quad (7)$$

tenable, and by sending them back into the peak cluster hypotheses space (see Fig 32, Module (2)).

Note that, criterion (7) guarantees that only the peak cluster hypotheses with maximum conditional probability  $p(x_{\max j} | x_{\max 1} \dots, x_{\max j-1}, x_{\max j+1}, \dots, x_{\max i}, \mathbf{y}, S)$  are kept in the partial peak cluster list  $(x_{\max 1}, \dots, x_{\max i})$ . Since  $p(x_{\max j} | x_{\max 1} \dots, x_{\max j-1}, x_{\max j+1}, \dots, x_{\max i}, \mathbf{y}, S) \propto p(x_{\max 1}, \dots, x_{\max i} | \mathbf{y}, S)$  for all  $x_{\max j} \in (x_{\max 1}, x_{\max 2}, \dots, x_{\max i})$ , this approximately maximizes the joint conditional probability  $p(x_{\max 1}, x_{\max 2}, \dots, x_{\max i} | \mathbf{y}, S)$ . As a result, the following equation is approximately tenable.

$$\prod_{j=1}^i \arg \max_{x_j \in X_j} p(X_j | x_1, x_2, \dots, x_{j-1}, \mathbf{y}, S) \approx \arg \max_{x_1, x_2, \dots, x_i \in X_1, X_2, \dots, X_i} p(X_1, X_2, \dots, X_i | \mathbf{y}, S)$$

### 5.3 Estimating Probability with Chemical and NMR Knowledge

The searching routine requires the estimation of the conditional probability  $p(x_i | \mathbf{y}, S)$ ,  $x_i \in H$  and  $p(x_j | x_1 \dots, x_{j-1}, x_{j+1}, \dots, x_i, \mathbf{y}, S)$ ,  $x_j \in (x_1, x_2, \dots, x_i)$  (see 4.4.4.2).  $p(x_i | \mathbf{y}, S)$  is interpreted as the likelihood of the peak cluster hypothesis  $x_i$  to be in the optimal peak cluster list for the given input NMR spectrum and the theoretical multiplet distribution list. This likelihood is determined by using NMR and chemical knowledge such as  $x_i$ 's consistency with each subset of  $\mathbf{y}$  in chemical shifts, proton number, multiplicity, coupling constants,  $x_i$ 's fitness to the input spectrum,  $x_i$ 's reliability, etc. As already shown in section 4.4.3,  $p(x_i | \mathbf{y}, S)$  is estimated by the product of  $x_i$ 's structure matching score and spectrum matching score. Formally, we have

$$\bar{p}(x_i | \mathbf{y}, S) = \theta_{x_i, \text{structure}} \times \theta_{x_i, \text{spectrum}} \quad (8)$$

Furthermore,  $\theta_{x_i, \text{structure}}$  is estimated as the maximum matching score between  $x_i$  and every subset of  $\mathbf{y}$ , which is denoted as  $y_j, j \in (1, \dots, m!)$ , in chemical shift, proton number, multiplicity, and coupling constants. Formally, we have

$$\theta_{x_i, \text{structure}} = \max(\theta_{x_i}^{y_1}, \dots, \theta_{x_i}^{y_j}, \dots, \theta_{x_i}^{y_{m!}})$$

$$\text{with } \theta_{x_i}^{y_j} = f(\theta_{x_i,cs}^{y_j}, \theta_{x_i,pn}^{y_j}, \theta_{x_i,M}^{y_j}, \theta_{x_i,J}^{y_j}) = \theta_{x_i,cs}^{y_j} \times \theta_{x_i,pn}^{y_j} \times \theta_{x_i,M}^{y_j} \times \theta_{x_i,J}^{y_j} \quad (9)$$

Here,  $\theta_{x_i,cs}^{y_j}$  is a measure for the matching between  $x_i$  and  $y_j$  in chemical shift.  $\theta_{x_i,pn}^{y_j}$  is a measure for the matching between  $x_i$  and  $y_j$  in proton number.  $\theta_{x_i,M}^{y_j}$  is a measure for the matching between  $x_i$  and  $y_j$  in multiplicity.  $\theta_{x_i,J}^{y_j}$  is a measure for the matching between  $x_i$  and  $y_j$ 's coupling constants.

To simplify the computation, we assume that each measure independently influences the structure matching measure  $\theta_{x_i}^{y_j}$  in (9).

$\theta_{x_i,spectrum}$  is estimated as the product of  $x_i$ 's fitness to the input spectrum, and  $x_i$ 's reliability. Formally, we have

$$\theta_{x_i,spectrum} = f(\theta_{x_i,sf}, \theta_{x_i,reli}) = \theta_{x_i,sf} \times \theta_{x_i,reli} \quad (10)$$

Here,  $\theta_{x_i,sf}$  is a measure to scale  $x_i$ 's fitness to the spectrum  $S$ , and  $\theta_{x_i,reli}$  is a measure to scale  $x_i$ 's chance to be a simple, clean, non-overlapped experimental multiplet. Note, to simplify the computation, we assume each measure independently influences the spectrum matching measure  $\theta_{x_i,spectrum}$  in (10).

Similarly,  $p(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_i, y, S)$ ,  $x_j \in (x_1, x_2, \dots, x_i)$  is decomposed as the product of a structure matching score and a spectrum matching score. The structure matching score and the spectrum matching score are different from  $\theta_{x_i,structure}$  and  $\theta_{x_i,spectrum}$ , since the computation of the scores of  $x_j$  requires the consideration of the matching measure for other peak clusters  $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_i)$  in the peak cluster list. To distinguish the difference, we denote the structure matching score as  $\theta'_{x_j,structure}$ , and denote the spectrum matching score as  $\theta'_{x_j,spectrum}$ . In addition, the factors are introduced to punish the previous deselect of  $x_j$  from the peak cluster list (see 4.4.3 and Fig 33 Module (2)). Due to the independency of the structure measurement and the spectra measurement, different factors are used to punish the structure matching score and the spectra matching score respectively. Formally, we have

$$\bar{p}(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_i, y, S) = (\theta'_{x_j,structure} \times (1 - bias_{str})^{cnt_{str\_rej}}) \times (\theta'_{x_j,spectrum} \times (1 - bias_{spe})^{cnt_{spe\_rej}}) \quad (11)$$

Here,  $bias_{str}$  and  $bias_{spe}$  are hyper parameters to define the level of the punishment in structure matching and spectrum matching respectively.  $cnt_{str\_rej}$  denotes the times  $x_j$  was deselected from the peak cluster list due to structural inconsistencies.  $cnt_{spe\_rej}$  denotes the times of deselecting  $x_j$  from the peak cluster list due to spectrum inconsistencies.

Furthermore,  $\theta'_{x_j,structure}$  is estimated as the maximum matching score between  $x_j$  and any subset  $y_k$   $k \in (1, \dots, m!)$  of  $y$  in chemical shift, proton number, multiplicity, coupling constants, and connectivity. Formally, we have

$$\theta'_{x_j,structure} = \max(\theta_{x_j}^{y_{1'}}, \dots, \theta_{x_j}^{y_{k'}}, \dots, \theta_{x_j}^{y_{m!'}}) \quad (12)$$

$$\text{with } \theta_{x_j}^{y_{k'}} = f(\theta_{x_j,cs}^{y_k}, \theta_{x_j,M}^{y_k}, \theta_{x_j,J}^{y_k}, \theta_{x_j,con}^{y_k}, \theta_{x_j,pn}^{y_k}) = \theta_{x_j,cs}^{y_k} \times \theta_{x_j,M}^{y_k} \times \theta_{x_j,J}^{y_k} \times \theta_{x_j,con}^{y_k} \times \theta_{x_j,pn}^{y_k} \quad (13)$$

Here,  $\theta_{x_j, con}^{y_k, x_i, y}$  is a measure of the connectivity consistency among  $x_1, \dots, x_i, y_1, \dots, y_m$  by assigning  $x_j$  to  $y_k$ .

$\theta_{x_j, spectrum}'$  is estimated as the product of  $x_i$ 's fitness to the “difference” spectrum (see below), and  $x_i$ 's reliability. Formally, we have

$$\theta_{x_j, spectrum}' = f(\theta_{x_i, sf}', \theta_{x_i, reli}) = \theta_{x_i, sf}' \times \theta_{x_i, reli} \quad (14)$$

Here,  $\theta_{x_i, sf}$  is a measurement to scale  $x_i$ 's fitness to the “difference” spectrum. Whereas the “difference” spectrum is defined as the absolute difference between the input NMR spectrum and the pseudo spectrum constructed with peak cluster list  $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_i)$ .

With slight modifications, formula (11) and (12) are directly usable for the estimation of the peak cluster hypothesis's relative confidence score for quantification (see 4.4.5). By denoting the peak cluster hypothesis  $x_i$ 's relative confidence score as  $RCS_{x_i}$ , we obtain

$$RCS_{x_i} = (RCS_{x_i, structure} \times (1 - bias_{str})^{cnt_{str, rej}}) \times (\theta_{x_j, spectrum}' \times (1 - bias_{spe})^{cnt_{spe, rej}}) \quad (15)$$

$$RCS_{x_i, structure} = \max(\theta_{x_j}^{y_{1'}}, \dots, \theta_{x_j}^{y_{k'}}, \dots, \theta_{x_j}^{y_{m'}}) - secondmax(\theta_{x_j}^{y_{1'}}, \dots, \theta_{x_j}^{y_{k'}}, \dots, \theta_{x_j}^{y_{m'}}) \quad (16)$$

### 5.3.1 Computing $\theta_{x_i, cs}^{y_j}$

$\theta_{x_i, cs}^{y_j}$  denotes the measure of the consistency between the peak cluster hypothesis  $x_i$  and the theoretical multiplet distribution subset  $y_j$  in chemical shift. Formally we have

$$\theta_{x_i, cs}^{y_j} = f(cs_{x_i}, cs_{y_j}, csl_{y_j}, csh_{y_j}) \quad (17)$$

Here,  $cs_{x_i}$  denotes chemical shift of peak cluster hypothesis  $x_i$ 's,  $cs_{y_j}$  denotes the average chemical shift of the theoretical multiplet distribution subset  $y_j$ ,  $csl_{y_j}$  and  $csh_{y_j}$  denote the lowest and the highest end of  $y_j$ 's chemical shift, respectively.

Spectroscopists utilize several empirical rules to evaluate the consistency between  $x_i$  and  $y_j$  in chemical shift, which are described below.

- (1) If  $x_i$  and  $y_j$  are consistent in chemical shift, the probability of the experimental chemical shift  $cs_{x_i}$  fall into the chemical shift range  $[csl_{y_j}, csh_{y_j}]$  is high (> 95%).
- (2)  $csl_{y_j}$  and  $csh_{y_j}$  can be asymmetric to  $cs_{y_j}$
- (3)  $cs_{y_j}$  has the highest probability density in range  $[csl_{y_j}, csh_{y_j}]$ .

To model these empirical rules, the beta function is used to concretely compute  $\theta_{x_i, cs}^{y_j}$ . This results in:

$$\theta_{x_i,cs}^{y_j} = \begin{cases} \text{beta}(cs_{x_i}, csl_{y_j}, csh_{y_j}, \alpha, \beta) & , csl_{y_j} < cs_{x_i} < csh_{y_j} \\ \varepsilon & , \text{others} \end{cases} \quad (18)$$

with  $\text{beta}(cs_{x_i}, csl_{y_j}, csh_{y_j}, \alpha, \beta)$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left( \frac{cs_{x_i}}{csh_{y_j} - csl_{y_j}} - \frac{csl_{y_j}}{csh_{y_j} - csl_{y_j}} \right)^{\alpha-1} \left( 1 - \left( \frac{cs_{x_i}}{csh_{y_j} - csl_{y_j}} - \frac{csl_{y_j}}{csh_{y_j} - csl_{y_j}} \right) \right)^{\beta-1} \quad \text{where,}$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Here,  $\alpha$  and  $\beta$  are hyper-parameters, which are computed with  $cs_{y_j}, csl_{y_j}, csh_{y_j}$ .  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

Note, under the premise to satisfy above empirical rules, the choice of the function to calculate  $\theta_{x_i,cs}^{y_j}$  is arbitrary. Other functions e.g. triangle or ladder function can be used here instead.

### 5.3.2 Computing $\theta_{x_i,pn}^{y_j}$

With the integration proton ratio given, the proton number of the peak cluster hypothesis can be computed directly. On the other hand, the proton number of the given theoretical multiplet distribution subset is simply the sum of the proton number of all theoretical multiplet distributions in the subset. With both proton numbers calculated, a strict decision rule is used to measure the consistency between  $x_i$  and  $y_j$ 's proton numbers. Formally, we have

$$\theta_{x_i,pn}^{y_j} = f(pn_{x_i}, pn_{y_j})$$

$$= \begin{cases} 1 - \varepsilon, & pn_{x_i} = pn_{y_j} \\ \varepsilon, & pn_{x_i} \neq pn_{y_j} \end{cases} \quad (19)$$

Here,  $pn_{x_i}$  denotes the proton numbers of the peak cluster hypothesis  $x_i$ ,  $pn_{y_j}$  denotes the proton numbers of the theoretical multiplet distribution subset  $y_j$ .  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

### 5.3.3 Computing $\theta_{x_i,M}^{y_j}$

Different from  $\theta_{x_i,cs}^{y_j}$  and  $\theta_{x_i,pn}^{y_j}$ ,  $\theta_{x_i,M}^{y_j}$  is computed only if it is possible to match multiplicity between  $x_i$  and  $y_j$ . While it is impossible to match  $x_i$  and  $y_j$ 's multiplicity,  $\theta_{x_i,M}^{y_j}$  is set as the default value 1. Specifically, two conditions are required to be satisfied for computing  $\theta_{x_i,M}^{y_j}$ . (1) The peak cluster hypothesis  $x_i$  should appear as a clear first order multiplet pattern. In other words, the multiplet

hypothesis space of the peak cluster hypothesis should contain a first-order multiplet interpretation, which can explain the majority signal of the peak cluster hypotheses. (2) The theoretical multiplet distribution subset should contain only one theoretical multiplet distribution. The first condition limits the candidates to the peak cluster hypotheses, which contain clear first-order multiplicity pattern. At the same time, it excludes the peak cluster hypotheses which are the signals from second (high)-order multiplets. The second condition limits the candidates to peak cluster hypotheses, which is one-to-one mapped to the theoretical multiplet distribution. Obviously, in case that a peak cluster hypothesis is matched to multiple theoretical multiplet distributions (which mean that the observed peak cluster hypothesis is the overlapping of multiple multiplets), reliable multiplicities cannot be extracted from the peak cluster hypothesis. Therefore, the match on the multiplicity should not be computed.

With above two conditions satisfied, a strict decision rule is used to measure the consistency between  $x_i$  and  $y_j$ 's multiplicities. Formally, we have

$$\begin{aligned} \theta_{x_i, M}^{y_j} &= f(nc_{x_i}, nc_{y_j}) \\ &= \begin{cases} 1 - \varepsilon, & nc_{x_i} = nc_{y_j} \\ \varepsilon, & nc_{x_i} \neq nc_{y_j} \end{cases} \quad (20) \end{aligned}$$

Here,  $nc_{x_i}$  denotes the number of couplings of the multiplet hypothesis in the peak cluster hypothesis  $x_i$ ,  $nc_{y_j}$  denotes the number of coupling of the theoretical multiplet distribution in  $y_j$ .  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

### 5.3.4 Computing Coupling Constant Measure $\theta_{x_i, J}^{y_j}$

Since the coupling constant is the quantity bounded with multiplicity, similar to  $\theta_{x_i, M}^{y_j}$ ,  $\theta_{x_i, J}^{y_j}$  is only computed when above two conditions in 5.3.3 are satisfied. Specifically,  $\theta_{x_i, J}^{y_j}$  denotes the measure of the consistency between the multiplet hypothesis in  $x_i$  and the theoretical multiplet distribution  $y_j$  in coupling constants. First, both the coupling constants of the multiplet hypothesis in  $x_i$  and the predicted coupling constants of the theoretical multiplet distribution  $y_j$  are sorted by the numerical size of the coupling constants. Next,  $n$  one-to-one consistent mappings are built between the sorted coupling constants of  $x_i$  and  $y_j$ .  $n$  is the number of couplings of  $x_i$  or  $y_j$ . Then  $\theta_{x_i, J}^{y_j}$  is represented as the product of  $n$  consistent measurements defined upon the  $n$  one-to-one consistent mapping between  $x_i$  and  $y_j$ . Formally, we have

$$\theta_{x_i, J}^{y_j} = \theta_{x_i, J_1}^{y_j} \times \theta_{x_i, J_2}^{y_j} \times \dots \times \theta_{x_i, J_k}^{y_j} \times \dots \times \theta_{x_i, J_n}^{y_j} \quad (21)$$

Here,  $\theta_{x_i, J_k}^{y_j}$  denotes the consistent measurement between the  $k$ th coupling constant of  $x_i$  and the  $k$ th coupling constant of  $y_j$ . Then,  $\theta_{x_i, J_k}^{y_j}$  can be written as:

$$\theta_{x_i J_k}^{y_j} = f(J_{x_i}^k, J_{y_j}^k, J_{y_j}^k, J_{y_j}^k) \quad (22)$$

Here,  $J_{x_i}^k$  denotes of the multiplet hypothesis in  $x_i$ 's  $k$ th coupling constant,  $J_{y_j}^k$  denotes the theoretical multiplet distribution  $y_j$ 's  $k$ th predicted coupling constant,  $J_{y_j}^k$  denotes  $y_j$ 's  $k$ th coupling constant range low end,  $J_{y_j}^k$  denotes  $y_j$ 's  $k$ th coupling constant range high end.

Spectroscopists utilize several empirical rules to evaluate  $\theta_{x_i J_k}^{y_j}$ , which are described below.

- (1) If  $x_i$  and  $y_j$ 's  $k$ th coupling constant is consistent, the probability of the experimental coupling constant  $J_{x_i}^k$  falls into the coupling constant range  $[J_{y_j}^k, J_{y_j}^k]$  is high (> 95%).
- (2)  $J_{y_j}^k$  and  $J_{y_j}^k$  are asymmetric to  $J_{y_j}^k$ .
- (3)  $J_{y_j}^k$  has the highest probability density in range  $[J_{y_j}^k, J_{y_j}^k]$ .

To model these empirical rules, the beta function is used to concretely compute  $\theta_{x_i J_k}^{y_j}$ . Formally, we have

$$\theta_{x_i J_k}^{y_j} = \begin{cases} \text{beta}(J_{x_i}^k, J_{y_j}^k, J_{y_j}^k, \alpha, \beta) & , \quad J_{y_j}^k < J_{x_i}^k < J_{y_j}^k \\ \epsilon & , \quad \text{others} \end{cases} \quad (23)$$

with,  $\text{beta}(J_{x_i}^k, J_{y_j}^k, J_{y_j}^k, \alpha, \beta)$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left( \frac{J_{x_i}^k}{J_{y_j}^k - J_{y_j}^k} - \frac{J_{y_j}^k}{J_{y_j}^k - J_{y_j}^k} \right)^{\alpha-1} \left( 1 - \left( \frac{J_{x_i}^k}{J_{y_j}^k - J_{y_j}^k} - \frac{J_{y_j}^k}{J_{y_j}^k - J_{y_j}^k} \right) \right)^{\beta-1}$$

where,  $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ .

Here,  $\alpha$  and  $\beta$  are hyper parameters, which could be optimized by  $J_{y_j}^k, J_{y_j}^k, J_{y_j}^k$ .  $\epsilon$  is a small positive real number used to model the noise in the measurement channel.

Note, under the premise to satisfy above empirical rules, the choice of the function to calculate  $\theta_{x_i J_k}^{y_j}$  is arbitrary. Other functions e.g. triangle or ladder function can be used here instead.

### 5.3.5 Computing Coupling Connectivity Measure $\theta_{x_i, con}^{y_j, x, y}$

To estimate  $\theta_{x_i, con}^{y_j}$ , first, all possible one-to-one assignments between  $(x_1, x_2, \dots, x_{i-1})$  and  $(y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_m)$  need to be established. Here,  $m \geq i$ . By adding the additional mapping between  $x_i$  and  $y_j$ , all possible one-to-one mappings between  $(x_1, x_2, \dots, x_i)$  and  $(y_1, y_2, \dots, y_m)$  are constructed. Formally, we denote the ensemble of all possible assignments as  $\mathbf{X}$ , and  $\mathbf{X}$  is decomposable as  $(X_1, X_2, \dots, X_m)$ . Obviously, the size of  $\mathbf{X}$  is  $C_{i-1}^{m-1}$ . Next, a theoretical connectivity

matrix is constructed from the theoretical multiplet distribution list  $y_1, y_2, \dots, y_m$ . We denote it as  $M^y$ . Specifically,  $M_{p,q}^y$ , the element of the  $p$ th row and the  $q$ th column of  $M^y$ , represents the existence of a coupling from theoretical multiplet distribution  $y_p$  to  $y_q$ . (Note, one represents the existence, and zero represents the nonexistence.) Similarly, an experimental connectivity matrix is constructed for each  $x \in X$ . We denote it as  $M^x$ , while  $M_{p,q}^x$ , the element of the  $p$ th row and the  $q$ th column of  $M^x$ , represents the number of experimental couplings from experimental multiplet hypotheses  $x_p$  to  $x_q$ . Specifically,  $M_{p,q}^x$  is estimated by counting the number of the couplings in  $x_p$ , which have the coupling constant equal to a coupling constant of  $x_q$ . With above matrixes, the connectivity consistency analysis is easily implemented by comparing the corresponding numbers of  $M^x$  and  $M^y$ . If  $M_{p,q}^x \geq M_{p,q}^y$  for all  $p$  and  $q$  where both  $x_p$  and  $x_q$  are existed and assigned to  $y_p$  and  $y_q$ , the assignment of  $x_i$  to  $y_j$  is consistent with  $x_1, x_2, \dots, x_{i-1}, y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_i$  in connectivity. Formally, we have

$$\theta_{x_i, con}^{y_j} = f(x_i, M^x, M^y) = \begin{cases} 1 - \varepsilon, & \exists x \in X, M^x \geq M^y \\ \varepsilon, & \text{Others} \end{cases} \quad (24)$$

Note,  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

The existence symbol in formula (24) shows that the exhausted search through  $X$  is unnecessary. The search stops whenever  $\exists x \in X, M^x \geq M^y$  is satisfied. To further reduce the computational complexity,  $\prod_{k=1}^i \theta_{x_k, CS}^{y_k} \times \theta_{x_k, M}^{y_k} \times \theta_{x_k, J}^{y_k}$  is computed for  $x \in X$  to define an order through  $X$ . In this way, the assignment  $x$ , which has large individual matching score on chemical shift, proton number, multiplicity and coupling constants, is estimated first.

### 5.3.6 Spectrum Fitting Score $\theta_{x_i, sf}$ and $\theta'_{x_i, sf}$

$\theta_{x_i, sf}$  is used to measure in how far the peak cluster hypothesis  $x_i$  matches the input 1D  $^1H$  NMR spectrum. This is measured by  $x_i$ 's integration on the real spectrum. Formally, we have

$$\theta_{x_i, sf} = I_{x_i} \quad (25)$$

Note,  $I_{x_i}$  denote the integration of  $x_i$  on the real spectrum  $S$  normalized over the peak cluster hypotheses space.

Similarly,  $\theta'_{x_i, sf}$  is used to measure in how far the peak cluster hypothesis  $x_i$  matches the “difference” spectrum. The procedure to compute the “difference” spectrum is described below. Specifically,

- Set the “difference” spectra  $D$  equal to the real spectrum  $S$ .
- Loop through  $(x_1, x_2, \dots, x_{i-1})$  with  $j$
- Subtract the signal of peak cluster hypothesis  $x_j$  from  $D$  to compute the new “difference” spectrum.
- Repeat b.

A hybrid rule is used to measure in how far the peak cluster hypothesis  $x_i$  matches the “difference” spectrum. Specifically, if all regions of the “difference” spectrum that the peak cluster hypothesis  $x_i$  covers contain NMR signals,  $\theta'_{x_i, sf}$  is measured by  $x_i$ ’s integration into the “difference” spectrum. Else, a small positive real number  $\varepsilon$  is assigned to  $\theta'_{x_i, sf}$ . Formally, we have

$$\theta'_{x_i, sf} = \begin{cases} I'_{x_i}, & \text{if } D \text{ contains full signals of } x_i \\ \varepsilon, & \text{Others} \end{cases} \quad (26)$$

Note,  $I'_{x_i}$  denotes the integration of  $x_i$  on the “difference” spectrum  $D$  normalized over the peak cluster hypotheses space,  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

### 5.3.7 Reliability Score $\theta_{x_i, reli}$

$\theta_{x_i, reli}$  is used to measure spectroscopists’ subjective appraisal of a peak cluster hypothesis  $x_i$ . Intuitively, spectroscopists prefer prominent, simple, “clean” experimental peak clusters. Here, “clean” means non-overlapping, low-baseline, etc. This can be described with  $x_i$ ’s fitness to the spectrum,  $x_i$ ’s average line width,  $x_i$ ’s amplitude asymmetric level,  $x_i$ ’s base line level,  $x_i$ ’s out overlapping level,  $x_i$ ’s multiplicity complexity,  $x_i$ ’s peak reliability, etc. Correspondingly, a group of factors are defined to estimate  $\theta_{x_i, reli}$ . Formally, we have

$$\theta_{x_i, reli} = \theta_{x_i, sf} \times \theta_{x_i, lw} \times \theta_{x_i, asym} \times \theta_{x_i, M-sim} \times \theta_{x_i, ol\_out} \times \theta_{x_i, bs} \times \theta_{x_i, peak\_reli} \times \dots \quad (27)$$

Here,

(1)  $\theta_{x_i, asym}$  is a factor to measure  $x_i$ ’s level of amplitude asymmetry. Specifically,  $\theta_{x_i, asym}$  is proportional to the sum of the absolute amplitude difference between every symmetric peak pair in  $x_i$ .

(2)  $\theta_{x_i, M-sim}$  is a factor to measure  $x_i$ ’s multiplicity complexity. Specifically, a rough empirical rule is used here to assign  $\theta_{x_i, M-sim}$  a score. Formally we have

$$\theta_{x_i, M-sim} = \begin{cases} 1, & 0 \leq x_i \text{’s number of significant multiplet interpretation} \leq 2 \\ 0.5, & 2 < x_i \text{’s significant multiplet interpretation} \leq 3 \\ 0.25, & 3 < x_i \text{’s significant multiplet interpretation} \leq 4 \\ \varepsilon, & \text{others.} \end{cases} \quad (28)$$

Here,  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

(3)  $\theta_{x_i, ol\_out}$  is a factor to measure the overlapping level between  $x_i$  and the neighboring peak clusters. Specifically, the following procedure is used to estimate  $\theta_{x_i, ol\_out}$ .

- a. The nearest local minimum points or the inflexion points on both sides of  $x_i$  are detected.
- b. The ratio of the amplitude of the most left peak of  $x_i$  and the nearest local minimum point (or the inflexion point) on the left side, and the ratio of the amplitude of the most right peak of



$x_i$  and the nearest local minimum point (or the inflexion point) on the right side are computed and averaged.

c. Normalize  $x_i$ 's average amplitude ratio through the peak cluster hypotheses space. The normalized average amplitude ratio then is used as the estimation of  $\theta_{x_i,ol\_out}$ .

(4)  $\theta_{x_i,bs}$  is a factor to measure  $x_i$ 's relative baseline level. Specifically,  $x_i$ 's baseline level is computed by averaging the amplitudes of the nearest local minimum or inflexion points.  $\theta_{x_i,bs}$  is estimated as the ratio between  $x_i$ 's baseline level and the spectra's noise level. Note, the normalization is implemented through the peak cluster hypotheses space to restrict the value of  $\theta_{x_i,bs}$  in  $[0, 1]$ .

(5)  $\theta_{x_i,sf}, \theta_{x_i,lw}, \theta_{x_i,peak\_reli}$  (see 5.3.8)

Note, the independent assumption over  $\theta_{x_i,sf}, \theta_{x_i,lw}, \theta_{x_i,asym}, \theta_{x_i,M-sim}, \theta_{x_i,ol\_out}, \theta_{x_i,bs}$  is inaccurate. For example,  $\theta_{x_i,ol\_out}$  and  $\theta_{x_i,bs}$  rely on each other. A Bayesian network could be introduced here to model the interdependence among  $\theta_{x_i,sf}, \theta_{x_i,lw}, \theta_{x_i,asym}, \theta_{x_i,M-sim}, \theta_{x_i,ol\_out}, \theta_{x_i,bs}$ .

### 5.3.8 Solvent likelihood $\theta_{x_i,so}$

Going back to solvent detection (see 4.4.1), a confidence score is computed for each peak cluster hypothesis in the peak cluster hypothesis space to model its likelihood to be a solvent signal. In this sector, we formalize the solvent likelihood score computation procedure.

Specifically,  $\theta_{x_i,so}$  is used to measure the likelihood of the peak cluster hypothesis  $x_i$  to be a solvent signal. According to spectroscopists' experience, depending on the type of the solvent, different factors influence spectroscopists' recognition decision. As an example, the estimation of the likelihood of  $x_i$  to be a H2O signal or a DMSO signal is described in this sector. Formally, we denote the H2O likelihood as  $\theta_{x_i,H2O}$ , and the DMSO likelihood as  $\theta_{x_i,DMSO}$ .

A set of empirical rules are used to estimate  $\theta_{x_i,H2O}$ . Specifically, (1) the H2O signal appears in the chemical shift range of 3.0ppm – 4.9ppm. (2) The number of split peaks in the H2O signal is likely between 1 and 3. (3) The H2O signal has a wide line width. (4) The peaks of the H2O signal are overlapping. (5) The H2O signal does not have satellite peaks, etc. Correspondingly, a group of factors are defined to estimate  $\theta_{x_i,H2O}$ . Formally, we have

$$\theta_{x_i,H2O} = \theta_{x_i,H2O-cs} \times \theta_{x_i,H2O-M} \times \theta_{x_i,H2O-pn} \times \theta_{x_i,lw} \times \theta_{x_i,ol\_int} \times \theta_{x_i,sate} \times \theta_{x_i,peak\_reli} \times \dots \quad (29)$$

Here,

(1)  $\theta_{x_i,H2O-cs}$  is a factor to measure  $x_i$ 's likelihood to be in the H2O signal's chemical shift range. This is computed by fitting a ladder distribution (see 4.4.1.2).

(2)  $\theta_{H2O-M}$  is used to evaluate  $x_i$ 's likelihood to be a H2O signal with its peak number. Specifically, a rough empirical rule is used here to assign  $\theta_{H2O-mult}$  a score. Formally we have

$$\theta_{H2O-M} = \begin{cases} 1, & \text{if number of peaks in } x_i \text{ equals to one.} \\ 0.5, & \text{if number of peaks in } x_i \text{ equals to two.} \\ 0.25, & \text{if number of peaks in } x_i \text{ equals to three.} \\ \varepsilon, & \text{others.} \end{cases} \quad (30)$$

Here,  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

(3)  $\theta_{H2O-pn}$  is used to evaluate  $x_i$ 's likelihood to be a H2O signal with its proton number (See 4.4.1.2).

(4)  $\theta_{x_i, lw}$  is a factor to measure  $x_i$ 's line width. Specifically,  $\theta_{x_i, lw}$  is proportional to the derivation of  $x_i$ 's average line width from the average line width of the input spectra  $S$ .

(5)  $\theta_{x_i, ol\_int}$  is a factor to measure the overlapping level among the peaks of  $x_i$ . Specifically, the following procedure is used to estimate  $\theta_{x_i, ol\_int}$ .

- The local minimum points or the inflexion points between the peaks of  $x_i$  are detected.
- The ratio of the amplitude of the detected local minimum point (or the inflexion point) and the average amplitude of the neighboring peaks are computed.
- The product of the amplitude ratios is used as the estimation of  $\theta_{x_i, ol\_int}$ .

(6)  $\theta_{x_i, sate}$  is a factor to punish  $x_i$  for having the satellite peaks. Specifically, we have

$$\theta_{x_i, sate} = \begin{cases} (1 - \varepsilon), & x_i \text{ does not have satellite peaks} \\ \varepsilon, & x_i \text{ has satellite peaks} \end{cases} \quad (31)$$

Here,  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

(7)  $\theta_{x_i, peak\_reli}$  is a factor to measure the reliability of  $x_i$ 's peaks. It is estimated as a product of the confidence scores of  $x_i$ 's peaks.

Note, the independent assumption of  $\theta_{x_i, H2O-cs}$ ,  $\theta_{x_i, H2O-mult}$ ,  $\theta_{x_i, H2O-pn}$ ,  $\theta_{x_i, lw}$ ,  $\theta_{x_i, ol\_int}$ ,  $\theta_{x_i, sate}$ ,  $\theta_{x_i, peak\_rl}$  is inaccurate. For example,  $\theta_{x_i, lw}$  have direct influence on  $\theta_{x_i, ol\_int}$ . A Bayesian network could be introduced here to model the interdependence among  $\theta_{x_i, H2O-cs}$ ,  $\theta_{x_i, H2O-mult}$ ,  $\theta_{x_i, H2O-pn}$ ,  $\theta_{x_i, lw}$ ,  $\theta_{x_i, ol\_int}$ ,  $\theta_{x_i, sate}$ .

Similarly, a set of empirical rules are used to estimate  $\theta_{x_i, DMSO}$ . Specifically, (1) the *DMSO* signal appears in the chemical shift range of 2.0ppm – 3.0ppm. (2) The *DMSO* signal has certain multiplicity, e.g. most likely to be a quintuplet or a doublet of triplet. (3) The *DMSO* signal has certain coupling constant from 1.4Hz to 2.1Hz. (4) The *DMSO* signal has satellite peaks, etc. Correspondingly, a group of factors are defined to estimate  $\theta_{x_i, DMSO}$ . Formally, we have

$$\theta_{x_i, DMSO} = \theta_{x_i, DMSO-cs} \times \theta_{x_i, DMSO-M} \times \theta_{x_i, DMSO-pn} \times (1 - \theta_{x_i, sate}) \times \theta_{x_i, peak\_reli} \times \dots \quad (32)$$

Here,

(1)  $\theta_{x_i, DMSO}$  is a factor to measure  $x_i$ 's likelihood to be in the *DMSO* signal's chemical shift range. This is computed by fitting a ladder distribution (see 4.4.1.1).

(2)  $\theta_{x_i, DMSO-M}$  is used to evaluate  $x_i$ 's likelihood to be a *DMSO* signal with its multiplicity. Specifically, a rough empirical rule is used here to assign  $\theta_{x_i, DMSO-M}$  a score. Formally we have

$$\theta_{x_i, DMSO-M} = \begin{cases} 1, & \text{if } x_i\text{'s multiplicity is a quintuplet or a doublet of triplet.} \\ 0.5, & \text{if } x_i\text{'s multiplicity is a triplet.} \\ 0.25, & \text{if } x_i\text{'s multiplicity is a doublet or a singleton.} \\ \varepsilon, & \text{others.} \end{cases} \quad (33)$$

Here,  $\varepsilon$  is a small positive real number used to model the noise in the measurement channel.

(3)  $\theta_{x_i, DMSO-pn}$  is used to evaluate  $x_i$ 's likelihood to be a *DMSO* signal with its proton number (See 4.4.1.1).

(4)  $\theta_{x_i, sate}$ ,  $\theta_{x_i, peak\_rl}$  (see above).

## **Chapter 6 Experiments**

In this chapter we introduce the experimental setup we utilized to evaluate the automatic structural verification system. This is followed by presenting the experimental results and discussions about them. Emphatically, the experiments are specifically designed to evaluate the performance of the system in term of decision accuracy and consistency with human experts.

### **6.1 Experimental Setup**

To evaluate the performance of the automatic structural verification system, as the premise, we firstly have to answer the question about what is the consistency between the structure and the 1D <sup>1</sup>H NMR spectrum. In our opinion, the consistency between the spectrum and the structure means that the structure is uniquely explainable with the given spectrum. With this premise, the consistency decision making relies on answering the following two questions.

- (1) Does the spectrum explain the proposed structure?
- (2) Does the spectrum only explain the proposed structure?

If the answers to both questions are affirmative, from a practical point of view we say that the spectrum and the structure are consistent to each other.

With the above understanding of consistency between the spectrum and the structure, the accuracy of the structural consistency verification system could be warranted by controlling two types of errors. They are :

- a. the spectrum-structure pair is consistent, but the system judges that they are inconsistent (the first type of error).
- b. the spectrum-structure pair is inconsistent, but the system judges that they are consistent (the second type of error).

Obviously, good accuracy of the system means minimizing both types of errors. The estimation of these two types of errors gives us the first measurement of the system's performance.

In order to push the system into practice to replace human spectroscopists, it is important to convince spectroscopists by showing them the detail assignments between the structure and the spectrum. Obviously, high consistency between the assignments of the system to that of the spectroscopists will convince them of the reliability of the system, and thereby influence the business decision in the management level of the pharmaceutical industry. Hence, the consistency between the assignments of the system and spectroscopists gives us the second measurement of the system's performance.

### 6.1.1 Evaluation Criteria

To control both types of errors and the consistency between the system's assignments and that of spectroscopists, three criteria are defined. They are – the False Negative Rate (FN), the False Positive Rate (FP) and the Consistency Rate (CR).

Specifically, in our problem setup, FN is defined as the percentage of cases where the system's decision is inconsistent in all consistent spectrum-structure test pairs. Formally, we have

$$FN = \lim_{n_{CH} \rightarrow \infty} \frac{n_{CH}^{IS}}{n_{CH}}. \quad (34)$$

Here  $n_{CH}$  denotes the number of experimental test cases where spectrum-structure pairs are consistent,  $n_{CH}^{IS}$  denotes the number of cases where spectrum-structure pairs are consistent, but the system decides that they are inconsistent.

FP is defined as the percentage of cases where the system's decision is consistent in all inconsistent spectrum-structure test pairs. Formally, we have

$$FP = \lim_{n_{IH} \rightarrow \infty} \frac{n_{IH}^{CS}}{n_{IH}}. \quad (35)$$

Here  $n_{IH}$  denotes the number of cases where spectrum-structure pairs are inconsistent,  $n_{IH}^{CS}$  denotes the number of cases where spectrum-structure pairs are inconsistent, but the system decides that they are consistent.

CR is defined as the percentage of the system's assignments which is consistent with the spectroscopists' assignments in all system's assignments. Formally, we have

$$CR = \lim_{n_{TS} \rightarrow \infty} \frac{n_{TS}^{HS}}{n_{TS}}. \quad (36)$$

Here  $n_{TS}$  denotes the total number of the system's assignments, where  $n_{TS}^{HS}$  denotes the number of the system's assignments which is consistent with the spectroscopists' assignments.

In practice, it is impossible to accurately calculate the value of these criteria, since the calculation of them requires infinite test cases. Instead, we estimate the values of the criteria by utilizing big (but finite) test datasets. Correspondingly, the estimation formulas of the criteria are defined as following.

$$FN' = \frac{n_{CH}^{'IS}}{n_{CH}} \quad (37)$$

where  $FN'$  is the estimation of FN. Here  $n_{CH}'$  denotes the total number of cases where spectrum-structure pairs are consistent in the test dataset,  $n_{CH}^{'IS}$  denotes the number of cases where spectrum-structure pairs are consistent, but the system decides that they are inconsistent in the test dataset.

$$FP' = \frac{n'_{IH}^{CS}}{n'_{IH}} \quad (38)$$

where  $FP'$  is the estimation of FP. Here  $n'_{IH}$  denotes the total number of cases where spectrum-structure pairs are inconsistent in the test dataset,  $n'_{IH}^{CS}$  denotes the number of cases where spectrum-structure pairs are inconsistent, but the system decides that they are consistent in the test dataset.

$$CR' = \frac{n'_{TS}^{HS}}{n'_{TS}} \quad (39)$$

where  $CR'$  is the estimation of CR. Here  $n'_{TS}$  denotes the total number of system's assignments in the test dataset, where  $n'_{TS}^{HS}$  denotes the number of system's assignments which is consistent with the spectroscopists' assignments in the test dataset.

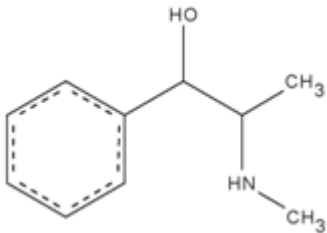
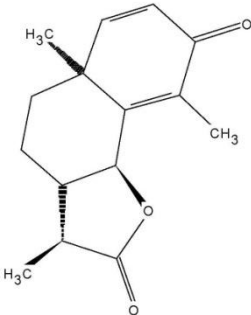
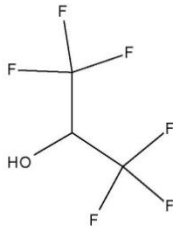
## 6.1.2 Evaluation Data

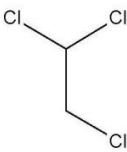
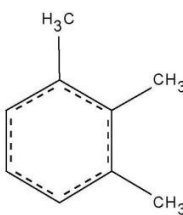
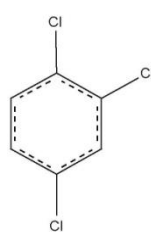
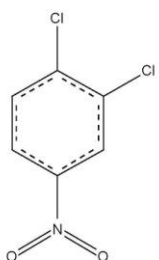
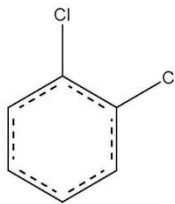
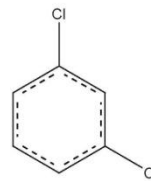
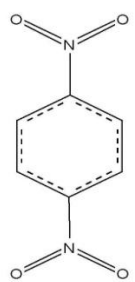
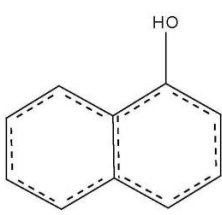
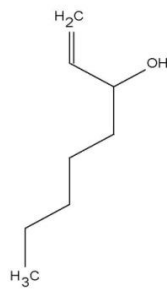
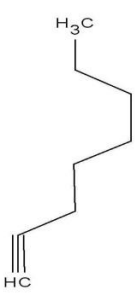
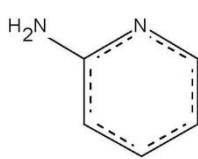
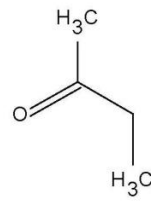
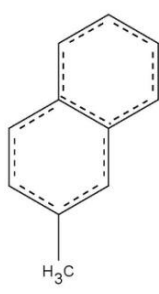
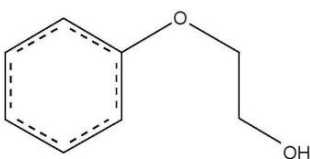
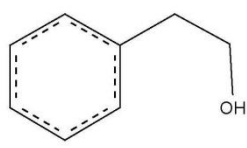
In this subsection, we introduce the datasets we used to evaluate the performance of the system.

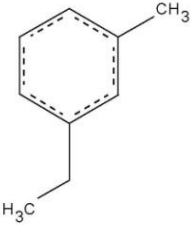
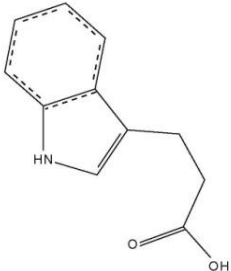
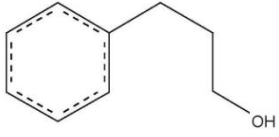
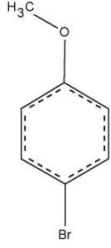
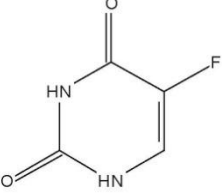
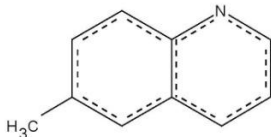
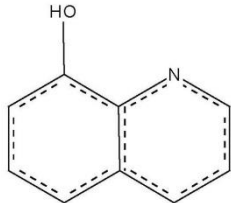
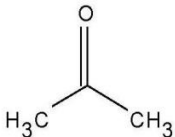
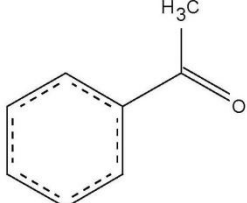
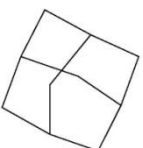
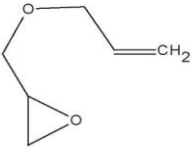
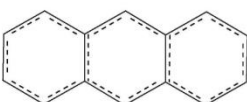
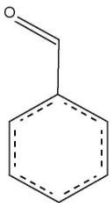
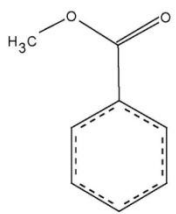
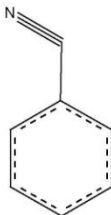
### 6.1.2.1 Real Compounds and Their Spectra

To accurately estimate the criteria we defined in 6.1.1, we need a reasonably large test dataset, which is infeasible to be acquired in practice. Due to the cost involved in doing so, to make the most with our limited budget, 85 real compounds with known 2D structure (which contain some amount of unknown impurities) were bought and their 1H 1D NMR spectra were acquired by our industrial cooperator. All compounds were diluted in DMSO, and were measured with 400MHz NMR spectrometer to acquire their spectra. The list of the compounds used in the evaluation is shown in Table

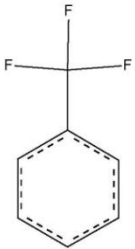
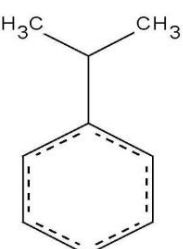
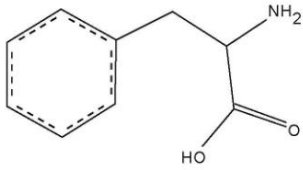
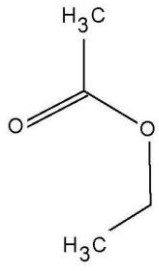
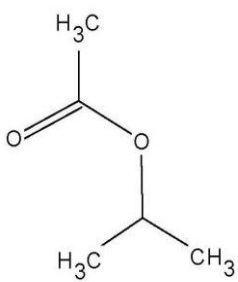
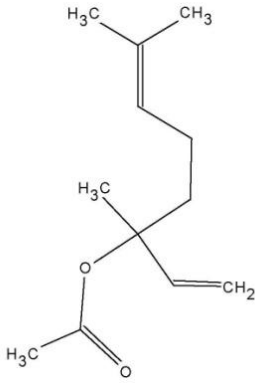
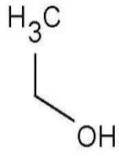
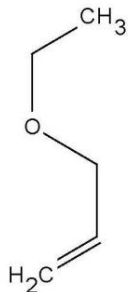
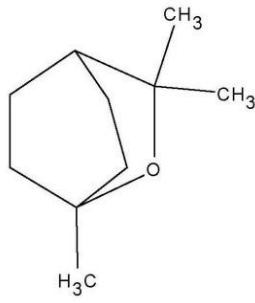
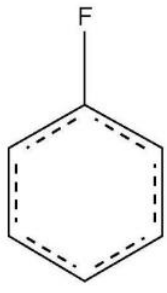
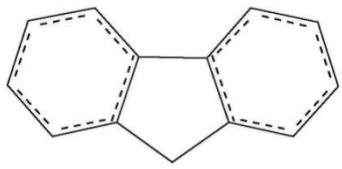
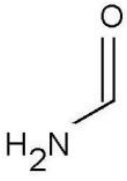
4.

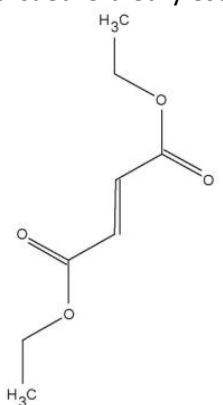
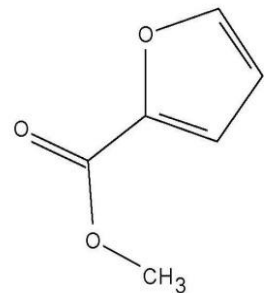
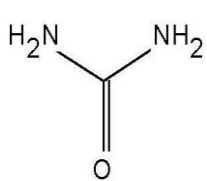
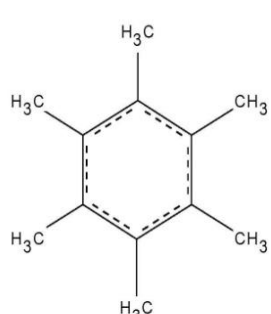
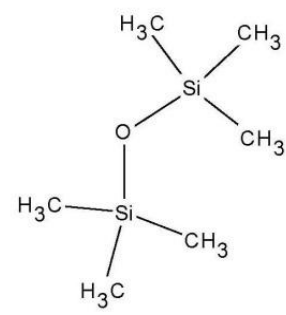
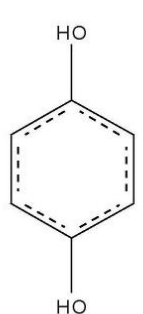
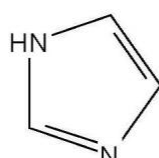
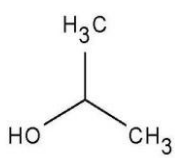
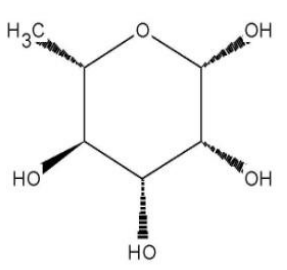
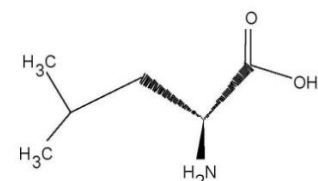
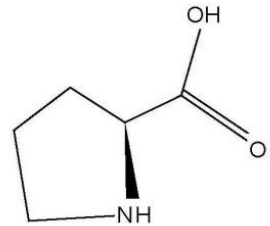
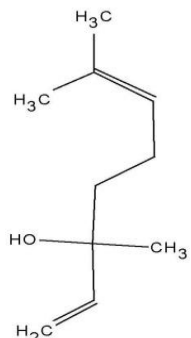
<p>+Pseudoephedrin</p> 	<p>--alpha-Satonin</p> 	<p>1,1,1-3,3,3-Hexafluor-2-propanol</p> 
--	--	---

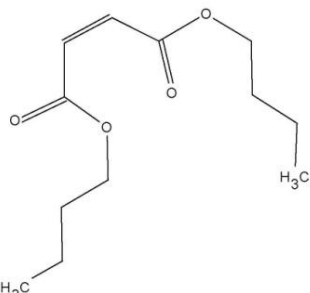
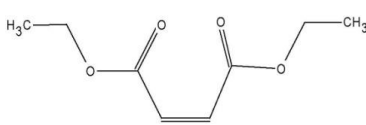
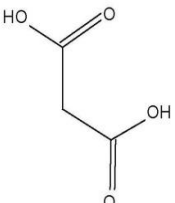
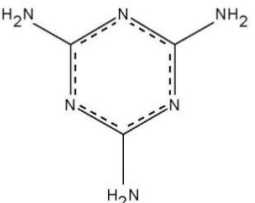
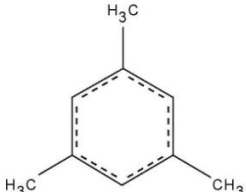
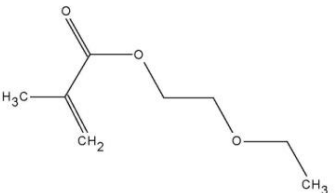
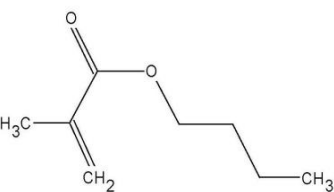
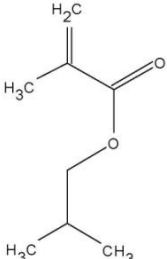
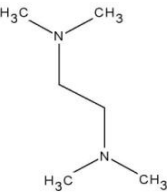
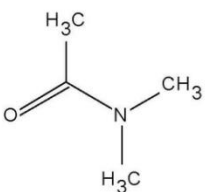
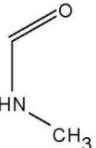
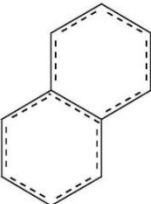
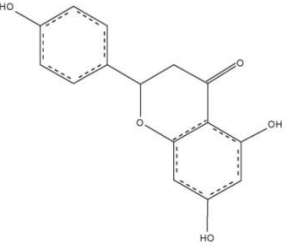
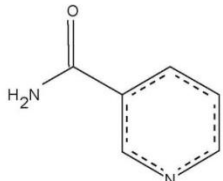
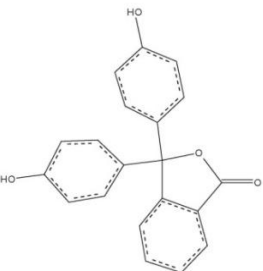
<b>1,1,2-Trichlorethan</b> 	<b>1,2,3-Trimethylbenzol</b> 	<b>1,2,4-Trichlorbenzol</b> 
<b>1,2-Dichlor-4-nitrobenzol</b> 	<b>1,2-Dichlorbenzol</b> 	<b>1,3-Dichlorbenzol</b> 
<b>1,4-Dinitrobenzol</b> 	<b>1-Naphthol</b> 	<b>1-Octen-3-ol</b> 
<b>1-Octyne</b> 	<b>2-Aminopyridin</b> 	<b>2-Butanon</b> 
<b>2-Methyl-naphthalin</b> 	<b>2-Phenoxyethanol</b> 	<b>2-phenyl-ethylakohol</b> 

<p>3-Ethyltoluol</p> 	<p>3-Indolepropionicacid</p> 	<p>3-Phenyl-propylalkohol</p> 
<p>4-Bromanisol</p> 	<p>5-Fluorouracil</p> 	<p>6-Methyl-chinolin</p> 
<p>8-Hydroxy-chinolin</p> 	<p>Aceton</p> 	<p>Acetophenon</p> 
<p>Adamantan</p> 	<p>Allylglycidether</p> 	<p>Anthracen</p> 
<p>Benzaldehyd</p> 	<p>Benzoesauremethylester</p> 	<p>Benzonitril</p> 



<p>Benzotrifluorid</p> 	<p>Cumol</p> 	<p>D,L-Phenylalanin</p> 
<p>Essigester</p> 	<p>Essigsäure-isopropyl-ester</p> 	<p>Essigsäurelinalylester</p> 
<p>Ethanol</p> 	<p>Ethylallyl-ether</p> 	<p>Eucalyptol</p> 
<p>Fluorbenzol</p> 	<p>Fluoren</p> 	<p>Formamid</p> 

<p>Fumarsaeure-diethylester</p> 	<p>Furan-2-carbonsaeuremethylester</p> 	<p>Harnstoff</p> 
<p>Hexamethylbenzol</p> 	<p>Hexamethyldisiloxan</p> 	<p>Hydrochinon</p> 
<p>Imidazol</p> 	<p>Isopropanol</p> 	<p>L-+-Rhamnose-Monohydrat</p> 
<p>L-Leucin</p> 	<p>L-Prolin</p> 	<p>Linalool</p> 

<b>Maleinsaeure-dibutylester</b> 	<b>Maleinsaeure-diethylester</b> 	<b>Malonsaeure</b> 
<b>Melamin</b> 	<b>Mesiylen</b> 	<b>Methacrylsaeure-2-ethoxyethylester</b> 
<b>Methacrylsaeure-butylester</b> 	<b>Methacrylsaeure-isobutylester</b> 	<b>N,N,N,N-Tetramethyl-ethylendiamin</b> 
<b>N,N-Dimethylacetamid</b> 	<b>N-Methylformamid</b> 	<b>Naphthalin</b> 
<b>Naringenin</b> 	<b>Nicotinsaeureamid</b> 	<b>Phenolphthalein</b> 

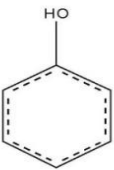
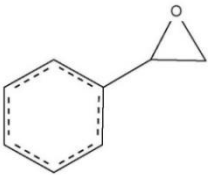
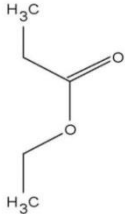
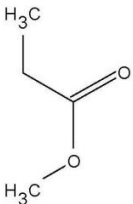
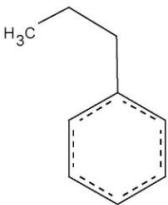
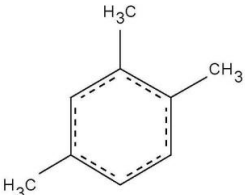
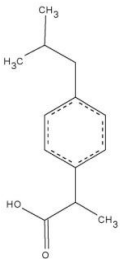
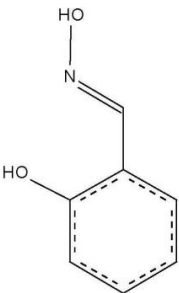
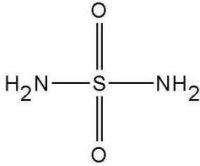
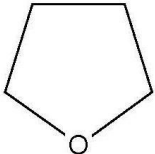
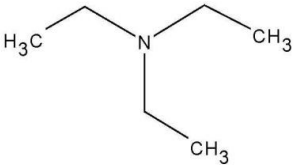
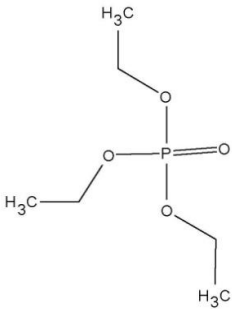
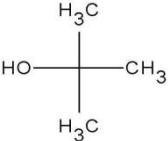
<p>Phenol</p> 	<p>Phenylethylenoxid</p> 	<p>Propionsaeureethylester</p> 
<p>Propionsaeuremethylester</p> 	<p>Propylbenzol</p> 	<p>Pseudocumol</p> 
<p>S+-2-4- Isobutylphenylpropionsaeure</p> 	<p>Salicylaldoxim</p> 	<p>Sulfamid</p> 
<p>THF</p> 	<p>Triethylamin</p> 	<p>Triethylphosphat</p> 
<p>tert-Butylalkohol</p> 		

Table 4 List of Compounds Used in Evaluation

### 6.1.2.2 Simulated Spectra and Theoretical Multiplet Distribution Lists

Since we only had a limited number of real compounds and thereby real spectra, which have limited representativeness of spectral variation in term of spectrum baseline, multiplet overlapping, high-order multiplet, signals of impurities, etc, the estimation  $FN'$  and  $FP'$  on 85 real compounds and their spectra may not be convincing.

To increase the reliability of estimations on both  $FN'$  and  $FP'$ , an artificial dataset of simulated spectra and their corresponding consistent theoretical multiplet distribution lists were automatically generated by our industrial cooperator with a simulation program. We were not informed about the approach used to implement the simulation program, to prevent us from “cheating”. What we did know is that the program randomly changes the level of spectrum baseline, the level of multiplet overlap, the number of high-order multiplet, and the number of impurity signal, etc. In addition, some simulated spectra are randomly selected and shown to several top NMR spectroscopists to be confirmed regarding their quality and usability.

Specifically, two setups are used to generate the simulated spectra and their corresponding theoretical multiplet distribution lists. In the first setup (easy setup), the maximum number of the theoretical multiplet distributions (chemical equivalent protons) are controlled to be 16. This is the setup which domain experts (spectroscopists) believe could be used to simulate compounds with regular complexity. Fig 35 shows a randomly selected example of the simulated spectrum under the first setup, and Fig 36 shows its corresponding theoretical multiplet distribution list.

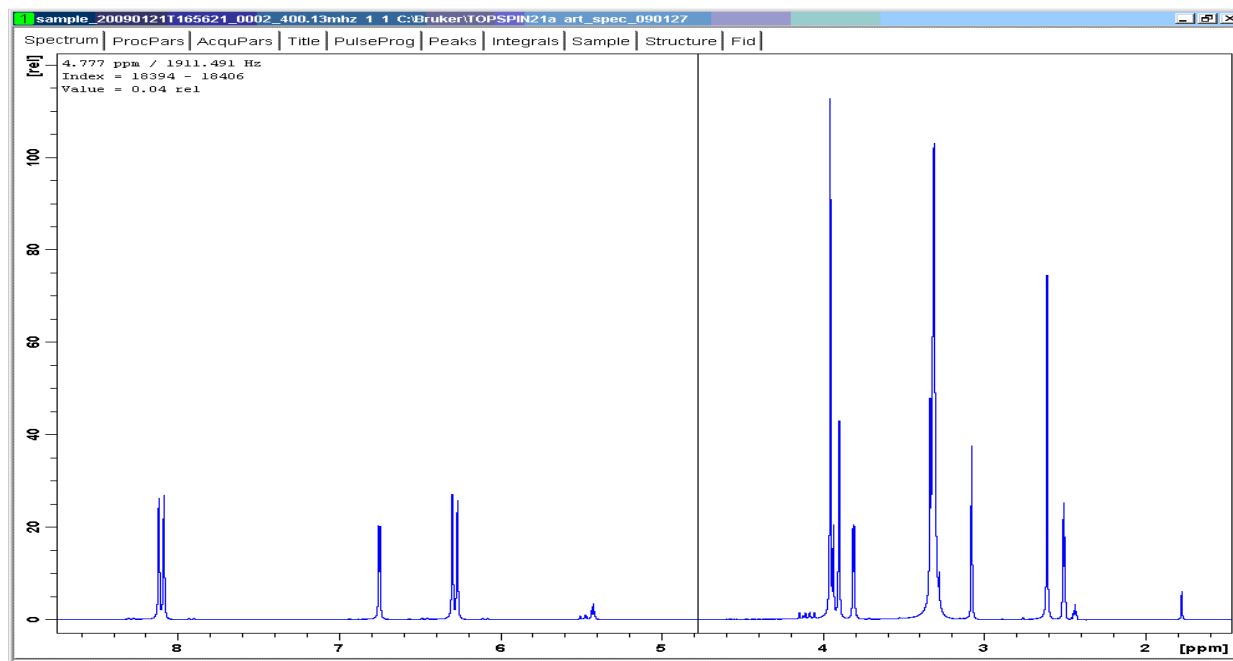


Fig 35 An example of a simulated spectrum (the first setup)

```
# Generated: 21-Jan-2009
# Equivalent protons
#
$1      number  shift  shift_range  J      coupling_range  M      complex  connection
      identifier      proton_number  J_het  J_het_range  J_het_connection  M_het  %_het
      linewidth
```

#	[-] [-]	[ppm] [-]	[ppm] [%]	[Hz] [Hz]	[Hz]	[-]	[-]	[-]	[-]	[-]	[Hz]	[Hz]
1	1 (100,150)	4.044 J(1,C13)	(3.24,4.84) 2	N/A 1.07	N/A 0.5-4	1		1	N/A	H1	3	120
2	2 120	6.170 (100,150)	(5.37,6.97) J(2,C13)	0.66 2	(2.34,3.66) 1.07		2	2	J(2,3)	H2		1
3	2 120	6.170 (100,150)	(5.37,6.97) J(2,C13)	0.55 2	(2.45,3.55) 1.07		2	2	J(2,4)	H2		1
4	2 120	6.170 (100,150)	(5.37,6.97) J(2,C13)	12.85 2	(9.85,15.85) 1.07		2	2	J(2,5)	H2		1
5	3 120	6.170 (100,150)	(5.37,6.97) J(3,C13)	0.66 2	(2.34,3.66) 1.07		2	2	J(3,2)	H3		1
6	3 120	6.170 (100,150)	(5.37,6.97) J(3,C13)	12.85 2	(9.85,15.85) 1.07		2	2	J(3,4)	H3		1
7	3 120	6.170 (100,150)	(5.37,6.97) J(3,C13)	0.55 2	(2.45,3.55) 1.07		2	2	J(3,5)	H3		1
8	4 120	8.139 (100,150)	(7.34,8.94) J(4,C13)	0.55 2	(2.45,3.55) 1.07		2	2	J(4,2)	H4		1
9	4 120	8.139 (100,150)	(7.34,8.94) J(4,C13)	12.85 2	(9.85,15.85) 1.07		2	2	J(4,3)	H4		1
10	4 120	8.139 (100,150)	(7.34,8.94) J(4,C13)	0.66 2	(2.34,3.66) 1.07		2	2	J(4,5)	H4		1
11	5 120	8.139 (100,150)	(7.34,8.94) J(5,C13)	12.85 2	(9.85,15.85) 1.07		2	2	J(5,2)	H5		1
12	5 120	8.139 (100,150)	(7.34,8.94) J(5,C13)	0.55 2	(2.45,3.55) 1.07		2	2	J(5,3)	H5		1
13	5 120	8.139 (100,150)	(7.34,8.94) J(5,C13)	0.66 2	(2.34,3.66) 1.07		2	2	J(5,4)	H5		1
14	6 (100,150)	2.441 J(6,C13)	(1.64,3.24) 2	N/A 1.07	N/A 0.5-4	1		1	N/A	H6	2	120
15	7 120	3.839 (100,150)	(3.04,4.64) J(7,C13)	2.14 2	(0.86,5.14) 1.07		2	1	J(7,8)	H7		1
16	8 120	3.045 (100,150)	(2.25,3.85) J(8,C13)	2.14 2	(0.86,5.14) 1.07		2	1	J(8,7)	H8		1
17	8 120	3.045 (100,150)	(2.25,3.85) J(8,C13)	2.14 2	(0.86,5.14) 1.07		2	1	J(8,9)	H8		1
18	9 120	6.569 (100,150)	(5.77,7.37) J(9,C13)	2.14 2	(0.86,5.14) 1.07		2	1	J(9,8)	H9		1
19	10 (100,150)	4.116 J(10,C13)	(3.32,4.92) 2	N/A 1.07	N/A 0.5-4	1		1	N/A	H10	1	120
20	11 (100,150)	3.337 J(11,C13)	(2.54,4.14) 2	N/A 1.07	N/A 0.5-4	1		1	N/A	H11	1	120
21	12 120	3.649 (100,150)	(2.85,4.45) J(12,C13)	12.83 2	(9.83,15.83) 1.07		2	1	J(12,13)	H12		1
22	13 N/A	1.440 N/A	(-2.64,2.24) N/A	12.83 2	(9.83,15.83) N/A		2	1	J(13,12)	H13		1
23	14 (100,150)	3.053 J(14,C13)	(2.25,3.85) 2	N/A 1.07	N/A 0.5-4	1		1	N/A	H14	1	120
#	IDENTICAL CHEMICAL SHIFTS AND J COUPLINGS											
#	If the chemical shifts are identical, shift ranges, proton numbers, J(het), M(het), %(het), linewidths and											
#	reliabilities also need to be identical.											
#	If Js are identical, coupling ranges and Ms also need to be identical											
#												
\$2	ep_no_1		ep_no_2									
CS	2	=	3									
CS	4	=	5									
\$3	J_1		J_2									
J	J(2,3)	=	J(3,2)									
J	J(2,4)	=	J(3,5)									
J	J(2,5)	=	J(3,4)									
J	J(4,2)	=	J(5,3)									
J	J(4,3)	=	J(5,2)									
J	J(4,5)	=	J(5,4)									
#												
\$4	CHIRAL CENTERS:											
#												
CC	N/A											
#												
\$5	THROUGH SPACE COUPLINGS:											
#												
TSC	N/A											
#												
\$6	TAUTOMERISM:											
#												
TA	N/A											

Fig 36 An example of a simulated theoretical multiplet distribution list (the first setup)

In the second setup (difficult setup), the maximum number of the theoretical multiplet distributions are controlled to be 25. This is the setup which domain experts (spectroscopists) believe could be used to simulate compounds with higher complexity. Fig 37 shows a randomly selected example of the simulated spectrum under the second setup, and Fig 38 shows its corresponding theoretical multiplet distribution list.

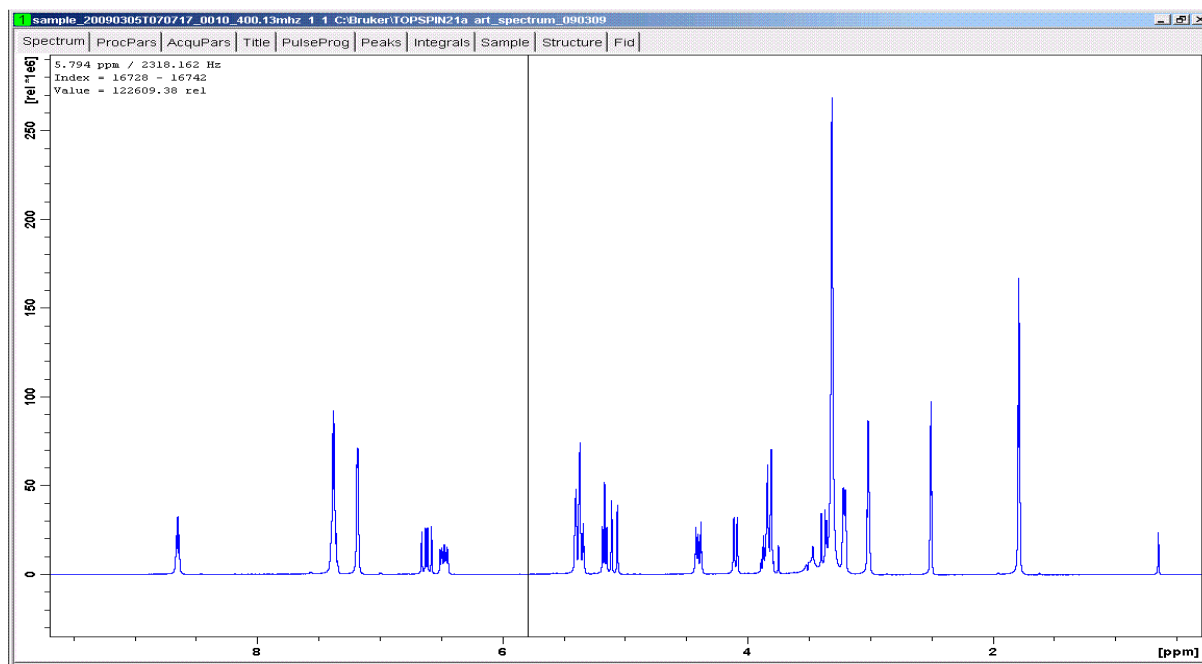


Fig 37 An example of a simulated spectrum (the second setup)

# Generated: 05-Mar-2009  
# Equivalent protons

#	number	shift	shift_range	J	coupling_range	M	complex	connection		
#	identifier	[ppm]	proton_number	J_het	J_het_range	J_het_connection		M_het	%_het	[Hz]
#	[#]	[ppm]	[ppm]	[Hz]	[#]	[#]	[#]	[#]	[#]	[Hz]
#	[#]	[#]	[#]	[#]	[#]	[#]	[#]	[#]	[#]	[#]
1	1	1.430	(0.63,2.23)	2.82	(0.00,5.82)	3	1	J(1,2)	H1	3
120		(100,150)	J(1,C13) 2	1.07	0.5-4					
2	2	2.704	(1.90,3.50)	2.82	(0.00,5.82)	4	1	J(2,1)	H2	2
120		(100,150)	J(2,C13) 2	1.07	0.5-4					
3	3	3.307	(2.51,4.11)	2.82	(0.00,5.82)	2	1	J(3,4)	H3	2
120		(100,150)	J(3,C13) 2	1.07	0.5-4					
4	3	3.307	(2.51,4.11)	5.85	(2.85,8.85)	2	1	J(3,5)	H3	2
120		(100,150)	J(3,C13) 2	1.07	0.5-4					
5	3	3.307	(2.51,4.11)	0.73	(0.00,3.73)	2	1	J(3,6)	H3	2
120		(100,150)	J(3,C13) 2	1.07	0.5-4					
6	4	5.587	(4.79,6.39)	2.82	(0.00,5.82)	3	1	J(4,3)	H4	1
120		(100,150)	J(4,C13) 2	1.07	0.5-4					
7	4	5.587	(4.79,6.39)	10.91	(7.91,13.91)	2	1	J(4,6)	H4	1
120		(100,150)	J(4,C13) 2	1.07	0.5-4					
8	5	5.109	(4.31,5.91)	5.85	(2.85,8.85)	3	1	J(5,3)	H5	1
120		(100,150)	J(5,C13) 2	1.07	0.5-4					
9	6	4.341	(3.54,5.14)	0.73	(0.00,3.73)	3	1	J(6,3)	H6	1
120		(100,150)	J(6,C13) 2	1.07	0.5-4					
10	6	4.341	(3.54,5.14)	10.91	(7.91,13.91)	2	1	J(6,4)	H6	1
120		(100,150)	J(6,C13) 2	1.07	0.5-4					
11	7	8.316	(7.52,9.12)	4.72	(1.72,7.72)	2	1	J(7,8)	H7	1
120		(100,150)	J(7,C13) 2	1.07	0.5-4					
12	7	8.316	(7.52,9.12)	6.63	(3.63,9.63)	2	1	J(7,10)	H7	1
120		(100,150)	J(7,C13) 2	1.07	0.5-4					
13	8	6.705	(5.90,7.50)	4.72	(1.72,7.72)	2	1	J(8,7)	H8	1
120		(100,150)	J(8,C13) 2	1.07	0.5-4					
14	8	6.705	(5.90,7.50)	16.15	(13.15,19.15)	2	1	J(8,9)	H8	1
120		(100,150)	J(8,C13) 2	1.07	0.5-4					
15	8	6.705	(5.90,7.50)	6.63	(3.63,9.63)	2	1	J(8,10)	H8	1
120		(100,150)	J(8,C13) 2	1.07	0.5-4					
16	9	3.537	(2.74,4.34)	16.15	(13.15,19.15)	2	1	J(9,8)	H9	1
120		(100,150)	J(9,C13) 2	1.07	0.5-4					
17	9	3.537	(2.74,4.34)	5.85	(2.85,8.85)	2	1	J(9,10)	H9	1
120		(100,150)	J(9,C13) 2	1.07	0.5-4					
18	9	3.537	(2.74,4.34)	15.03	(12.03,18.03)	2	1	J(9,11)	H9	1
120		(100,150)	J(9,C13) 2	1.07	0.5-4					
19	10	3.919	(3.12,4.72)	6.63	(3.63,9.63)	2	1	J(10,7)	H10	1
120		(100,150)	J(10,C13) 2	1.07	0.5-4					
20	10	3.919	(3.12,4.72)	6.63	(3.63,9.63)	2	1	J(10,8)	H10	1
120		(100,150)	J(10,C13) 2	1.07	0.5-4					
21	10	3.919	(3.12,4.72)	5.85	(2.85,8.85)	2	1	J(10,9)	H10	1
120		(100,150)	J(10,C13) 2	1.07	0.5-4					
22	11	4.138	(3.34,4.94)	15.03	(12.03,18.03)	2	1	J(11,9)	H11	1
120		(100,150)	J(11,C13) 2	1.07	0.5-4					
23	12	4.080	(3.28,4.88)	2.91	(0.00,5.91)	2	1	J(12,13)	H12	1
120		(100,150)	J(12,C13) 2	1.07	0.5-4					

**112 · Automatic Verification of Small Molecule Structure with One Dimensional Proton Nuclear Magnetic Resonance Spectrum**

24	12	4.080	(3.28,4.88)	1.31	(0.00,4.31)	2	1	J(12,15)	H12	1
	120	(100,150)	J(12,C13) 2	1.07	0.5-4					
25	13	5.765	(4.96,6.56)	2.91	(0.00,5.91)	2	1	J(13,12)	H13	1
	120	(100,150)	J(13,C13) 2	1.07	0.5-4					
26	13	5.765	(4.96,6.56)	13.77	(10.77,16.77)	2	1	J(13,14)	H13	1
	120	(100,150)	J(13,C13) 2	1.07	0.5-4					
27	14	3.276	(2.48,4.08)	13.77	(10.77,16.77)	2	1	J(14,13)	H14	1
	120	(100,150)	J(14,C13) 2	1.07	0.5-4					
28	14	3.276	(2.48,4.08)	17.23	(14.23,20.23)	2	1	J(14,15)	H14	1
	120	(100,150)	J(14,C13) 2	1.07	0.5-4					
29	15	3.220	(2.42,4.02)	1.31	(0.00,4.31)	2	1	J(15,12)	H15	0-1
	N/A	N/A	N/A 2	N/A	0.5-100					
30	15	3.220	(2.42,4.02)	17.23	(14.23,20.23)	2	1	J(15,14)	H15	0-1
	N/A	N/A	N/A 2	N/A	0.5-100					
31	16	6.884	(6.08,7.68)	4.90	(1.90,7.90)	2	2	J(16,17)	H16	1
	120	(100,150)	J(16,C13) 2	1.07	0.5-4					
32	16	6.884	(6.08,7.68)	2.67	(0.00,5.67)	2	2	J(16,18)	H16	1
	120	(100,150)	J(16,C13) 2	1.07	0.5-4					
33	16	6.884	(6.08,7.68)	2.60	(0.00,5.60)	2	2	J(16,19)	H16	1
	120	(100,150)	J(16,C13) 2	1.07	0.5-4					
34	17	7.619	(6.82,8.42)	4.90	(1.90,7.90)	2	2	J(17,16)	H17	1
	120	(100,150)	J(17,C13) 2	1.07	0.5-4					
35	17	7.619	(6.82,8.42)	7.12	(4.12,10.12)	2	2	J(17,18)	H17	1
	120	(100,150)	J(17,C13) 2	1.07	0.5-4					
36	17	7.619	(6.82,8.42)	2.60	(0.00,5.60)	2	2	J(17,20)	H17	1
	120	(100,150)	J(17,C13) 2	1.07	0.5-4					
37	18	7.196	(6.40,8.00)	2.67	(0.00,5.67)	2	1	J(18,16)	H18	1
	120	(100,150)	J(18,C13) 2	1.07	0.5-4					
38	18	7.196	(6.40,8.00)	7.12	(4.12,10.12)	2	1	J(18,17)	H18	1
	120	(100,150)	J(18,C13) 2	1.07	0.5-4					
39	18	7.196	(6.40,8.00)	2.67	(0.00,5.67)	2	1	J(18,19)	H18	1
	120	(100,150)	J(18,C13) 2	1.07	0.5-4					
40	18	7.196	(6.40,8.00)	7.12	(4.12,10.12)	2	1	J(18,20)	H18	1
	120	(100,150)	J(18,C13) 2	1.07	0.5-4					
41	19	6.884	(6.08,7.68)	2.60	(0.00,5.60)	2	2	J(19,16)	H19	1
	120	(100,150)	J(19,C13) 2	1.07	0.5-4					
42	19	6.884	(6.08,7.68)	2.67	(0.00,5.67)	2	2	J(19,18)	H19	1
	120	(100,150)	J(19,C13) 2	1.07	0.5-4					
43	19	6.884	(6.08,7.68)	4.90	(1.90,7.90)	2	2	J(19,20)	H19	1
	120	(100,150)	J(19,C13) 2	1.07	0.5-4					
44	20	7.619	(6.82,8.42)	2.60	(0.00,5.60)	2	2	J(20,17)	H20	1
	120	(100,150)	J(20,C13) 2	1.07	0.5-4					
45	20	7.619	(6.82,8.42)	7.12	(4.12,10.12)	2	2	J(20,18)	H20	1
	120	(100,150)	J(20,C13) 2	1.07	0.5-4					
46	20	7.619	(6.82,8.42)	4.90	(1.90,7.90)	2	2	J(20,19)	H20	1
	120	(100,150)	J(20,C13) 2	1.07	0.5-4					
47	21	6.948	(6.15,7.75)	12.59	(9.59,15.59)	2	1	J(21,22)	H21	1
	120	(100,150)	J(21,C13) 2	1.07	0.5-4					
48	21	6.948	(6.15,7.75)	19.83	(16.83,22.83)	2	1	J(21,23)	H21	1
	120	(100,150)	J(21,C13) 2	1.07	0.5-4					
49	22	5.408	(4.61,6.21)	12.59	(9.59,15.59)	2	1	J(22,21)	H22	1
	120	(100,150)	J(22,C13) 2	1.07	0.5-4					
50	22	5.408	(4.61,6.21)	0.27	(0.00,3.27)	2	1	J(22,23)	H22	1
	120	(100,150)	J(22,C13) 2	1.07	0.5-4					
51	23	4.895	(4.09,5.69)	19.83	(16.83,22.83)	2	1	J(23,21)	H23	1
	120	(100,150)	J(23,C13) 2	1.07	0.5-4					
52	23	4.895	(4.09,5.69)	0.27	(0.00,3.27)	2	1	J(23,22)	H23	1
	120	(100,150)	J(23,C13) 2	1.07	0.5-4					
#	IDENTICAL CHEMICAL SHIFTS AND J COUPLINGS									
#	If the chemical shifts are identical, shift ranges, proton numbers, J(het), M(het), %(het), linewidths and									
#	reliabilities also need to be identical.									
#	If Js are identical, coupling ranges and Ms also need to be identical									
#										
\$2	ep_no_1		ep_no_2							
CS	16	=	19							
CS	17	=	20							
\$3	J_1		J_2							
J	J(16,17)	=	J(19,20)							
J	J(16,18)	=	J(19,18)							
J	J(16,19)	=	J(19,16)							
J	J(17,16)	=	J(20,19)							
J	J(17,18)	=	J(20,18)							
J	J(17,20)	=	J(20,17)							
#										
\$4	CHIRAL CENTERS:									
#										
CC	N/A									
#										
\$5	THROUGH SPACE COUPLINGS:									
#										
TSC	N/A									
#										
\$6	TAUTOMERISM:									
#										
TA	N/A									

**Fig 38 An example of a simulated theoretical multiplet distribution list (the second setup)**



### 6.1.3 Experimental Design to Compute $FN'$ , $FP'$ and $CR'$

In this subsection, we introduce the approach to estimate the criteria defined in experimental setup.

#### 6.1.3.1 An approach to Compute $FN'$

Real compounds are measured with NMR spectrometers to get their NMR spectra. Naturally, each compound and its corresponding measured NMR spectrum form a consistent pair. By feeding all these consistent pairs into the system, we can calculate  $FN'$  on real compound dataset.

Similarly, the simulated spectrum and the corresponding theoretical multiplet distribution list are generated in pair, and from the principle of the simulation program (we know that) they are consistent. By feeding all these consistent pairs into the system, we get  $FN'$  on simulated dataset.

#### 6.1.3.2 An approach to compute $FP'$

To compute  $FP'$ , we need to generate enough inconsistent pairs. In a real compound dataset, a matrix of all possible structure-spectrum pairs is generated, and the consistent pairs are organized to the diagonal of the matrix. The off-diagonal elements of the matrix are all inconsistent structure-spectrum pairs. By feeding all these inconsistent pairs into the system, we can calculate  $FP'$  on the real compound dataset.

A similar matrix can also be built on the simulated dataset, where each element of the matrix represents a spectrum-theoretical multiplet distribution pair. By arranging the consistent pairs in the diagonal of the matrix, the off-diagonal of the matrix is composed of all inconsistent pairs. By feeding theses pairs into the system, we compute  $FP'$  on the simulated dataset.

#### 6.1.3.3 An approach to compute $CR'$

Computation of  $CR'$  relies on spectroscopists' manual interpretations, which makes it impossible to compute it in a big test dataset. In addition, simulated spectra are not generated from real compounds, instead they are only mapped to the simulated theoretical multiplet distribution lists. The theoretical multiplet distribution list is the intermediate result, which is supposed to be the output of the Molecular Interpreter (see 4.2) and thereby uninterpretable for human spectroscopists. Therefore, we only utilize our 85 consistent real compounds' structure-spectrum pairs to compute  $CR'$ .

Five human NMR spectroscopists worked together to manually assign NMR signals to all chemically equivalent protons in all 85 given consistent structure-spectrum pairs. On the other hand, the same 85 structure-spectrum pairs are fed into the system to give automatic assignments between the NMR signals and the chemically equivalent protons in each structure-spectrum pair. Then, each automatic assigned NMR signal- chemically equivalent protons pair is checked against to the assigned pairs given by human spectroscopists. If the pair is consistent with the assignment from spectroscopists, it will be counted as consistent assigned pair. Finally, the percentage of the consistent pairs in total automatic assigned pairs is computed as  $CR'$ .

## 6.2 Experimental Results

All experiments are run at a personal computer with Intel 2.00GHz processor, 2.00 GB RAM and Windows XP.

### 6.2.1 Experimental Results of Estimating False Negative Rate(FN)

In this section, we give the experiment results of the estimations of the False Negative on both real compound dataset and simulated datasets.

#### 6.2.1.1 Experimental Result on Real Compound Dataset

85 consistent structure-spectrum pairs are used to compute  $FN'$ . Experimental results are shown in Table5.

Input Consistent Spectrum-Structure Pairs	85
Predicted Consistent Spectrum-Structure Pairs	81
Predicted Inconsistent Spectrum-Structure Pairs	4
Estimated False Negative Rate ( $FN'$ )	0.047
Total Running Time	68 minutes
Average Running Time	48.0 Seconds

**Table 5 Experimental Result of Estimating FN on Real Compound Dataset**

### 6.2.1.2 Experimental Results of Simulated Dataset (Easy Setup)

100 consistent simulated spectrum-theoretical multiplet distribution list pairs, which are generated by the simulation program with the first setup, are fed into system to compute  $FN'$ . Experimental results are shown in Table 6.

Input Consistent Theoretical Multiplet-Structure Pairs	100
Predicted Consistent Theoretical Multiplet -Structure Pairs	94
Predicted Inconsistent Theoretical Multiplet -Structure Pairs	5
Crashed Pairs	1
Estimated False Negative Rate ( $FN'$ )	0.051
Total Running Time	11 hours 10 Minutes
Average Running Time	6.77 Minutes

**Table 6 Experimental Result of Estimating FN on Simulated Dataset (The First Setup)**

### 6.2.1.3 Experimental Results of Simulated Dataset (Difficult Setup)

925 consistent simulated spectrum -theoretical multiplet distribution list pairs, which are generated by the simulation program with the second setup, are fed into system to compute  $FN'$ . Experimental results are shown in Table 7.

Input Consistent Theoretical Multiplet-Structure Pairs	925
Predicted Consistent Theoretical Multiplet -Structure Pairs	864
Predicted Inconsistent Theoretical Multiplet -Structure Pairs	58
Crashed Pairs	3
Estimated False Negative Rate ( $FN'$ )	0.059
Total Running Time	155 hours 25 Minutes
Average Running Time	9.49 inutes

**Table 7 Experimental Result of Estimating FN on Simulated Dataset (The Second Setup)**

## 6.2.2 Experimental Results of Estimating False Positive Rate(FP)

In this section, we give the experimental results of the estimations of the False Positive on both real compound dataset and simulated datasets.

### 6.2.2.1 Experimental Results of Real Compound Dataset

85 compounds and 85 spectra pairs are used to build a 85×85 pairs matrix. The off-diagonal elements of the matrix generate 7140 inconsistent structure-spectrum pairs, which are fed into the system to compute  $FP'$ . Experimental results are shown in Table 8.

Input Inconsistent Spectrum-Structure Pairs	7140
Predicted Consistent Spectrum-Structure Pairs	234
Predicted Inconsistent Spectrum-Structure Pairs	6906
Estimated False Positive Rate ( $FP'$ )	0.033
Total Running Time	107 hours 7 minutes
Average Running Time	54.0 Seconds

**Table 8 Experimental Result of Estimating FP on Real Compound Dataset**

### 6.2.2.2 Experimental Results of Simulated Dataset (Easy Setup)

50 consistent spectrum-theoretical multiplet distribution list pairs are randomly selected without replacement from the 100 consistent spectrum-theoretical multiplet distribution list pairs generated with the first setup. Then, these theoretical multiplet lists and spectra are used to build a 50×50 pairs matrix. The off-diagonal elements of the matrix generate 2450 inconsistent spectrum-theoretical multiplet distribution list pairs, which are fed into the system to compute  $FP'$ . Correspondingly, the experimental results are shown in Table 9.

Input Inconsistent Theoretical Multiplet-Structure Pairs	2450
Predicted Consistent Theoretical Multiplet -Structure Pairs	7
Predicted Inconsistent Theoretical Multiplet -Structure Pairs	2443
Estimated False Positive Rate ( $FP'$ )	0.003
Total Running Time	7 days 23 Minutes
Average Running Time	4.68 Minutes

**Table 9 Experimental Result of Estimating FP on Simulated Dataset (The First Setup)**

### 6.2.2.3 Experimental Results of Simulated Dataset (Difficult Setup)

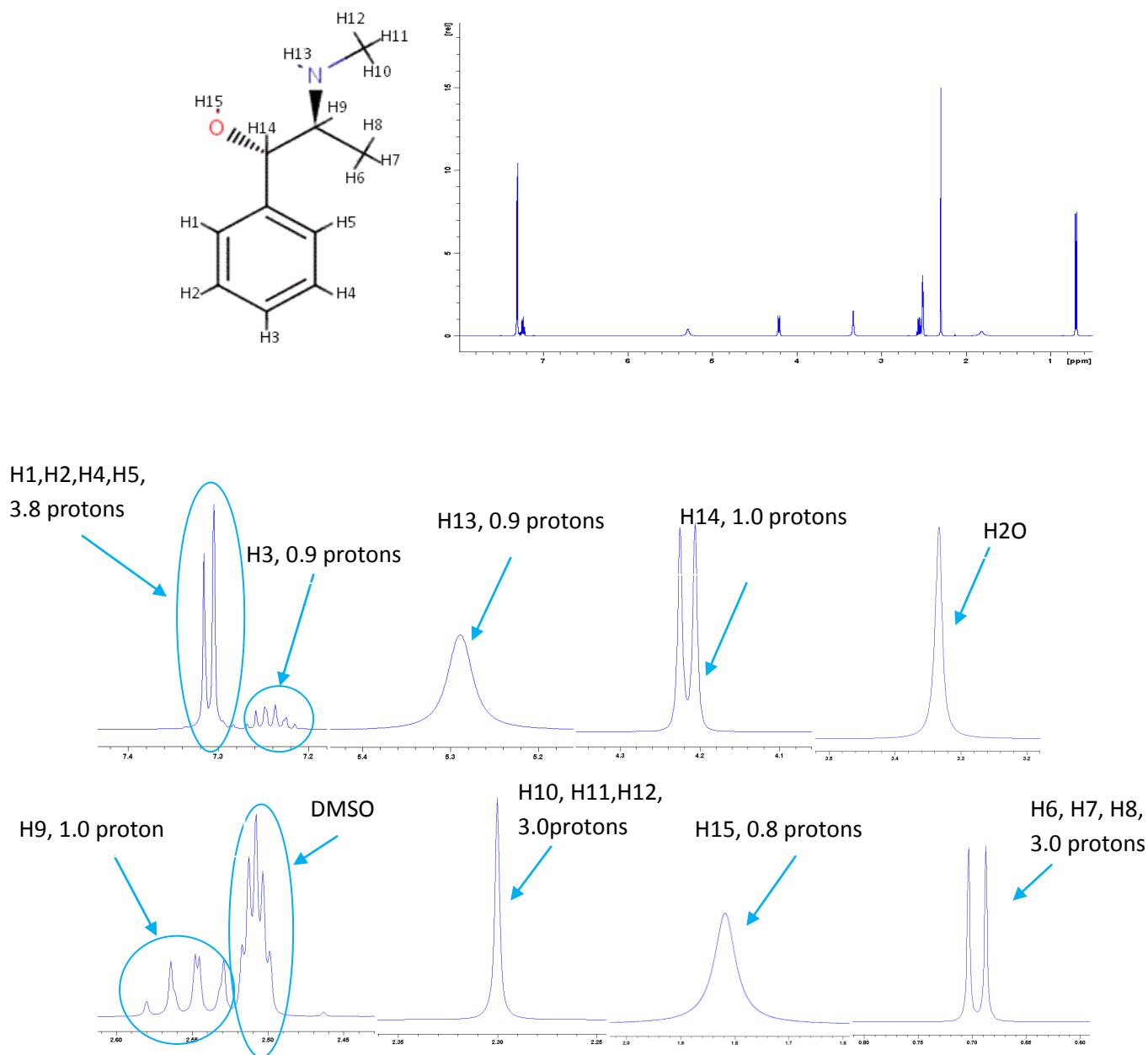
50 consistent spectrum-theoretical multiplet distribution list pairs are randomly selected without replacement from the 925 consistent spectrum-theoretical multiplet distribution list pairs generated with the second setup. Then, these spectrum-theoretical multiplet distribution list pairs are used to build a 50×50 pair matrix. The off-diagonal elements of the matrix generate 2450 inconsistent spectrum-theoretical multiplet distribution list pairs, which are fed into the system to compute  $FP'$ . Experimental results are shown in Table 10.

Input Inconsistent Theoretical Multiplet-Structure Pairs	2450
Predicted Consistent Theoretical Multiplet -Structure Pairs	27
Predicted Inconsistent Theoretical Multiplet -Structure Pairs	2423
Estimated False Positive Rate ( $FP'$ )	0.011
Total Running Time	96Days 15Hours 30 Minutes
Average Running Time	56.80 Minutes

**Table 10 Experimental Result of Estimating FP on Simulated Dataset (The Second Setup)**

### 6.2.3 Experimental Results of Estimating Consistent Rate (CR)

To clarify the meaning of assignment consistency between the system and human spectroscopists, an example of automatic assignments is demonstrated below. Specifically, automatic assignments between chemically equivalent protons and NMR signals of +-Pseudoephedrin by the system are shown in Fig 39, which demonstrates extremely high consistency between the assignments of the system and that of human spectroscopists (see Fig 19 for the assignments of the spectroscopists).



**Fig 39 Automatic assignments between NMR spectrum and structure of +Pseudoephedrin**

In 85 consistent structure-spectrum pairs (generated from real compound dataset), 81 pairs are considered to be consistent by the system. Therefore, the detail assignments of these 81 structure-spectrum pairs are used to compute  $CR'$ . The consistency analysis results are settled by human spectroscopists and presented in Table 11. Note, the detailed assignments of 81 structure-spectrum pair by the system are listed in Appendix B.

Names	DMSO	H2O	#As	correct As	comments	errors made
+Pseudoephedrin	1	1	8	8		0
--alpha-Satonin	1	1	9	5	unproblematic	4
1,1,1-3,3,3-Hexafluor-2-propanol	1	1	2	2		0
1,1,2-Trichlorethan	1	1	2	2		0
1,2,3-Trimethylbenzol	1	1	3	3		0
1,2,4-Trichlorbenzol	1	1	3	3		0
1,2-Dichlor-4-nitrobenzol	1	1	3	3		0
1,2-Dichlorbenzol	1	1	2	2		0
1,3-Dichlorbenzol	1	1	1	1	should be 2 PC	0
1,4-Dinitrobenzol	1	1	1	1		0
1-Naphthol	1	1	6	3		3
1-Octen-3-ol	1	1	6	5	make two PC @ ~5.0ppm	1
1-Octyne	1	1	5	5	do not seperate PCs	0
2-Aminopyridin	1	1	4	4		0
2-Butanon	1	1	3	3		0
2-Methyl-naphthalin	1	1	5	5		0
2-Phenoxyethanol	1	1	5	5		0
2-phenyl-ethylalkohol	1	1	5	5		0
3-Ethyltoluol	1	1	5	5		0
3-Indolepropionicacid	1	1	9	7		2
3-Phenyl-propylalkohol	1	1	6	4		2
4-Bromanisol	1	1	3	3		0
5-Fluorouracil	1	1	2	2	make two PC @ ~11.0ppm	0
6-Methyl-chinolin	1	1	6	6		0
8-Hydroxy-chinolin	1	1	6	6		0
Acetophenon	1	1	4	4		0
Adamantan	1	1	2	2		0
Allylglycidether	1	1	7	7		0
Anthracen	1	1	3	3		0
Benzaldehyd	1	1	4	4		0
Benzoesauremethylester	1	1	4	4		0
Benzonitril	1	1	3	1		2
Cumol	1	1	3	3	make two PC @ ~7.0ppm	0
D,L-Phenylalanin	1	1				0
Essigester	1	1	3	3		0
Essigsaeure-isopropyl-ester	1	1	3	3	baseline need to be improved.	0
Essigsaeurelinalylester	1	1	4	2		2
Ethanol	1	1	3	3		0
Eucalyptol	1	1	4	2	unproblematic	2
Fluorbenzol	1	1	2	2		0
Fluoren	1	1	5	1		4
Formamid	1	1	2	2	impurities??	0
Furan-2-carbonsaeuremethylest	1	1	4	4		0

Harnstoff	1	1	1	1		0
Hexamethylbenzol	1	1	1	1		0
Hexamethyldisiloxan	1	1	1	1	too big PC!	0
Hydrochinon	1	1	2	2		0
Imidazol	1	1	3	3		0
Isopropanol	1	1	3	3		0
L-+-Rhamnose-Monohydrat	1	1	9	7	OH exchanged; unproblematic	2
Linalool	1	1	9	9	impurity in PC @ 1.55	0
L-Leucin	1	1	6	5		1
L-Prolin	1	1	3	0		3
Maleinsaeure-dibutylester	1	1	5	5		0
Maleinsaeure-diethylester	1	1	3	3		0
Malonsaeure	1	1	1	1		0
Melamin	1	1	1	1		0
Mesiylen	1	1	2	2		0
Methacrylsaeure-2-ethoxyethylester	1	1	6	4		2
Methacrylsaeure-butylester	1	1	6	6		0
Methacrylsaeure-isobutylester	1	1	4	4		0
N,N,N,N-Tetramethyl-ethylendiamin	1	1	1	1		0
N,N-Dimethylacetamid	1	1	3	3		0
Naphthalin	1	1	2	2		0
Naringenin	1	1	9	6	seperate PC @ 3.25ppm; unproblematic	3
Nicotinsaeureamid	1	1	4	4	seperate PC @ 8.1 & 7.5 ppm	0
N-Methylformamid	0	1	2	2	seperate PC @ 8.0 & 2.6 ppm; DMSO not found	0
Phenol	1	1	3	3		0
Phenolphthalein	1	1	7	3		4
Phenylethylenoxid	1	1	3	3		0
Propionsaeureethylester	1	1	4	4		0
Propionsaeuremethylester	1	0	2	2		0
Propylbenzol	1	1	5	3	unproblematic	2
Pseudocumol	1	1	5	3	unproblematic; aromatics exchanged	2
S+-2-4-Isobutylphenylpropionsaeure	1	1	8	2	4 severe and 2 unproblematic errors; aromatics unp.	6
Salicylaldoxim	1	1	6	4	unproblematic	2
Sulfamid	1	1	1	1		0
tert-Butylalkohol	1	1	2	2		0
THF	1	1	2	2		0
Triethylamin	1	1	2	2	split PC @ DMSO	0



Triethylphosphat	1	1	2	2		0
			309	260	CR = 84.14%	49

**Table 11 Experimental Result of Estimating CR on Real Compound Dataset**

In Table 11, Table Item: “DMSO” presents the identification status of the DMSO signal by the system - “1” represents a correct identification, while “0” represents an incorrect identification. Table Item: “H<sub>2</sub>O” presents the identification status of the H<sub>2</sub>O signal by the system - “1” represents a correct identification, while “0” represents an incorrect identification. Table Item: “#As” presents the total number of assignments given by the system. Table Item: “correct As” presents the number of the assignments given by the system which are consistent with those of human spectroscopists. Table Item: “comments” presents the additional comments from spectroscopists above the system’s assignments. (refer to Appendix B for detail system’s assignments) Table Item: “error made” presents the number of assignments wrongly made by the system.

From Table 11, we see that there are totally 309 assignments which are made by the system upon 81 structure-spectrum pairs. Wherein, 260 assignments are consistent, and 49 assignments are inconsistent, and this gives us the estimation of CR as 84.14%. Note, in 49 inconsistent assignments, there are 19 cases commented as “unproblematic”, which means even the system gives the different assignments to the assignments of spectroscopists, these differences are reasonable and acceptable by spectroscopists. If we added these “unproblematic” cases into the consistent assignment set, we would have totally 279 consistent assignments in 309 system’s assignments. This would give us the estimation of CR as 90.29%. Nevertheless, both estimations give us a good indication to show the high consistency between the system and the spectroscopists.

## 6.3 Discussion of the Experimental Results

In this subsection, we discuss the experiment results along the decision accuracy, the time complexity and the consistency to human spectroscopists.

### 6.3.1 Decision Accuracy

Table 5 and 8 give us the estimated false negative rate (FN) of 0.047 and false positive rate (FP) of 0.033 on the real compound dataset. These results demonstrate that the two types of errors measured of the system of the real compound dataset are controlled within the 5% error rate bar. Hence, the accuracy of the system is significantly higher than 90%, which satisfies the goal (requirement) a and b defined in the beginning of Chapter 3. Note, the benchmark of 90% accuracy was defined through careful discussion among NMR spectroscopists and compound library management experts from our industrial cooperator and several pharmaceutical companies.

Furthermore, Table 6 and 9 give us the estimated false negative rate (FN) of 0.051 and false positive rate (FP) of 0.003 of the simulated dataset (easy setup). Table 7 and 10 give us the estimated false negative rate (FN) of 0.059 and false positive rate (FP) of 0.013 of the simulated dataset (difficult setup). These additional experimental results over simulated datasets further confirm that both types of errors of the system are well within the 5% error rate bar, and thereby prove the accuracy of the system to be above 90%.

With above results, we have experimentally confirmed that the system reaches the decision accuracy which satisfies the goal set by industrial participants.

### 6.3.2 Time Complexity

Table 5 shows that given the structure and the spectrum consistent, average running time per spectrum is about 22.3 seconds, and Table 8 shows that given the structure and the spectrum inconsistent, average running time per spectrum is about 32.4 seconds. Both time expenses well-satisfy the requirements of compound library management.

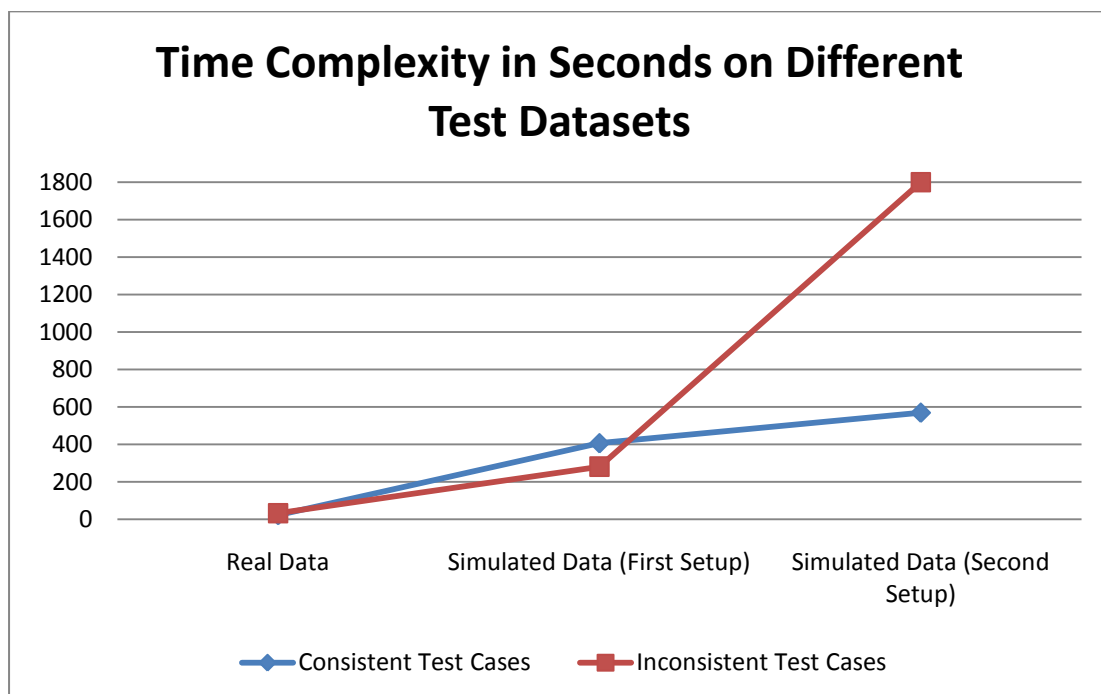
Currently the pharmaceutical industry utilizes Liquid Chromatography-Mass Spectrometry (LC-Mass) technique as the standard approach for automatic quality control of their compound libraries. The advantage of the technique is that Mass spectra are automatically interpretable, and in fact the automatic interpretation itself cost instantaneous time (within a second). However, Mass spectra technique suffers from supplying limited information about molecular structure, and from its inability for quantification (see Chapter 1). Therefore, it cannot supply enough accuracy for structure verification tasks. In addition, LC-Mass is a slow technique and takes on average 8 minutes to acquire the Mass spectrum of the sample to the best of our knowledge. This is mainly contributed to the slow infiltration of Liquid Chromatography. Comparatively, NMR requires less than a minute to acquire a 1D  $^1\text{H}$  NMR spectrum. This gives us a margin of about 7 minutes for automatic NMR spectrum molecular structural consistency analysis. Therefore, the average time expense of the system essentially demonstrates the speed advantage of NMR technique compared to that of LC-Mass technique. From the experimental results, we see that the system takes roughly a half minute on average to execute a structure NMR spectrum verification task of the real compound test dataset. This is the time expense, which is dramatically smaller than 7 minutes, and therefore strongly demonstrates the feasibility that NMR-based automatic structure verification is faster than that of LC-Mass based technique.

For the simulated dataset (easy setup), the average running time for a consistent test case is about 6.77 minutes, and the average running time for an inconsistent test case is about 4.68 minutes. The average time expenses on the simulated dataset (easy setup) are dramatically higher than that on the real compound dataset. This is due to the complexity of the simulated dataset which is designed to surpass that of the real compound dataset. With this additional complexity, the system often takes more time to search for a reasonable solution. Nevertheless, average time expense on the simulated dataset (easy setup) is still within the 7 minutes time margin, and thereby comparable to that of LC-Mass technique.

For the simulated dataset (difficult setup), the average running time for a consistent test case is about 9.49 minutes, and the average running time for an inconsistent test case is about an hour. The average time expense on the simulated dataset (difficult setup) becomes higher than that of the LC-Mass approach. This especially happens to the inconsistent test cases. However, for the following reasons, we believe that it is still acceptable for the pharmaceutical industry, and in practical application, the average time expense should be significantly smaller than an hour:

1. The simulated dataset (difficult setup) is designed to simulate very complex chemical compounds, which don't often appear in the routine compound library management.
2. The experiment was conducted on a PC with 2.00GHz computational power, which is significantly slower than the computer used to control NMR spectrometer and process NMR data.
3. Advances in computational speed according to Moore's law should half the execution time every 18 months.

Considering these three factors, we believe that the time expense of the system is not a big issue. With some suitable investment into computer hardware, even the time expense of the system on complex chemical compound can be effectively controlled on the level of the time expense of LC-Mass technique.



**Fig 40 Average Time Expenses on Different Datasets**

An observation on the time expenses is that the average running time on the real compound dataset is significantly smaller than the running time on the simulated dataset (easy setup). The same also applies to the running time on the simulated dataset (easy setup) is significantly small than the running time on the simulated dataset (difficult setup)(see Fig 40). This is due to the fact that the simulated datasets are designed to simulate more complex compounds than compounds we meet in the compound library management environment. This complexity of the compound results in a larger search space built by the system, and thereby increases the time consumption of the heuristic search (optimization) approach. This conjecture could be further confirmed by the fact that the running time for the simulated dataset (difficult setup) is longer than the running time for the simulated dataset (easy setup).

Another observation of the time expenses is that for the same dataset, the average running time for the consistent test cases often is shorter than the running time for the inconsistent test cases. This trend is not very obvious for the real compound dataset and the simulated dataset (easy setup), but very visible for the simulated dataset (difficult setup). Probably the search space for the real and the (easy setup) simulated dataset is reasonably small, so that the time expense is not big, and the difference between partially searching the space for a solution and completely searching the space for a solution are not significantly big either. Therefore, the time expense over the whole structural verification tasks is mainly contributed by the NMR spectrum interpretation (e.g. peak picking, multiplicity analysis) instead of searching for a consistency analysis solution itself. On the other hand, the simulated dataset (difficult setup) represents more complex compounds. This induces the system to build a larger search space, which makes a complete search impossible. Consequently the time expense on the consistent test cases is significantly smaller than that of the inconsistent cases. Specifically, a consistent test case implies the existence of a solution in the search space. A well-designed heuristic search can find a solution quickly without scanning the whole space. Conversely, an inconsistent case implies the nonexistence of the solution in the search space. Thereby, no matter how good the heuristic approach is, it has to scan the whole space before being able to confirm the nonexistence of the solution. In practice, the search space could be big enough to make a whole scan of it impossible. Nonetheless, the search heuristic will still scan a significantly large part of the space before it decides to give up.

In summary, the difference of the search efficiency of the consistent test cases and the inconsistent cases directly causes the different time expenses (especially of the complex compounds) of the consistent test dataset and the inconsistent test dataset. This conclusion is experimentally confirmed by the time expenses on the simulated data (difficult setup), where for a consistent case, the average running time is 9.49 minutes, but for an inconsistent case, the average running time is about an hour.

### 6.3.3 Assignment Quality, Consistency between the System and Spectroscopists

Table 11 give us the estimated consistency rate (CR) of 84.14% on the 81 real compound structure-spectrum pair. This result demonstrates that the system is highly consistent with human

spectroscopists in detail NMR property analysis and assignments. Could we optimistically understand this result as the system reaches 84% capacity of human spectroscopists on structural verification tasks?

The answer is “No”. Specifically, the consistency rate (CR) gives us a good indicator to measure the consistency between the system and spectroscopists. But it is not the only indicator. There are additional indicators, which could be used to measure the consistency between the system and spectroscopists. To list some of them, we could measure the consistency between the system and spectroscopists on spectral baseline selection, on peak cluster identification, on peak cluster multiplicity analysis, on impurity identification, etc. Note, in Table 11, the comments of spectroscopists are given to indicate the deviation of the system from spectroscopists on these measurements which are beyond the CR.

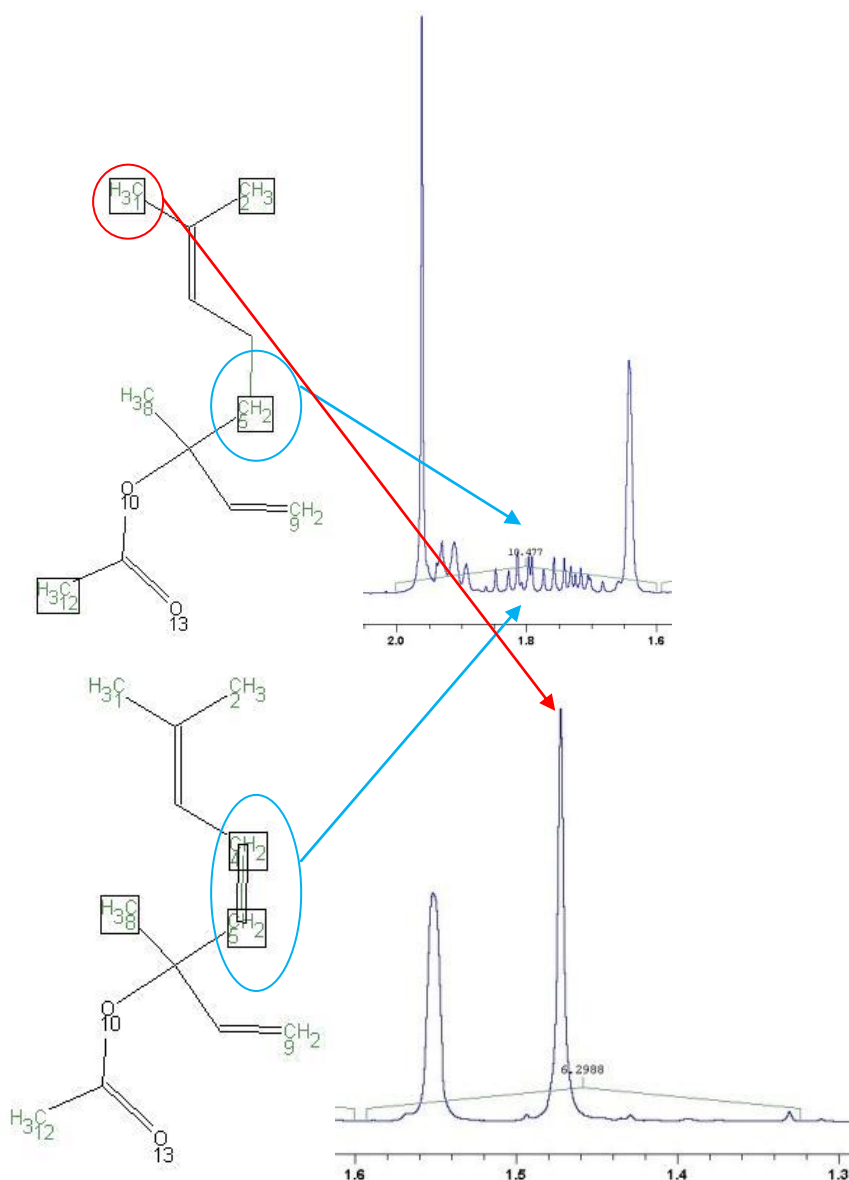
In addition, even if we rely on CR to evaluate the consistency between the system and spectroscopists, we still don't know whether the CR of 84.14% has reasonable persuasion to convince spectroscopists, since there are no quantitative requirements about consistency rates, which has been proposed by our industrial cooperation partner, and has historically not made record. However, through more than three years negotiation with NMR spectroscopists, we clearly understand that the only way to completely convince them of the effectiveness of the system is to show them that the system does the job exactly like them. From this point of view, the system needs to be continually improved to convince spectroscopists, since there is still a margin of about 15% mistakes. Nonetheless, the consistency rate of 84.14% on the assignments does motivate all spectroscopists involved in the project to believe that it is possible to reach a higher consistency rate (e.g. 95%) with some suitable improvements. A consistency rate (CR) of 95% (as what they believe) should be enough to convince the decision makers in the management level of pharmaceutical companies to decide using the system as the supplement of NMR spectrometer to replace LC-Mass technique on automatic structural verification tasks of compound library management.

Table 11 shows that the system makes 49 inconsistent assignments. Except 19 “unproblematic” cases, there are a total of 30 cases where the system makes incorrect assignments. To pursue the reason, the majority mistakes of these 30 cases come from insufficient J-coupling analysis. Specifically, there are two scenarios where the system would not execute J-coupling analysis:

1. While a peak cluster hypothesis is mapped to multiple theoretical multiplet distributions, the system would expect that multiple first-order multiplets overlap altogether in the spectrum so that the experimental multiplicity of individual multiplet becomes unsolvable. Therefore, the system won't execute J-coupling analysis on the peak cluster hypothesis.
2. While a peak cluster hypothesis is mapped to a theoretical multiplet distribution, which is generated by the chemically equivalent but magnetically inequivalent protons, the system would expect that the peak cluster hypothesis shows the high-order multiplet pattern which is beyond the first-order multiplet analysis. Therefore, the system won't execute J-coupling analysis on the peak cluster hypothesis.

Compared with the system, spectroscopists show a more flexible pattern recognition ability, which helps them to reduce the ambiguity when they meet the two scenarios. To illustrate these advantages of spectroscopists, and reveal the weaknesses of the system, we utilize two examples to explain the two scenarios. Specifically, in Fig 41, the system assigns proton groups 1, 2, 5, 12 to the

peak cluster on the top, and assigns proton groups 3, 4, 5 to the peak cluster on the bottom. Clearly, this example belongs

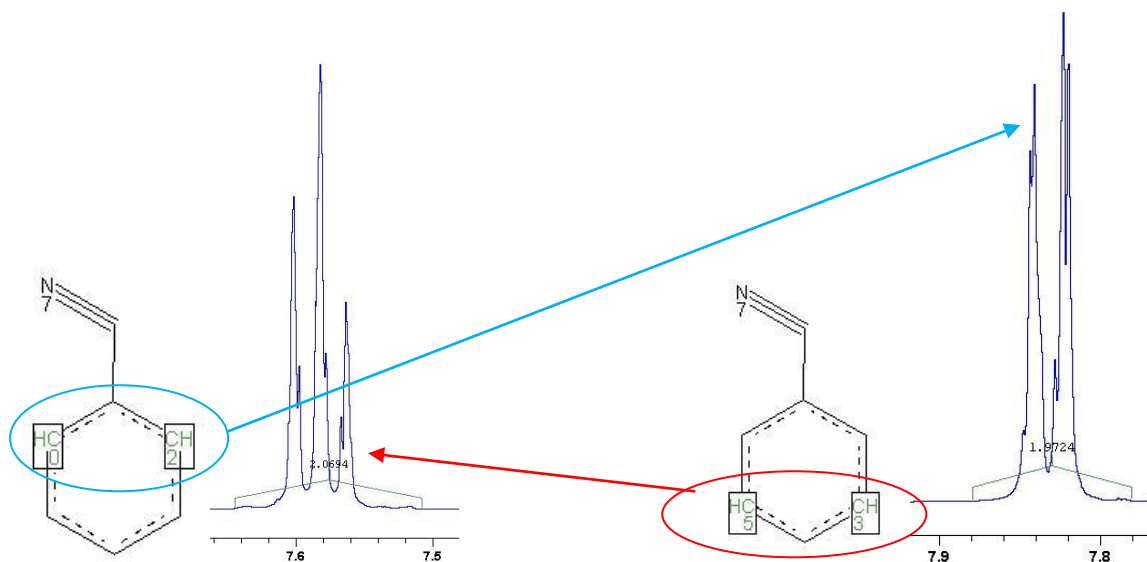


**Fig 41** Wrong assignments by the system, and their corrections on Essigsaeurelinallylester

to the first scenario, where multiple proton groups are assigned to a single peak cluster. Therefore, the system only relies on chemical shift and proton numbers to do the assignments. Unfortunately, all these proton groups are shown in similar chemical shift positions, and thereby make them undistinguishable with the information on chemical shift. As a result, any grouping of these proton groups which are consistent on proton numbers are considered as the reasonable assignments by the system. The assignments shown in Fig 41 are one of them and are indeed consistent in both chemical shift and proton number. However, through discussion with spectroscopists, we know that

the assignments are wrong. Specifically, spectroscopists will assign proton group 1 to the right peak of the peak cluster at the bottom, and assign proton group 4 and 5 to the middle part of the peak cluster at the top. This is owed to spectroscopists' ability to check multiplicities on overlapped multiplets. To pursue the reason, from the structure spectroscopists know that proton group 4 and 5 are coupled to each other to show complex multiplicity patterns. Therefore, it is impossible to assign them to the peak cluster at the bottom, which only shows two singleton patterns. On the other hand, the signal patterns in the middle of the peak cluster on the top shows complex multiplet pattern, and thereby it is reasonable to assign these proton groups to it. With this additional check, spectroscopists end with the correct assignments, while the system makes the wrong assignments.

In Fig 42, the system assigns proton group 0, 2 to the peak cluster at the left, and assigns proton group 3, 5 to the peak cluster at the right. Clearly, this example belongs to the second scenario, where chemically equivalent but magnetically inequivalent protons are assigned to a single peak cluster. Therefore, the system only relies on chemical shift and proton numbers to do the assignments. Unfortunately, two groups (0, 2 and 3, 5) are identical with the measurement only upon chemical shift and proton number. This makes them undistinguishable by the system. Consequently, the system will arbitrarily select assignments among them. In contrast, spectroscopists can identify the subtle difference between the two groups. Specifically, with the help of a J-coupling analysis, spectroscopists know that there is an additional proton which will cause the splitting of the NMR signal of proton group 3, 5, and therefore make the NMR signal of proton group 3, 5 showing more complex multiplet pattern to that of proton group 0, 2. Clearly, the peak cluster at the left shows a triplet-like signal pattern, which is more complex than that of the peak cluster at the right, which shows a doubleton-like pattern. This gives spectroscopists enough evidence to assign proton group 0, 2 to the peak cluster at the right, and assign proton group 3, 5 to the peak cluster at the left.



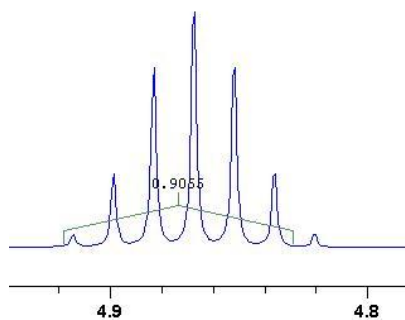
**Fig 42 Wrong assignments by the system, and their corrections on Benzonitril**

The above two examples show that to reach the level of the accuracy of human spectroscopists, more advanced signal pattern recognition techniques (which are beyond first-order multiplet analysis) are needed to be designed and added into the system. We believe that a good starting point would be to start a first-multiplicity analysis upon the overlapped NMR signals and magnetically inequivalent NMR signals.

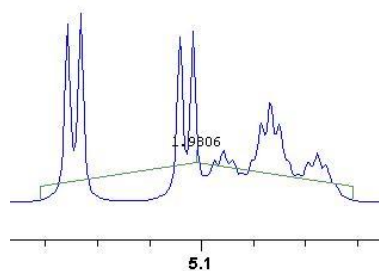
Based on the distribution of errors in the real dataset (see Table 11), we can safely say that the system can reach the consistent rate (CR) of 95% in the real dataset with above improvement. Anyhow, a lot more research is needed to be done before the system is mature enough for the practical evaluation, and we will leave the discussion of them to future work in Chapter 9.

In addition, as the supplement of the experimental result shown in Table 11, there are some other inconsistencies between the system and spectroscopists, which can't be measured by the consistency rate (CR). Fig 43 to Fig 45 gives us three examples. In Fig 43, the inconsistency between the system and spectroscopists comes from the baseline identification. With baseline set too high, the system does not put the most right peak in Fig 43 to the peak cluster. As the result, even the assignments of the system were correct, spectroscopists would still doubt about the correctness of the system, if they saw the case in Fig 43. Similarly, in Fig 44, the peak cluster should be split to two peak clusters, and in Fig 45, the two peak clusters should not be split. These deviations of the system also hamper spectroscopists' confidence about the system's performance.

From these counterexamples, we summarize that to build a mature system, which is accepted by spectroscopists, a lot of detail engineering works still needs to be done to further improve the components of the system, which cause these deviations.

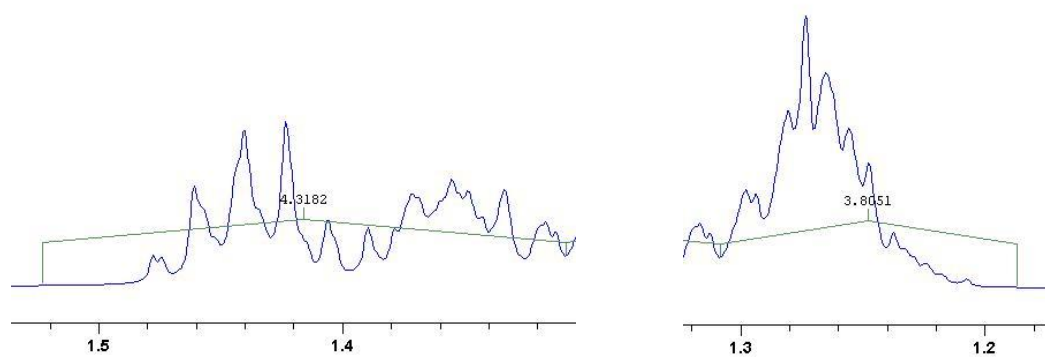


**Fig 43 Baseline Problem on Essigsaeure-isopropyl-ester**



**Fig 43 Peak Cluster should Split on Linalool**





**Fig 44 Peak Cluster should not Split on 1-Octyne**



## **Part III**

### **Contribution, Limitation, Future Work and Conclusion**



## **Chapter 7 Contribution**

This chapter introduces the potential impact of our work on the NMR and pharmaceutical industry. As the contribution to applied computer science, it also proposes a human-logic based optimization strategy, and compares it with several classical optimization approaches. We hope that the human logic based strategy could be utilized by other computational participants to solve similar problems in various application domains, especially in fields where computer could be used to replace human experts.

### **7.1 Impact for NMR and Pharmaceutical Industries**

In this section, we explain the potential impact of the successful system on NMR and pharmaceutical industry.

#### **7.1.1 Impact on the NMR industry**

NMR is the most comprehensive technology for molecular structure identification in modern world. For a long time it suffered the disadvantages of a high price, high operational cost and complex spectrum interpretation procedure. High price and high operational cost keep it away from extensive routine molecular structure identification tasks in drug discovery, drug production and drug quality control processes. The fact that NMR interpretation relies on highly educated and experienced people limits its application in universities and research institutions. As the result, most of molecular structure identification and verification tasks in production and QA/QC processes still utilize other (cheaper but easier to operate) analytical instrumentation techniques such as LC-Mass techniques.

In order to popularize NMR technique into the routine molecular structure identification tasks, over the past 30 years, NMR manufactures have been constantly improving the spectrometer hardware, and reducing production cost. As a result, the low end NMR spectrometer has a price comparable to the price of LC-Mass spectrometer. With the breakthrough in detector (probe) technology, the operational cost of NMR spectrometer is also dramatically reduced. Through careful and detailed market research, NMR manufactures get the conclusion that the total cost (including spectrometer price and operational cost) of low end NMR is reduced to the same level as the cost of LC-Mass technology. Therefore, the only bottleneck left is the complexity of NMR spectrum interpretation. To take over this big market, which used to belong to LC-Mass technology, NMR manufactures need an automatic NMR spectrum interpretation software to reduce requirements on human quality. Ideally, they want the spectrum interpretation to be fully automatic so that the requirements on human spectroscopists are reduced to a minimum. Since >99% of the structure identification or verification tasks in routine production and QA/QC processes utilize 1D <sup>1</sup>H NMR spectrum, the primary goal of NMR manufactures is focused on automation of 1D <sup>1</sup>H NMR spectrum interpretation.

In the range of this thesis, we built a fully automatic molecular structure 1D <sup>1</sup>H NMR spectrum verification system, which includes the fully automatic 1D <sup>1</sup>H NMR spectrum interpretation, and fully automatic NMR signal-structure proton assignments. The experimental results show that the system reaches the decision accuracy which is acceptable to be used as the automatic structure verification tool in industry. It also demonstrates that the total time expense of NMR acquisition and interpretation is comparable to the time expense of the LC-Mass technology. In addition, to some extent, it exhibits the consistency between the system and spectroscopists. Conclusively, as the prototype, the system proves the feasibility of automating the structural verification procedure, and thereby taking over the last barrel of applying NMR technology in routine structural verification tasks. Based on our system, reliable commercial software are under development, and are designed to be embedded into NMR spectrometer control software to ease the complexity of spectrum interpretation and structural verification. To summarize, we hope that by adding this software into the NMR system, NMR technology could be pushed to replace LC-Mass in structure verification tasks of routine production and QA/QC process, and finally increase the market share and application scope of NMR technology in life science industry.

### 7.1.2 Impact on the Pharmaceutical Industry

It is crucial to guarantee the effectiveness of the drug discovery process to insure the quality of the compound library. It is the long term interest for the compound library management participants to seek new approaches to improve the quality of the compound library. NMR technology has obvious superiorities for structure verification tasks. Therefore, it is used as the arbitrate technology to supplement the analysis of LC-Mass technology. Due to the expensive human effort consumption, majority structural verification tasks still rely on LC-Mass technology. As a result, there is long term desirability to automate the NMR spectrum verification process in the compound library management.

With the experimental results of our system, it seems that the system can reach 90% of the human spectroscopists' consistency analysis decision accuracy. And the experimental results also show more than 80% consistency between human spectroscopists and the system in assignments. This result demonstrates that the system could be used to replace human spectroscopists in structure verification tasks to a great extent. Therefore, the automation based on the system is close to be mature enough to be used to dramatically reduce the human efforts in the structural verification process. With this new automation, it is possible to use NMR to replace LC-Mass for routine structure verification tasks in compound library management. As the result, relying on NMR technology and high consistency between the automatic NMR spectrum analysis system and NMR spectroscopists, the quality control level of compound library will be qualitatively improved. Finally, this will in turn improve the effectiveness and the efficiency of the drug discovery process.

## 7.2 Contribution to Computer Science

The core of the system is an optimization routine. The optimization is based on mimicking spectroscopists' human decision logic, which distinguishes itself from other optimization approaches. In fact, some optimization approaches (e.g. simulation annealing, Markov chain Monte Carlo, etc.), have been utilized to address the problem in the past, but failed. We believe that the inabilities of these optimizations are due to the simplicity of their embedded heuristics design, and the lack of human-like reconsideration mechanism. Therefore, the optimization is designed to get over these inabilities. To explain these characteristics of the optimization in detail, we illustrate our optimization policies with an example in 7.2.1. In 7.2.2, we analyze the difference between our optimization policies and other optimization methodologies.

### 7.2.1 Human Logic Based Optimization – a Demonstration

This section shows an example to explain how the optimization process works. Specifically, in Fig 46, we abstract the problem setup by omitting its NMR interpretation. Therein, Input List 1 and Input List 2 represent two sets of elements which need to be matched to each other, while there are additional constraints defined on Input List 2. (Note, to map this abstract setup to our NMR structural verification problem, Input List 1 represents the peak cluster hypothesis space, Input List 2 represents the theoretical multiplet distribution list, while constraints defined on Input List 2 represents connectivity among theoretical multiplet distributions. Therefore, the match between Input List 1 and Input List 2 represents searching for reasonable consistent assignments between the peak cluster hypotheses space and the theoretical multiplet distribution list.) To illustrate the work flow of the optimization, Fig 47 to Fig 52 demonstrate a simulation of a sequential match between Input List 1 and Input List 2. Specifically, in Fig 47, pair-wise matches between Input List 1 and Input List 2 select the best matched pair (A, 1) as the initial part of the solution. Then, in Fig 48, the algorithm continually searches for the next best matched pair (D, 4) and adds it to the solution. Note, there is no constraint defined between 1 and 4, and thereby the solution (A, 1) (D, 4) is still consistent. Continually, in Fig 49, the algorithm keeps searching for the next best matched pair (C, 3+5) and adds it to the solution. But now, there are constraints defined between 1 and 3, 3 and 4, 3 and 5, and thereby the consistencies between A and C, C and D need to be checked. Due to the fact that A and C are not matched, pairs (A, 1) and (C, 3+5) are deleted from the solution. As the result, only pair (D, 4) is left in the solution. Note, pairs (A, 1) and (C, 3+5) are deleted from the solution, but are not deleted from the search space. Instead their priorities to be reselected into the solution are reduced. This is designed to mimic human's logic of reconsideration. Next, in Fig 50, the algorithm finds the next best matched pair (B, 3) from the search space, and adds it to the solution. With a constraint defined between 3 and 4, A and D are checked and found to be consistent. Therefore, the solution now includes (D, 4) and (B, 3). In Fig 51, (A, 2+5) is added to the solution. In Fig 52, pair (C, 1) is added into the solution. Since all elements in Input List 2 are reasonably explained by elements in Input List 1, and constraints defined on Input List 2 are satisfied by the elements of Input List 1, match is complete. Hence, the complete solution is shown in Fig 53.

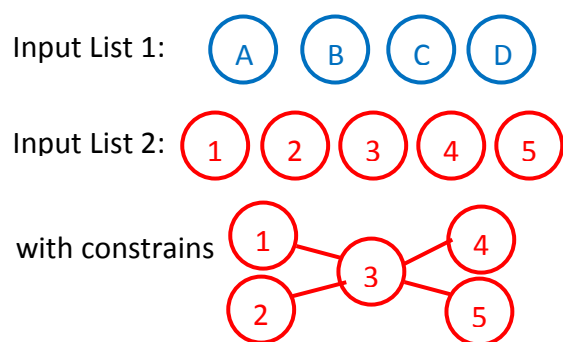
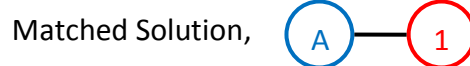


Fig 45 Demo –Input (a)

Step 1:

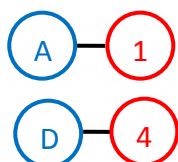


Result, Consistent.

Fig 46 Demo –Step1 (b)

Step 2:

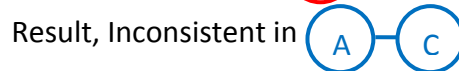
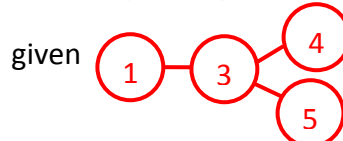
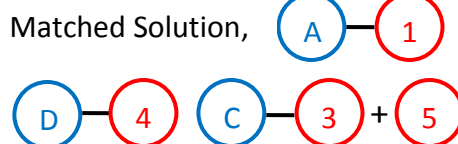
Matched Solution,



Result, Consistent.

Fig 47 Demo –Step2 (c)

Step 3:



Delete



Fig 48 Demo –Step3 (d)



Step 4:

Matched Solution,



Check



given

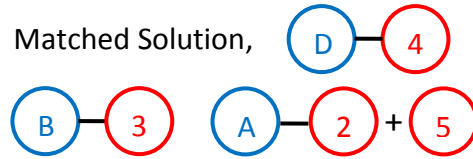


Result, Consistent.

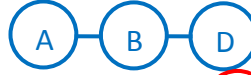
Fig 49 Demo –Step4 (e)

Step 5:

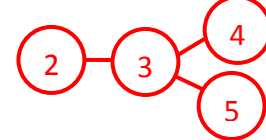
Matched Solution,



Check



given

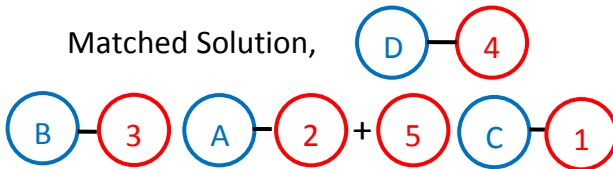


Result, Consistent.

Fig 50 Demo –Step5 (f)

Step 6:

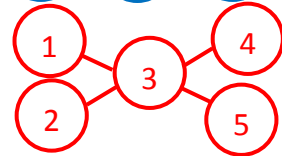
Matched Solution,



Check



given



Result, Consistent.

Fig 51 Demo –Step6 (g)

Output:

Final Solution,

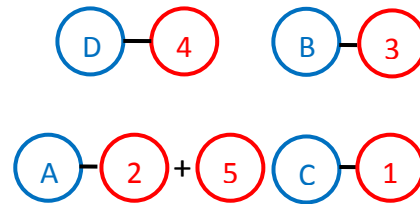


Fig 52 Demo –Output (h)

From the above demonstration, we summarize that there are three characteristics of the human logic based optimization, which distinguishes it from other optimization approaches.

- The human logic based optimization sequentially builds a solution instead of searching for a solution in the solution space. This is similar to majority heuristic search algorithms like those that are designed for finding the shortest path in a graph (see (Russell, et al., 2002)).
- The human logic based optimization contains a mechanism to “shrink” the solution. This mechanism is similar to the back-tracking mechanism embedded in the deep-first search algorithm (Sedgewick, 1997). But they are different in essence. Specifically, in the back-tracking mechanism, the solution is “shrunk” by returning back along the previous path. In

the human based logic, the solution could be “shrunk” to a status which is never been searched before (For the detail discussion of the difference, see 7.2.2).

- c. In the human based logic, the deleted part of the solution could be reconsidered again. This mechanism is similar to random walk police adopted by most stochastic optimization algorithms, since any part of searching space has chance to be traveled again. (Note, we assume that the searching space is connected, see 7.2.2.) But they are distinguished from each other in principal. With the random walk police, a dice is thrown in each searching status to decide which status to go next. But with the human based logic, revisiting a previously visited status is based on the maximum likelihood heuristics, and there is no random component involved.

## 7.2.2 Human Logic Based Optimization versus Classical Optimization

In this subsection, we analyze the difference of the human logic based optimization from some classic optimization methods.

### 7.2.2.1 Representation of Problem as Graph Search

Given the problem setup shown in Fig 46, optimization approaches convert the problem setup to a heuristic search. Here, heuristic search means a cluster of search strategies which utilizes problem-specific knowledge to make the search of the solution efficiently (Russell, et al., 2002).

Obviously, the search strategy design relies on the structure of the search space. Therefore, to discuss the advantages and the disadvantages of different search strategies, the first task should be to reasonably define a search space to represent the problem. With the problem setup in Fig 46, we believe that there are three ways to arrange the structure of the search space. To make it easily understand, we continue to use the example in 7.2.1 to illustrate the structure of the search space. Note, we represent the problem setup presented in Fig 46 to Fig 54, with some simplifications: we limit the problem to only contain one-to-one mappings.

#### Search Space Structure I

An undirected graph is built to represent the searching space, where each possible pair between Input List 1 and Input List 2 is represented as a graph node. With this structure of the search space, the problem of building a solution for the setup in Fig 54 is converted to the problem of searching for a reasonable path in the graph (see Fig 55).

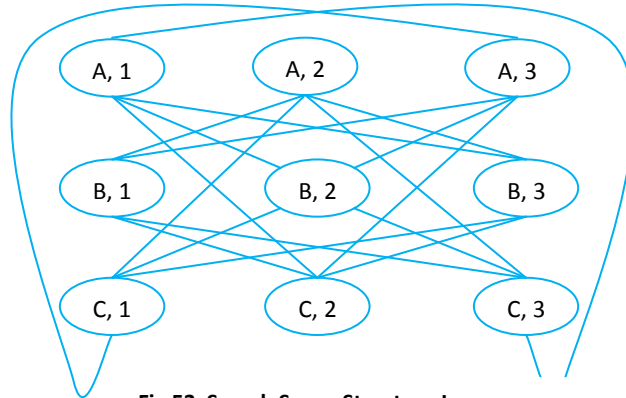
Input List 1: **A** **B** **C**

Input List 2: **1** **2** **3**

with constraints



**Fig 54 Problem Setup**



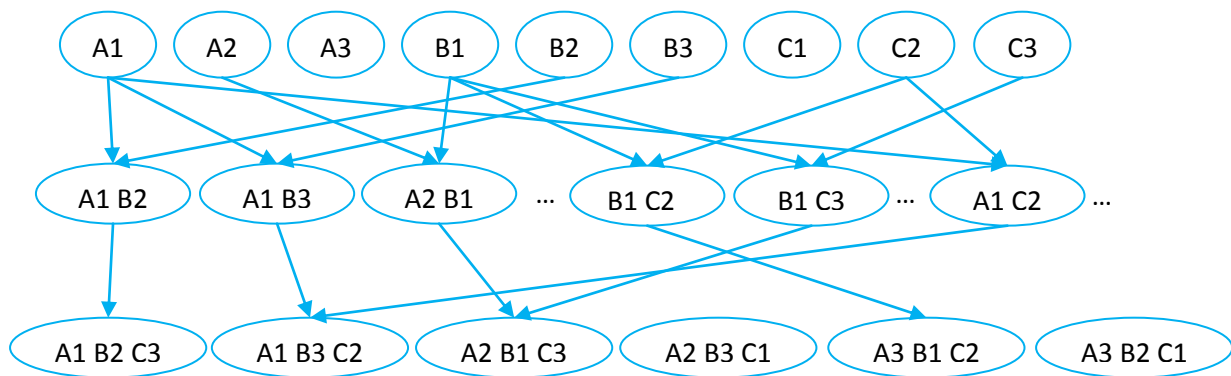
**Fig 53 Search Space Structure I**

### Search Space Structure II

A directed graph is built to represent the searching space, where a graph node represents a possible subset of all pairs. Note,

- (1) In the graph there are one-to-one mappings between all graph nodes and all possible subsets of all pairs.
- (2) There is no circle in the graph.

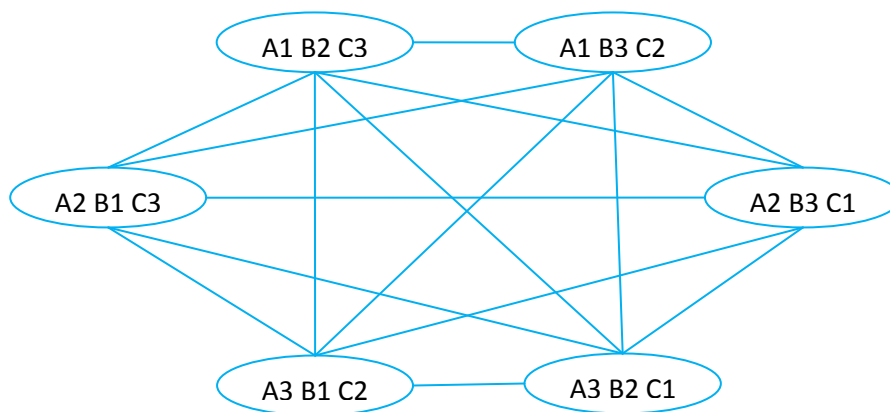
With this structure of the search space, the problem of building a solution for the setup in Fig 43 is converted to the problem of starting from a reasonable “source” node to search for a reasonable “sink” node in the graph (see Fig 56).



**Fig 55 Search Space Structure II**

### Search Space Structure III

A fully connected undirected graph is built to represent the searching space, where each graph node represents a complete match between List1 and List2 (a possible solution). All these solution nodes are connected altogether to construct a fully connected graph. Then, the problem of building a solution for the setup in Fig 54 is converted to the problem of searching for the optimal solution node in the fully connected graph (see Fig 57).



**Fig 56 Search Space Structure III**

#### 7.2.2.2 Difference between Human Logic Based Optimization and Best First Search

Generally speaking, best first search describes a subset of the general Tree-Search or Graph-Search algorithm in which the next tree or graph node is selected to the searching track based on an evaluation function, which is designed by applying problem-specific knowledge, in order to find a solution efficiently. For example, both greedy search (Russell, et al., 2002) and A\* search (Russell, et al., 2002) belong to the range of best first search.

Best first search is a fundamental search technology, which is universally applicable to different constructions of search space. Therefore, it can be used to search for a solution in any of three graph representations of our problem setup. Greedy search and A\* search are two typical best first search approaches, and therefore, we focus our discussion on their utilities in our problem setup. First, we explain the difference between greedy search and A\* search. Greedy search is a primitive technology, which selects the next graph node to add into the searching path only based on the information in the candidate nodes. Specifically, the candidate node which maximizes the utility of reaching the searching goal will be selected into the search path. Clearly, the weakness of this searching strategy is that it completely ignores the information in the past search path. This “Markov” property makes it incompetent for tasks of searching for an optimal path in graph, where all graph nodes belonging to the path have to be evaluated together, and this requires a memory of previous historical search paths. Comparatively, the evaluation (heuristic) function of A\* search combines the

information in both the candidate node and the previous searching track to determine the next graph node to be added into the searching path. This instrument makes A\* search competent for searching the optimal path in graph.

By reviewing the three graph representations of our problem setup, we know:

- (1) In the first representation (Fig 55), the solution is represented as a path. Clearly, with the instrument to consider a whole path altogether, A\* search excels greedy search.
- (2) In the second representation (Fig 56), solutions are represented by “sink” nodes. Search is arranged to always start from some “source” nodes, follow direction of edges to propagate to a “sink” node. The graph is subtly arranged in a way that the previous search track is recorded in the current graph node. By moving the information of historical searching track into graph node, “Markov” property is assigned to the graph representation itself, and this makes the mechanism of A\* search to consider the historical searching track unnecessary. Therefore this representation is indifferent to the choice of greedy search or A\* search, and both approaches “degenerate” to a hill climbing policy.
- (3) In the third representation (Fig 57), each node represents a possible solution, which is independent from other solutions. This representation essentially describes the solution space, and thereby search on this representation becomes comparing among different solutions. The comparison is naturally pair-wise (Markov). This makes greedy search and A\* search identical to each other, and both of them “degenerate” to local greedy search (hill climbing).

In our problem setup, to use greedy search in the first graph representation, the solution is built by sequentially adding new graph nodes to the searching path. Since the search path is ignored (lost) while selecting a new graph node, there is no way to check the constraints defined on Input List 2. With these checks missed, the built solution could be totally wrong. As a conclusion, greedy search is not suitable for searching the solution in the first graph representation.

To use greedy search in the second and the third graph representations, there is no principal limitation on checking constraints defined on Input List 2, due to the subtle arrangement of the search space structure. However, greedy search (hill climbing) is notorious for trapping to local minimum or local maximum, which makes it unlikely to find the optimal solution (global minimum or global maximum). This especially happens when the searching graph (space) is big.

To use A\* search in the first graph representation, its heuristic design allows it to consider the search path while selecting new graph nodes. This makes it possible to check constraints defined on Input List 2 in the problem setup. But, A\* search is still an approach to sequentially build a solution. Therefore, before the complete solution is found, in each searching step the searching path built so far only represents a part of the solution. Hence, only a subset of constraints defined on Input List 2 is checked. With some constraints unchecked, it is impossible to accurately estimate the distance of the currently built partial solution to the final solution (goal), and it could happen that the currently built partial solution is totally wrong. In computer terms, we say that we cannot build an admissible heuristics (Russell, et al., 2002) for A\* search for our problem setup. Without this warranty, it is possible that A\* search converges to a solution far away from the optimal solution.

Compared to the best first search, human logic based approach works on the second graph representation. To get over the problem of trapping in the local minimum or the local maximum, the human logic based approach extends the standard greedy search approach by adding two additional mechanisms into it.

- (1) It adds a mechanism to jump from the current graph node to another suitable graph node. Here we emphasize that **the jump is not arbitrary, and it is guided by shrinking the searching path represented by the graph node. This distinguishes it from a random walk policy.** Note, in the second representation, each node represents a searching path.
- (2) A previously visited graph node has the chance to be visited again, while this chance get smaller each times the graph node is revisited. **This distinguishes it from random walk policy, when the probability of visiting a graph node is fixed.**

We believe that with these two additional mechanisms, the greedy search does mimic human's logic in solving the problem setup defined in Fig 46 and Fig 56. In addition, the experimental result in NMR structural verification problem has demonstrated the utility of the human logic based optimization approach.

### 7.2.2.3 Difference between Stochastic Optimization and Human Logic Based Optimization

#### Stochastic Optimization Type I

Stochastic optimization is a cluster of heuristic searching algorithms which works in space with Markov property (Russell, et al., 2002) i.e. the second and the third graph representations of our problem setup. Stochastic optimization is designed to get over the problem of greedy search, where the local minimum or the local maximum is often found as the solution instead of the global minimum or the global maximum. To address the problem, stochastic optimization introduces a random walk mechanism into greedy search to "escape" from the local minimum or the local maximum. Specifically, its workflow is described as following:

- (1) Randomly select a "source" node in the graph as the initial current status node.
- (2) From the current status node, hill climbing (best first search heuristic) is used to select the "best" neighboring nodes as the new current status node.
- (3) Repeat (2) until the current status node reaches a local maximum or minimum. Here, by hill climbing, the search cannot find the next "best" status node (neighboring node), and therefore it is "trapped" in the local maximum or minimum. Then, random walk mechanism starts to pick up the next status node by "flipping a coin".
- (4) Go to (2) to continue the search.

Above mentioned is the principal of classical stochastic optimization approaches e.g. well-known simulation annealing (Russell, et al., 2002). Clearly, with this random walk mechanism, search is

possible to hurdle local maximums or minimums, and converge to the global maximum or minimum (optimal solution).

There are several mutations of the above optimization strategy. For example, instead of starting the random component at the time that greedy search traps in a local minimum or maximum, a stochastic component could be combined and used together with “high climbing” component in every search step. Specifically, instead of selecting the “best” neighboring node in above step (2), the algorithm could select several “best” neighboring nodes, and then let the stochastic component select a node from this set of “best” neighboring nodes as the next current status node, with probability proportional to their fitness to be “best” neighboring node. Formally, this mutation is named stochastic hill climbing (Russell, et al., 2002). Another example is that we can start several (k) greedy searches simultaneously from different starting points. Then, in each searching step, a set of “best” neighboring nodes are generated for each greedy search. All nodes that belong to these sets are added together into a single “best” node set. Next, the stochastic component selects k new current status from the set, with the probability proportional to each node’s fitness to be “best” neighboring node. This forms the principal of the group of optimization approaches e.g. stochastic beam search (Russell, et al., 2002), genetic programming (Russell, et al., 2002), etc.

## Stochastic Optimization Type II

In addition to the optimization approaches mentioned above, there is another type of stochastic optimization methodology, which is totally based on random walk strategy. It is particularly suitable to work in the third graph representation, in which the solution is selected from the solution space. Specifically, the work flow of the approach is:

- (1) The walk starts randomly from a node of the graph.
- (2) A stochastic component is used to select a neighboring node to walk with the probabilities proportional to the fitness of neighboring nodes.
- (3) Repeat (2) for a number of iteration. Then the statistics of number of times each node of the graph are visited are recorded. The distributions of the statistics are then used for determining which graph node is the optimal solution.

This is the principal of a cluster of optimizations named Markov Chain Monte Carlo approaches.

To summarize the group of the first type stochastic optimizations, all of them use a random component to supplement the greedy search strategy. Obviously, the Markov property of the first type stochastic optimization limits its applicability on the first graph representation. To the second graph representation, the first type stochastic optimization also shows its inability. Specifically, in the second graph representation, the graph is organized as a directed graph with multiple paths from “source” nodes to “sink” nodes. It could happen that some of these paths never intersect. Lacking connectivity, neither random component nor hill climbing component can guide the search to “jump” from the current path to any other paths. Hence the first type stochastic optimization is lack of ability to reach the whole search space of the second graph representation, and so as to lack of ability to build the optimal solution in this graph representation. To the third graph representation,

the first type stochastic optimization is perfectly matched, since there is no limitation on its Markov property, and there is no problem of ransacking the search space, either. In fact, the stochastic optimization is designed to work in the solution space e.g. the third graph representation. However, there is the problem of utilizing the first type stochastic optimization in our problem setup. Specifically, the efficiency of the first type stochastic optimization relies on the selection of the starting point, and the effect is especially significant while the search space is big. If it happens that the starting point is close to the optimal solution, the optimization often converges to the real (optimal) solution quickly. Otherwise, the algorithms could become extremely slow. Note, this low efficiency mainly originates from the uncertainty of the stochastic component. Unfortunately, to our problem setup in Fig 54, the good searching starting point is difficult to be acquired. This is due to the constraints defined on Input List 2. Without considering these additional constraints, any best first search approaches would fail to create a good starting point. To consider these constraints, the best optimization approach to build a solution is the human logic based approach (see 7.2.2.2). In fact the human based logic can give a reasonable solution, which makes the further optimization in the solution space unnecessary. This gives us an argument why we believe that utilizing human based logic to directly build the optimal solution is the better policy than the policy of searching the optimal solution in the solution space.

To summarize the group of the second type stochastic optimization, it relies on Monte Carlo sampling. In principal, Monte Carlo sampling requires high amount of instances (ideally infinite instances) to give the reliable estimation. This makes the optimization based on Monte Carlo sampling extremely slow. As the conclusion, we don't believe that the second type optimization approach is a practical choice for the problem setup.

Compared to stochastic optimization, there is no random component in the human logic based optimization. Instead a human-mimicking reconsideration mechanism is adopted to avoid trapping in the local minimum or maximum. We believe that this reconsideration (controlled "jump" in the second graph representation plus reduced chance to revisit the previously visited nodes) mechanism is a better choice than the blind random selection, and thereby makes the human logic based optimization more efficient than the stochastic optimization under our problem setup (see Fig 46 and Fig 54)). Theoretically, by excluding the random component, the human logic based optimization could be unified under the maximal likelihood principal (see Chapter 5).

### 7.2.3 Summary of Human Logic Based Optimization

In 7.2.2, we introduced some classical optimization methods and explained the problem of applying them to our problem setup. Combined with 7.2.1, we also explained the advantage of utilizing the human logic based approach to approach this problem setup. Specifically, we explained the difference between the human logic based approach and those traditional optimization approaches, and show that the human logic based approach works on the second graph representation, and it contains the mechanism to jump under control among nodes in the graph, and has the flexibility to revisit the previous visited graph nodes. Practical experience shows that this approach solves the



problem of NMR structure verification very well, while other classical optimizations have been shown useless. We hope that our approach to NMR structure verification problem and the resultant human logic based optimization could be helpful for other practical participants who also face the scenario to design the automatic system to replace human beings. Especially, we hope that other participants could add our human logic based optimization into their optimization tool box, and apply it to other similar problems from different application domains.

## **Chapter 8 Limitation**

In this chapter we explain the limitation of our NMR structure verification system, and explain where we need to improve it.

### **8.1 Limitation in Technology**

In this section, we explain the limitation of our current system and methodologies.

#### **8.1.1 Problems of Isomere, Conformer, and Hetero Coupling**

1D <sup>1</sup>H NMR spectrum is a powerful tool for molecular structure verification, which is capable to be used independently to identify the structure of most compounds (>99%) through sufficient interpretation. The automatic system we built inherits majority interpretation skills from NMR spectroscopists e.g. chemical shift analysis, proton number analysis, J-coupling analysis, etc. But there are still some skills of spectroscopists missed in the system. Most important among them is the skill to identify Isomeres, Conformers, or hetero couplings from the molecular structure, and utilize this additional information to improve the quality of the structural verification procedure. Isomeres, Conformers, or hetero couplings are terms used by spectroscopists. Intuitively, they represent the scenario, where a molecule has a unique 2D structure, but has two (multiple) 3D structures. These different 3D constructions of the molecule will produce different NMR signals, and thereby experimentally observable 1D <sup>1</sup>H NMR spectrum is actually the mix of all of these NMR signals. Only with 2D molecular structure as the input of the system, the subtle difference of 3D constructions of the molecule is invisible by the system, and thereafter this additional complexity in the 1D <sup>1</sup>H NMR spectrum is unexpected by the system. As the result, Isomeres, Conformers, or hetero couplings will cause the system to wrongly convert the structural verification decision from consistency to inconsistency, and thereby it deteriorates the accuracy of the system.

Through discussing with NMR spectroscopists, we understand that experienced NMR spectroscopists have the ability to identify Isomeres, Conformers, or hetero couplings from some molecules by only looking at their 2D structures and to precisely predict the corresponding changes in the 1D <sup>1</sup>H NMR spectrum. We believe that by computerizing this human expertise and adding them into the system we can keep improving the accuracy of the system.

### 8.1.2 Keeping Improving Assignment Accuracy

To push the automatic molecular structure NMR spectrum verification system into practice, we need to keep convincing NMR spectroscopists with the assignment consistency between the system and NMR spectroscopists. As we discussed in 7.3.3, our evaluation of assignment accuracy of the system on the limited dataset is about 84%, and there is a margin of 16% to be improved. In addition, we expect that under practical application environment, the assignment accuracy would decrease. These facts motivate us to keep improving the system's assignments accuracy. As we explained in 7.3.3, the majority of inconsistencies between the system and NMR spectroscopists on the test dataset comes from lack of pattern recognition and pattern matching ability of the system on overlapping first-order multiplets and on high-order multiplet. Clearly, to increase the assignment accuracy, a first improvement we could implement is to do some first-order multiplet recognition upon the overlapping NMR signals, and use them for assignments. This will reduce the assignment errors originated from the overlapping of first-order multiplets. Second, some NMR signal shape pattern recognition techniques could be added into the system to identify high-order multiplets from the spectrum, and thereupon to reduce the assignment error originated from the magnetic inequivalence of protons. Obviously, to improve the assignment accuracy, the new pattern recognition and pattern matching techniques mentioned above should be added into the system.

### 8.1.3 Adding 2D <sup>1</sup>H NMR and 1D <sup>13</sup>C NMR Interpretation

1D <sup>1</sup>H NMR spectrum technique is the main work horse for molecular structure verification. However, no technique is "omnipresent". There exist some molecules which cannot be identified by the 1D <sup>1</sup>H NMR spectrum alone even by top experts in the NMR structural verification field. If this incidentally happens, NMR spectroscopists turn to rely on other NMR techniques such as 2D <sup>1</sup>H NMR spectrum and/or <sup>13</sup>C NMR spectrum to supplement the <sup>1</sup>H 1D NMR structural verification process. Obviously, to automate 2D <sup>1</sup>H NMR and <sup>13</sup>C NMR structure verification will improve the accuracy of the system, and further push the system into the real industrial application beyond compound library management.

### 8.1.4 Combining the Structure Verification of NMR Spectrum with Mass Spectrum

More information means more accuracy. Even though mass spectrum is simpler in principal and gives only limited information for structural verification, it is based on a totally different principle. Absorbing the ability of mass spectrum structure verification into the system could also remedy the system's limitation in certain environments. Depending on the potential application, this merge is going to improve the performance of the automatic verification system, and make the system more reliable to face possible new challenges emerging from small molecular structure verification tasks.

## 8.2 Limitation of the Experiment

In this subsection, we explain the limitation of our experimental methods.

### 8.2.1 Limited Representativeness of Simulated Dataset

Since we have a limited quantity of real compounds, we used simulated data (spectrum and theoretical multiplet distribution list pairs) to evaluate the accuracy of the system. Here the evaluation of the accuracy means the estimations of two types of errors. However, even though the simulated spectra are specially designed to simulate all possible scenarios, which can happen in the 1D  $^1\text{H}$  NMR spectrum including existence of high order multiplets, existence of overlapping of first order multiplets, existence of impurities, shape and position variance of solvent signal, variance of baseline, and NMR spectroscopists have been involved to control the quality of the simulated spectra, it is still possible to doubt the estimated values of two types of errors. To reliably estimate the two types of error so as to the accuracy of the system, we need to test the system against millions of real compounds. Only compound libraries of pharmaceutical companies have the size of millions of compounds. But they are inventories of the pharmaceutical companies, which are unavailable to the public. In addition, even if we have access to use these compound libraries, intentionally acquiring NMR spectra of all these compounds is a huge amount of human work. Therefore, the optimal policy is to merge the estimation of the system's accuracy into the routine structural verification tasks of compound library management. From this point of view, we need cooperation from some pharmaceutical companies. At the moment, our industrial cooperator – a NMR manufacturer starts negotiating with some pharmaceutical companies. From their feedback, we understand that pharmaceutical companies are interested in our work and are eager to test the system in the practical application environment of compound library management.

### 8.2.2 Limited Representativeness of Real Compound Dataset

The real compound dataset is used to evaluate both accuracy of the system and assignment consistency between the system and NMR spectroscopists. Obviously, the quality of the estimation relies on the representativeness of the dataset. From the practical application point of view, the real compound dataset we used is a little bit utopian. Specifically, there are no examples of Isomere, Conformer or the hetero coupling in the dataset. There are no examples of compounds which cannot be identified by applying  $^1\text{H}$  1D NMR interpretation alone, either. We know that the probabilities of the above two scenarios happening in the practical structure verification tasks are low, but nevertheless existence of these compounds would deteriorate the performance of the system. Without the dataset, which could equably represent these two scenarios and so other

scenarios we have not yet expected, the estimation of the system's performance could not represent the system's behavior in the practical application environment. To pursue the better estimation of the system's performance, we return to the solution we proposed for taking over the problem of limited representativeness of simulated datasets (see 8.2.1). While the new deal is settled with the pharmaceutical company to allow us to access its compound library, we are able to evaluate the system's performance in the practical application environment. In addition, we are also able to measure the damnification of the scenarios e.g. Isomere, Conformere, or the hetero coupling on the system performance, and accordingly design new mechanisms to deal with it if necessary.

## **8.3 Limitation in Industrialization**

In this subsection, we explain the limitation of applying the current system to the practical application environments within the pharmaceutical industry.

### **8.3.1 NMR Automation Hardware**

To implement a practical automatic structure verification solution, solving the problem of automatic structure verification is only a part of the whole solution. To realize the automatic structure verification in practice, it requires additional NMR hardware e.g. physical sample buffer and automatic sample feeding robot arms. In addition, the system needs to be seamlessly embedded into NMR spectrometer control software so that the automatic structure verification is integrated into the spectrum acquisition process to give the consistency analysis on time. Obviously, all these require NMR manufacturers to invest on both developing the automation hardware and reengineering the NMR spectrometer control software. In fact, synchronous to our project, another project is executed in our cooperating NMR manufacturing site to develop an automatic sample feeding mechanisms and embed it into NMR spectrometers. At the moment, a software team from the NMR manufacture is designing the interface between our system and the NMR spectrometer control software. With above projects finished, as an independent system, NMR spectrometer is ready to accept fully automatic molecular structural verification tasks.

### **8.3.2 Link to Compound Library Management Automation**

To push the NMR based automatic structural verification solution to routine QA/QC in compound library management of the pharmaceutical industry, pharmaceutical companies need to make an effort to link its automatic compound sample management system to the automatic sample feeder of the NMR spectrometer. By adding this part, the automatic NMR structure verification system becomes fully interactive with the automatic compound library management system. At this step,

we reach to the milestone to practically test our structural verification system in the real application environment. Feedbacks from the evaluation will show us the direction of where and how to improve the system. We believe only under this track, NMR based automatic structural verification can truly become mature.

## **Chapter 9 Future Work**

In this chapter, we summarize the future work of both extending the system to a commercial product and pushing it to the practical application in compound library management, and potentially applying the methodologies we developed for the system to other applications in different fields.

### **9.1 Future Work in NMR/Pharmaceutical Industry**

As we discussed in the limitations (see Chapter 8), to pursue our final goal, we need to keep improving the system's performance. Specifically,

- (1) we need to extend the system's ability to detect and verify the molecular structures which contain Isomere, Conformer, hetero coupling or other unexpected characteristics.
- (2) we need to continually improve the system's assignment consistency to human spectroscopists. This could be improved by adding the mechanism of first-order multiplet analysis of overlapping NMR signals, the mechanism of identifying high-order multiplet from the spectrum, etc. into the system.
- (3) to build a complete structure verification system, we need to supplement the automatic 1D <sup>1</sup>H NMR structural verification system with both automatic 2D <sup>1</sup>H NMR structural verification procedure and <sup>13</sup>C NMR structure verification procedure.
- (4) Following this track, we could also combine NMR based structural verification procedure to that of mass spectrum based structural verification. Note, these supplements or combinations mentioned in (3) and (4) would have limited contributions to improving the accuracy of the system. Therefore, the decision of whether to implement these additions relies on the actual requirements for the accuracy of the applications.
- (5) we need to push the development of automation hardware in both NMR manufacturing and pharmaceutical companies. Without this hardware as mediums, automatic verification software itself can not accomplish the automatic structural verification task.
- (6) we need to push the evaluation of the system in the practical application environments. The system is complex enough so that there are errors, defects we have never expected. Abundant tests in practical application environments give us an opportunity to discover these errors and defects, so as to allow us to keep improving the system's performance. It is important because **only by passing these practical tests, we can conclude that the system is mature to be used in practice.**

## 9.2 Future Work in Applied Computer Science

In the majority of practical application fields, humans still show their superiority to the computer. To keep improving the efficiency of industrial production, there are increasing requirements in various industrial fields to utilize computer technology to replace humans in order to reduce costs and increase productivity. Our NMR structure verification problem only shows an example of these requirements. We believe that the human logic based approach we summarized from this particular problem could be easily transplanted to problems in other domains, where the computer technology is motivated to replace human beings e.g. automatic signal analysis from radar or sonar. To validate this hypothesis, we are eager to seek another problem from a different industrial field to apply our optimization approach. Further, relying on the experience of applying our approach in a second application field, we could start to distill the common features among the diversely subtle differences of the implementations for two application fields. These common features will supply a massive backbone for us to formalize our human logic based optimization algorithm. As the final goal, we would like to supply a mature new optimization algorithm to the applied computer science community.



## **Chapter 10 Conclusion**

Technical breakthroughs in NMR spectrometer (especially in NMR probe) over the past 30 years make it possible to directly apply NMR technology in QA/QC of compound library management in the pharmaceutical industry. Manual NMR spectrum interpretation becomes the only technical obstacle to prevent using NMR instead of LC-Mass for structural verification tasks in compound library management. This practical requirement motivated several attempts to automate molecular structure NMR spectrum verification. Unfortunately, these attempts are denied by the inspection of the practical application environment. As a result, NMR still sits as the arbitral method to supplement LC-Mass based automatic molecular structural verification process in practical application of compound library management.

To peek the rationales of these automatic NMR structural verification systems, they all use optimization methods to search reasonable assignments between the molecular structure and its 1D  $^1\text{H}$  NMR spectrum. However, the principles of these optimization procedures are widely divergent from the human spectroscopists' logic to do assignments. We believe this is the reason why these systems fail in the practical test.

Alternatively, in the scope of this thesis, we propose to design and implement a new molecular structure NMR spectrum verification system, which mimics human spectroscopists' logic of structure verification analysis. With three years efforts from both NMR spectroscopists and computer scientists, the system was built and demonstrated to behave similar to human logic in the structural verification task. Evaluated with both a real compound dataset and some simulated datasets, the system shows both high consistency analysis decision accuracy and high consistency to human spectroscopists in detail assignments. As the results, NMR spectroscopists involved in the project are convinced that the system shows better accuracy than previous structural verification solutions, and has a potential to reach structural verification decision accuracy of human spectroscopists. More importantly, the assignment report generated by the system gives NMR spectroscopists an opportunity to check the structural verification result of the system with their chemical knowledge. It is the first times that the structural verification system starts to "speak a common language" with NMR spectroscopists. The experimental result of high consistency between the assignments of the system and that of NMR spectroscopists deeply "touch" the spectroscopists involved in the project, and in fact builds their confidence in the system. To foresee the commercial merit of the system, our cooperator – the top NMR manufacturer has applied two patents to protect the core technology of the system, and has applied their effort to commercialize the system. Through their business channel, several pharmaceutical companies have shown their will to evaluate the system.

As we explained in the limitation (see Chapter 8), the evaluation in the pharmaceutical industry gives us the opportunity to test our system in the practical application environment. It is well known that only practice can validate the effectiveness of a theory. Therefore, before the system passes the evaluation under the practical application environment of the pharmaceutical industry, no one can predict the utility of our idea, methodology and system in practice. Nonetheless, we believe that our

human logic mimicking strategy supplies an alternative path to domain participants to approach the structural verification problem. With the feedback from the evaluation in the practical application environment, we have chance to continue improving the system until we reach our final goal – using automatic NMR structural verification system as the footstone of structural verification tasks in compound library management.

Overall, we hope that the approach we used to do the optimization could give applied computer science participants some hints to solve problems with similar characteristics in various domains. Specifically, for the problem of matching two sets of elements with additional constraints defined in one of them, the human logic based optimization (heuristic search) could give more flexibility and efficiency compared with other classical optimization approaches. We are particularly happy to see that the similar human logic based optimization is being applied for solving other practical problems in the near future.

## **Part IV**

## **Appendix**



## A. Glossary

### 1

**1D:** 1 dimension ,

**<sup>1</sup>H NMR:** NMR spectrum generated by measuring NMR signal of protons in compound.

**<sup>13</sup>C NMR:** NMR spectrum generated by measuring NMR signal of the isotope of carbon in compound.

### 2

**2D:** 2 dimension,

### 3

**3D:** 3 dimension,

## C

**chemical bond:** is the physical process responsible for the attractive interactions between atoms and molecules. See [http://en.wikipedia.org/wiki/Chemical\\_bond](http://en.wikipedia.org/wiki/Chemical_bond) for more information.

**chemical shift:** In nuclear magnetic resonance (NMR), the chemical shift describes the dependence of nuclear magnetic energy levels on the electronic environment in a molecule. The unit is ppm (parts per million) referring to the difference of the resonance frequency (in Hertz (Hz)) of a certain nucleus to a reference frequency (Hz). The chemical shifts in a <sup>1</sup>H Spectrum are typically in the range of +12 to -4 ppm,

**chemically equivalent functional group:** All protons of a molecule with the exact same chemical environment, eg, a Methyl group (CH<sub>3</sub>),

**chromophore:** a chromophore is part (or moiety) of a molecule responsible for its color.

**compound library management:** is one such field that attempts to manage and upkeep compound libraries as well as maximizing safety and effectiveness in their management.

**computational complexity:** The computational complexity of a problem is the number of steps that it takes to solve an instance of the problem as a function of the size of the input. It is roughly divided as linear, polynomial and exponential complexity,

**coupling connectivity:** The pair of protons or pair chemically equivalent proton groups interact with each other through the chemical bonds of a molecule and result in the splitting of the NMR signal,

**coupling constant:** The size of the splitting which occurs in a multiplet (difference in frequency measured in Hz between peaks), a typical coupling constant value is 7 Hz. In Fig6 a multiplet is shown. The distance in Hz between e.g. the most left peak and its direct neighbour is a coupling constant ,

## D

**divide and conquer algorithms:** Divide and conquer is an important algorithm design paradigm. It works by recursively breaking down a problem into two or more sub-problems of the same (or related) type, until these become simple enough to be solved directly. The solution to the sub-problems are then combined to give a solution to the original problem.,

**Dimethyl Sulfoxide (DMSO):** A solvent often used to store organic compounds of compound libraries in the liquid phase.

**dark region:** Any extraneous peaks from an 1D <sup>1</sup>H NMR Spectrum, which do not overlap significantly with signal peaks of the Molecule,

**deuterated solvents:** means the family of solvents in which the hydrogen atoms ("H") are replaced with deuterium (heavy hydrogen) isotope ("D").

## F

**false positive:** Plainly speaking, it occurs when we are observing a difference when in truth there is none. An example of this would be if a test shows that a woman is pregnant when in reality she is not.

**functional group:** in organic chemistry, functional groups are specific groups of atoms within molecules that are responsible for the characteristic chemical reactions of those molecules.

## G

**G protein-coupled receptor (GPCR):** is a large protein family of transmembrane receptor that senses molecules outside the cell and activate inside signal transduction pathways and ultimately cellular responses. Detail see [http://en.wikipedia.org/wiki/G\\_protein-coupled\\_receptor](http://en.wikipedia.org/wiki/G_protein-coupled_receptor).

**graph traveling algorithm:** It denotes algorithms, which could explore all graph nodes. Typical graph traveling algorithms include deep-first search, breath-first search, etc,

**greedy search:** It is a searching metaheuristic of making the locally optimum searching choice at each stage with the hope of finding the global optimum,

## H

**liquid chromatography:** High-performance liquid chromatography (or High pressure liquid chromatography, HPLC) is a form of column chromatography used frequently in biochemistry and analytical chemistry to separate, identify, and quantify compounds. HPLC utilizes a column that holds chromatographic packing material (stationary phase), a pump that moves the mobile phase(s) through the column, and a detector that shows the retention times of the molecules. Retention time varies depending on the interactions between the stationary phase, the molecules being analyzed, and the solvent(s) used.

**HPLC-MS:** Liquid chromatography-mass spectrometry (LC-MS, or alternatively HPLC-MS) is an analytical chemistry technique that combines the physical separation capabilities of liquid chromatography (or HPLC) with the mass analysis capabilities of mass spectrometry. LC-MS is a powerful technique used for many applications which has very high sensitivity and specificity. Generally its application is oriented towards the specific detection and potential identification of chemicals in the presence of other chemicals (impurities).

**High Throughput Screening (HTS):** is a method for scientific experimentation especially used in drug discovery. Specifically, using robotics, data processing, control software, liquid handling devices, and sensitive detectors, HTS quickly conducts millions of biochemical, genetic or pharmacological tests. Through the process, one can rapidly identify active compounds, antibodies or genes which modulate a particular biomolecular pathway. The results of these experiments provide starting points for drug design and for understanding the interaction or role of a particular biochemical process in biology.

**multiplet hypothesis's total amplitude:** Sum of amplitudes of all peaks belonging to the given multiplet hypothesis,

## I

**IR:** in the thesis, IR means Infrared spectroscopy (IR spectroscopy), which is the subset of spectroscopy that deals with the infrared region of the electromagnetic spectrum. It can be used to identify compounds or investigate sample composition.

## M

**Mass:** In the thesis, Mass means Mass spectrometry (MS), which is an analytical technique for the determination of the elemental composition of a sample or molecule. The MS principle consists of ionizing chemical compounds to generate charged molecules or molecule fragments and measurement of their mass-to-charge ratios.

**molar:** a unit of concentration, or molarity, of solutions equal to 1 mol/L

**Monte Carlo methods:** are a class of computational algorithms that rely on repeated random sampling to compute their results. Because of their reliance on repeated computation and random or pseudo-random numbers, Monte Carlo methods are most suited for calculation by a computer. Monte Carlo methods tend to be used when it is unfeasible or impossible to compute an exact result with a deterministic algorithm.

**multiplet:** The ensemble of all signals from a chemically equivalent functional group in a 1D <sup>1</sup>H NMR spectrum is called a multiplet. E.g. if the sum of all signals of a chemically equivalent functional group is two the multiplet would be called doublet, three a triplet etc.

**multiplet hypotheses space:** The ensemble of all possible experimental multiplets extracted from an 1D <sup>1</sup>H NMR Spectrum,

**multiplicity:** see number of couplings,

## N

**NMR:** Nuclear Magnetic Resonance,

**NP hard:** nondeterministic polynomial-time hard. In computational complexity theory, it denotes a group of problems which can not be solved in polynomial time,

**HSQC NMR:** 2D HSQC (Heteronuclear Single Quantum Coherence) experiment correlates chemical shifts of directly bound nuclei (i.e. two types of chemical nuclei). For example <sup>1</sup>H, <sup>15</sup>N-HSQC correlates chemical shifts within NH groups.

**non-deuterated DMSO:** DMSO (Dimethyl Sulfoxide) in which deuterium (heavy hydrogen) isotope ("D") are replaced with hydrogen atoms ("H"). In practical application, non-deuterated DMSO is cheaper than DMSO.

**nuclear spin:** It is an intrinsic quantum mechanical property of an atomic nucleus,

**number of couplings:** The number of protons interacting with the target proton through the chemical bonds of a molecule and results in the splitting of NMR signal,

## P

**peak clusters:** A peak cluster denotes an ensemble of positional symmetric peaks from an 1H 1D NMR Spectrum,

**protein kinase:** is a kinase enzyme that modifies other proteins by chemically adding phosphate groups to them (phosphorylation). Phosphorylation usually results in a functional change of the target protein by changing enzyme activity, cellular location, or association with other proteins. Detail see [http://en.wikipedia.org/wiki/Protein\\_kinase](http://en.wikipedia.org/wiki/Protein_kinase)

## Q

**quality assurance (QA):** refers to planned and systematic production processes that provide confidence in a product's suitability for its intended purpose. It is a set of activities intended to ensure that products (goods and/or services) satisfy customer requirements in a systematic, reliable fashion.

**quality control (QC):** is the branch of engineering and manufacturing which deals with assurance and failure testing in design and production of products or services, to meet or exceed customer requirements.

**quantification:** It is a procedure to determine the molar concentration of the main substance of a liquid state NMR sample, whereupon the solvent and impurities that are connected to the solvent are not considered as main substance,

## R

**R-group:** In a chemical structural formula, a generic substituent can be written as R. This is a generic placeholder which may replace any portion of the formula as the author finds convenient. Here a substituent means an atom or group of atoms substituted in place of a hydrogen atom on the parent chain of a hydrocarbon in organic chemistry and biochemistry.

## S

**satellite peaks:** They are signal peaks in a 1D  $^1\text{H}$  NMR spectrum created by direct bonding between protons and nuclear spin 1/2 particles e.g.  $^{13}\text{C}$ ,  $^{15}\text{N}$ , etc ,

**small molecule:** A small molecule is an organic compound which is not a polymer. Biopolymer (e.g. nucleic acids, proteins) often have much higher molecular weight than small molecules, but not necessarily. Small molecules are the main form of drugs. **structure verification:** It is a procedure to check if a given molecule structure is consistence with a given 1D  $^1\text{H}$  NMR spectrum,

## T

**theoretical multiplet distributions:** The theoretical multiplets with given chemical shift range, coupling constant ranges,

**theoretical multiplets:** The multiplets are interpreted from a given molecule. With NMR text book knowledge the appearance of each proton of the molecule in the 1D  $^1\text{H}$  NMR spectrum as a multiplet is estimated. ,

## X

**X-ray:** in the thesis, x-ray means X-ray crystallography, which is a method of determining the arrangement of atoms within a crystal, in which a beam of X-rays strikes a crystal and diffracts into many specific directions. From the angles and intensities of these diffracted beams, a crystallographer can produce a three-dimensional picture of the density of electrons within the crystal. By crystallizing compounds, it could be used to determine the three dimensional structure of the compounds.



## B. References

**ABRAHAM R J** A MODEL FOR THE CALCULATION OF PROTON CHEMICAL SHIFTS IN NON-CONJUGATED ORGANIC COMPOUNDS [Journal] // Progress in nuclear magnetic resonance spectroscopy. - 1999. - 2 : Vol. 35. - pp. 85-152.

**ACD** ACD/HNMR Predictor [Online] // Advanced Chemistry Development / prod. Predictor ACD/HNMR. - Advanced Chemistry Development Lab, 1996 - 2009. - [http://www.acdlabs.com/products/spec\\_lab/predict\\_nmr/hnmr/](http://www.acdlabs.com/products/spec_lab/predict_nmr/hnmr/).

**ACD** ACD/NMR Processor [Online] // Advanced Chemistry Development . - Advanced Chemistry Development Lab, 2005. - [http://www.acdlabs.com/products/spec\\_lab/exp\\_spectra/nmr\\_proc/](http://www.acdlabs.com/products/spec_lab/exp_spectra/nmr_proc/).

**Bailing Liu, Songjun Li and Jie Hu** Technological Advances in High-Throughput Screening [Journal] // American Journal of Pharmacogenomics. - 2004. - Vol. 4. - pp. 263-276.

**Barfield Michael and Karplus M** Valence-bond bond-order formulation for contact nuclear spin-spin coupling [Journal] // Journal of American Chemical Society. - January 1969. - 1 : Vol. 91. - pp. 1–10.

**Barnett Octo G [et al.]** DXplain An Evolving Diagnostic Decision-Support System [Journal] // The Journal of the American Medical Association. - 1987. - 1 : Vol. 258. - pp. 67-74.

**Barnett Octo** Medical Computing Lecture Note [Online] // Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT. - 2004. - <http://groups.csail.mit.edu/medg/courses/6872/2004/DXp%20HST%20Lec%2005.pdf>.

**Berger James O** Statistical Decision Theory and Bayesian Analysis (Springer Series in Statistics) [Book]. - [s.l.] : Springer , 1993. - ISBN-10 / ASIN: 0387960988 , ISBN-13 / EAN: 9780387960982.

**Berner Eta S** Clinical Decision Support Systems: Theory and Practice [Book]. - [s.l.] : Springer, 1998. - ISBN 0387985751, 9780387985756.

**Bravata Dena M [et al.]** Evaluating Detection and Diagnostic Decision Support Systems for Bioterrorism Response [Journal] // Emerg Infect Dis. - 2004. - 1 : Vol. 10.

**Bremser W** Hose — a novel substructure code [Journal] // Analytica Chimica Acta. - December 1978. - 4 : Vol. 103. - pp. 355-365.

**Burbaum Jonathan J and Sigal Nolan H** New technologies for high-throughput screening [Journal] // Current Opinion in Chemical Biology. - June 1997. - 1 : Vol. 1. - pp. 72-78.

**CambridgeSoft** ChemDraw [Online] // CambridgeSoft.com. - CambridgeSoft, 2009. - <http://www.cambridgesoft.com/software/ChemDraw/>.

**Castiglione F [et al.]** Toward a Generalized Algorithm for the Automated Analysis of Complex Anisotropic NMR Spectra [Journal] // Journal of Magnetic Resonance. - May 1998. - 1 : Vol. 132. - pp. 1-12.

**Chan James A and Hueso-Rodríguez Juan A** Compound library management. [Journal] // Methods in Molecular Biology. - 2002. - Vol. 190. - pp. 117-127.

**Committee on Innovations in Computing and Communications: Lessons from History National Research Council** Funding a Revolution: Government Support for Computing Research [Book]. - [s.l.] : The National Academies Press, 1999. - ISBN-10: 0-309-06278-0, ISBN-13: 978-0-309-06278-7.

**Corens David [et al.]** Liquid chromatography–mass spectrometry with chemiluminescent nitrogen detection for on-line quantitative analysis of compound collections: advantages and limitations [Journal] // Journal of Chromatography A. - November 2004. - 1-2 : Vol. 1056. - pp. 67-75.

**Duda Richard O, Hart Peter E and Stork David G** Pattern Classification [Book]. - [s.l.] : Wiley-Interscience, 2000. - 2 edition. - ISBN-10: 0471056693, ISBN-13: 978-0471056690.

DXplain [Online] // Wikipedia. - April 9, 2009. - <http://en.wikipedia.org/wiki/DXplain>.

**Efron Bradley and Tibshirani R J** An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability) [Book]. - [s.l.] : Chapman & Hall/CRC, 1994. - 1 edition. - ISBN-10: 0412042312, ISBN-13: 978-0412042317.

**Fang Liling [et al.]** Evaluation of Evaporative Light-Scattering Detector for Combinatorial Library Quantitation by Reversed Phase HPLC [Journal] // Journal of Combinatorial Chemistry. - 2000. - 3 : Vol. 2. - pp. 254–257.

**Feldman M J and Barnett G O** An approach to evaluating the accuracy of DXplain [Journal] // Computer Methods and Programs in Biomedicine. - August 1991. - 4 : Vol. 35. - pp. 261-266.

**Freund Yoav and Schapire Robert E** Experiments with a New Boosting Algorithm [Conference] // Machine Learning: Proceedings of the Thirteenth International Conference. - 1996. - pp. 148-156.

**Fürst Andràs and Pretsch Ernő** A computer program for the prediction of <sup>13</sup>C-NMR chemical shifts of organic compounds [Journal] // Analytica Chimica Acta. - 1990. - Vol. 229. - pp. 17-25.

**Fürst Andràs and Pretsch Ernő** Comprehensive parameter set for the prediction of the <sup>13</sup>C-NMR chemical shifts of sp<sup>3</sup>-hybridized carbon atoms in organic compounds [Journal] // Analytica Chimica Acta. - 1990. - Vol. 233. - pp. 213-222.

**GmbH Chemical Concepts** SpecInfo [Online] // Chemical Concepts / prod. SpecInfor. - Chemical Concepts GmbH, , 1998. - <http://specinfo.wiley-vch.de/ARCHIVE/>.

**Golotvin Sergey S [et al.]** Automated structure verification based on <sup>1</sup>H NMR prediction [Journal] // Magnetic Resonance in Chemistry. - February 2006. - 5 : Vol. 44. - pp. 524 - 538.

**Golotvin Sergey S [et al.]** Automated structure verification based on a combination of 1D <sup>1</sup>H NMR and 2D <sup>1</sup>H<sup>13</sup>C HSQC spectra [Journal] // Magnetic Resonance in Chemistry. - August 2007. - 10 : Vol. 45. - pp. 803 - 813.

**Golotvin Sergey, Vodopianov Eugene and Williams Antony** A new approach to automated first-order multiplet analysis [Journal] // Magnetic Resonance in Chemistry. - March 2002. - 5 : Vol. 40. - pp. 331 - 336.

**Griffiths Lee and Bright Jonathan D** Towards the automatic analysis of  $^1\text{H}$  NMR spectra: Part 3. Confirmation of postulated chemical structure [Journal] // Magnetic Resonance in Chemistry. - August 2002. - 10 : Vol. 40. - pp. 623 - 634.

**Griffiths Lee and Horton Rob** Towards the automatic analysis of  $^1\text{H}$  NMR spectra: Part 4 - Additional requirements of flow-NMR [Journal] // Magnetic Resonance in Chemistry. - September 2004. - 12 : Vol. 42. - pp. 1012 - 1021.

**Griffiths Lee and Horton Rob** Towards the automatic analysis of NMR spectra: Part 6. Confirmation of chemical structure employing both  $^1\text{H}$  and  $^{13}\text{C}$  NMR spectra [Journal] // Magnetic Resonance in Chemistry. - December 2005. - 2 : Vol. 44. - pp. 139 - 145.

**Griffiths Lee** Towards the automatic analysis of  $^1\text{H}$  NMR spectra [Journal] // Magnetic Resonance in Chemistry. - June 2000. - 6 : Vol. 38. - pp. 444 - 451.

**Griffiths Lee** Towards the automatic analysis of  $^1\text{H}$  NMR spectra: Part 2. Accurate integrals and stoichiometry [Journal] // Magnetic Resonance in Chemistry. - March 2001. - 4 : Vol. 39. - pp. 194 - 202.

**Griffiths Lee** Towards the automatic analysis of  $^1\text{H}$  NMR spectra: Part 5. Confirmation of chemical structure with flow-NMR [Journal] // Magnetic Resonance in Chemistry. - December 2005. - 1 : Vol. 44. - pp. 54 - 58.

**Hamper Bruce C [et al.]** High-Throughput  $^1\text{H}$  NMR and HPLC Characterization of a 96-Member Substituted Methylene Malonic Acid Library [Journal] // Journal of Combinatorial Chemistry. - 1999. - 2 : Vol. 1. - pp. 140–150.

**Hann Mike M and Oprea Tudor I** Pursuing the leadlikeness concept in pharmaceutical research [Journal] // Current Opinion in Chemical Biology. - June 2004. - 3 : Vol. 8. - pp. 255-263.

**Hastie T, Tibshirani R and Friedman J H** The Elements of Statistical Learning [Book]. - [s.l.] : Springer, 2003. - Corrected edition. - ISBN-10: 0387952845, ISBN-13: 978-0387952840.

**Heller Stephen** SpecTool [Journal] // Journal of Chemical Information and Modeling. - July 1994. - 4 : Vol. 34. - p. 1026.

**Hoye Thomas R and Zhao Hongyu** A Method for Easily Determining Coupling Constant Values: An Addendum to "A Practical Guide to First-Order Multiplet Analysis in  $^1\text{H}$  NMR Spectroscopy" [Journal] // Journal of Organic Chemistry. - May 2002. - 12 : Vol. 67. - pp. 4014–4016.

**Jansma Ariane [et al.]** Automated Microflow NMR: Routine Analysis of Five-Microliter Samples [Journal] // Analytical Chemistry. - August 2005. - 19 : Vol. 77. - pp. 6509–6515.

**Kalelkar Sandeep [et al.]** Automated Analysis of Proton NMR Spectra from Combinatorial Rapid Parallel Synthesis Using Self-Organizing Maps. [Journal] // Journal of Combinatorial Chemistry. - 2002. - 6 : Vol. 4. - pp. 622–629.

**Karplus Martin** THEORY OF PROTON COUPLING CONSTANTS IN UNSATURATED MOLECULES [Journal] // Journal of American Chemical Society. - 1960. - 16 : Vol. 82. - pp. 4431–4432.

**Karplus Martin** Vicinal Proton Coupling in Nuclear Magnetic Resonance [Journal] // Journal of American Chemical Society. - 1963. - 18 : Vol. 85. - pp. 2870–2871.

**Keeler James** Understanding NMR Spectroscopy [Book]. - [s.l.] : Wiley, 2005. - ISBN-10: 0470017872, ISBN-13: 978-0470017876.

**Keifer Paul A [et al.]** Direct-Injection NMR (DI-NMR): A Flow NMR Technique for the Analysis of Combinatorial Chemistry Libraries [Journal] // Journal of Combinatorial Chemistry. - 2000. - 2 : Vol. 2. - pp. 151–171.

**Kurzweil Ray** The Singularity Is Near: When Humans Transcend Biology [Book]. - [s.l.] : Penguin Books, 2006. - ISBN-10: 0143037889, ISBN-13: 978-0143037880.

**London S** DXplain: a Web-based diagnostic decision support system for medical students. [Journal] // Medical Reference Services Quarterly. - 1998. - 2 : Vol. 17. - pp. 17-28.

**Macnaughtan Megan A [et al.]** High-Throughput Nuclear Magnetic Resonance Analysis Using a Multiple Coil Flow Probe [Journal] // Analytical Chemistry. - August 2003. - 19 : Vol. 75. - pp. 5116–5123.

**Mitchell Tom M** Machine Learning [Book]. - [s.l.] : McGraw-Hill Science/Engineering/Math, 1997. - 1 edition. - ISBN-10: 0070428077, ISBN-13: 978-0070428072.

**Pearl Judea** Causality: Models, Reasoning, and Inference [Book]. - [s.l.] : Cambridge University Press, 2000. - ISBN-10: 0521773628, ISBN-13: 978-0521773621.

**PERCH** Pearch NMR Software [Online] // PERCH Solutions. - PERCH Solutions Ltd, 2005. - <http://www.perchsolutions.com/products/products.html>.

**Pierens Gregory K [et al.]** Determination of Analyte Concentration Using the Residual Solvent Resonance in <sup>1</sup>H NMR Spectroscopy [Journal] // Journal of Natural Products. - April 2008. - 5 : Vol. 71. - pp. 810–813.

**Pinciroli Vittorio [et al.]** Characterization of Small Combinatorial Chemistry Libraries by <sup>1</sup>H NMR. Quantitation with a Convenient and Novel Internal Standard [Journal] // Journal of Combinatorial Chemistry. - June 2001. - 5 : Vol. 3. - pp. 434–440.

**Popa-Burke Ioana G [et al.]** Streamlined System for Purifying and Quantifying a Diverse Library of Compounds and the Effect of Compound Concentration Measurements on the Accurate Interpretation of Biological Assay Results [Journal] // Analytical Chemistry. - 2004. - 24 : Vol. 76. - pp. 7278–7287.

**Press William H [et al.]** Numerical Recipes in FORTRAN 77: The Art of Scientific Computing (v. 1) [Book]. - [s.l.] : Cambridge University Press, 1992. - ISBN-10: 052143064X, ISBN-13: 978-0521430647.

**Pretsch E, Bühlmann P and Affolter C** Structure Determination of Organic Compounds: Tables of Spectral Data [Book]. - [s.l.] : Springer, 2004. - ISBN-10: 3540678158, ISBN-13: 978-3540678151.

**Pretsch Ernő, Bühlmann Philippe and Badertscher Martin** Structure Determination of Organic Compounds: Tables of Spectral Data [Book]. - [s.l.] : Springer, 2009. - 4th. - ISBN-10: 3540938095.

**Pretsch Ernő, Fürst Andràs and Robien Wolfgang** Parameter set for the prediction of the <sup>13</sup>C-NMR chemical shifts of sp<sup>2</sup>- and sp-hybridized carbon atoms in organic compounds [Journal] // *Analytica Chimica Acta*. - August 1991. - 2 : Vol. 248. - pp. 415-428.

**Prost Élise, Bourg Stéphane and Nuzillard Jean-Marc** Automatic first-order multiplet analysis in liquid-state NMR [Journal] // *Comptes Rendus Chimie*. - March-April 2006. - 3-4 : Vol. 9. - pp. 498-502.

**Russell Stuart and Norvig Peter** Artificial Intelligence: A Modern Approach [Book]. - [s.l.] : Prentice Hall, 2002. - 2 edition. - ISBN-10: 0137903952, ISBN-13: 978-0137903955.

**Schaller R B and Pretsch E A** A computer program for the automatic estimation of <sup>1</sup>H NMR chemical shifts. [Journal] // *Analytica Chimica Acta*. - 1994. - Vol. 290. - p. 295-302.

**Schaller Renate Bürgin, Arnold Cédric and Pretsch Ernő** New parameters for predicting <sup>1</sup>H NMR chemical shifts of protons attached to carbon atoms [Journal] // *Analytica Chimica Acta*. - August 1995. - 1 : Vol. 312. - pp. 95-105.

**Schaller Renate Bürgin, Munk Morton E and Pretsch Ernő** Spectra Estimation for Computer-Aided Structure Determination [Journal] // *Journal of Chemical Information and Modeling*. - 1996. - 2 : Vol. 36. - pp. 239-243.

**Schaller Renate Bürgin, Munk Morton E and Pretsch Ernő** Spectra Estimation for Computer-Aided Structure Determination [Journal] // *Journal of Chemical Information and Modeling*. - 1996. - 2 : Vol. 36. - pp. 239-243.

**Schröder Harald, Neidig Peter and Rossé Gérard** High-Throughput Structure Verification of a Substituted 4-Phenylbenzopyran Library by Using 2D NMR Techniques [Journal] // *Angewandte Chemie International Edition*. - October 2000. - 21 : Vol. 39. - pp. 3816 - 3819.

**Sedgewick Robert** Algorithms in C, Part 5: Graph Algorithms [Book]. - [s.l.] : Addison-Wesley Professional, 2001. - 3rd Edition. - ISBN-10:410768684757, ISBN-13: 978-0-7686-8475-9.

**Sedgewick Robert** Algorithms in C, Parts 1-4: Fundamentals, Data Structures, Sorting, Searching [Book]. - [s.l.] : Addison-Wesley Professional, 1997. - 3rd Edition. - ISBN-10: 0-201-31452-5, ISBN-13: 978-0-201-31452-6.

**Sepetov Nikolai and Issakova Olga** Analytical characterization of synthetic organic libraries [Book Section] // *Combinatorial chemistry and technology* / book auth. Miertus Stanislav and Fassina Giorgio. - [s.l.] : Marcel Dekker, 1999.

**Solomons T. W Graham and Fryhle Craig B** Organic Chemistry [Book]. - [s.l.] : Wiley, 2003. - 8 edition. - ISBN-10: 0471417998, ISBN-13: 978-0471417996.

**Steinbeck Christoph, Krause Stefan and Kuhn Stefan** NMRShiftDB-Constructing a Free Chemical Information System with Open-Source Components [Journal] // Journal of Chemical Information and Modeling. - 2003. - 6 : Vol. 43. - pp. 1733–1739.

**Steinmann Heinrich and Chorafas Dimitris N** [Book Section] // Expert Systems in Banking : A Guide for Senior Managers / book auth. Chorafas Dimitris N. - [s.l.] : Macmillan, 1991. - ISBN 10: 0333519396, ISBN 13: 9780333519394.

**System KnowItAll Informatics** Chemical Structure Drawing - ChemWindow [Online] // Bio-Rad Laboratories, Inc. - KnowItAll Informatics System, 2009. - <http://www3.bio-rad.com/>.

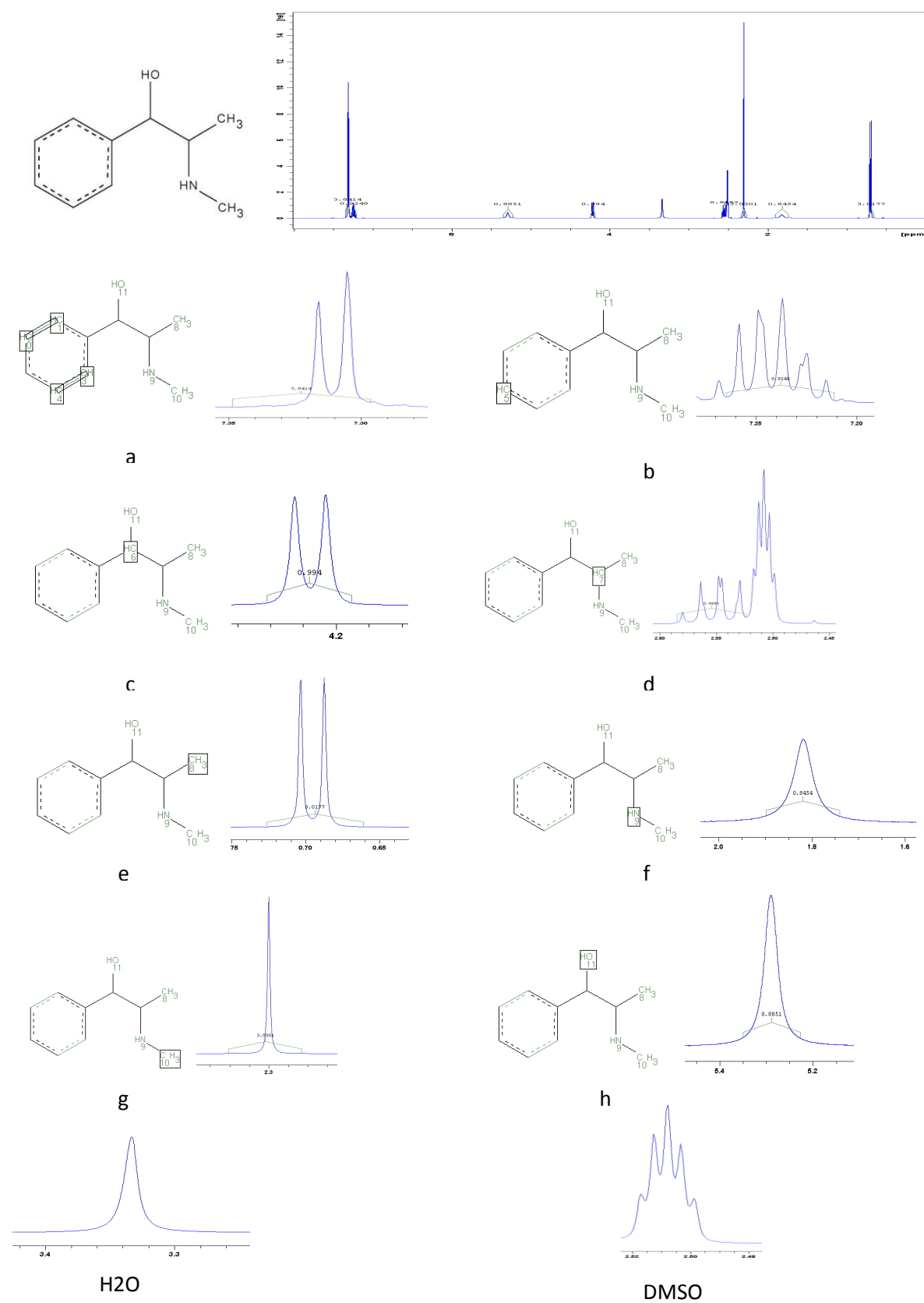
**Taylor Eric W [et al.]** Accelerating the Drug Optimization Process: Identification, Structure Elucidation, and Quantification of in Vivo Metabolites Using Stable Isotopes with LC/MSn and the Chemiluminescent Nitrogen Detector [Journal] // Analytical Chemistry. - 2002. - 13 : Vol. 74. - pp. 3232–3238.

**Wang H [et al.]** An eight-coil high-frequency probehead design for high-throughput nuclear magnetic resonance spectroscopy [Journal] // Journal of Magnetic Resonance. - October 2004. - 2 : Vol. 170. - pp. 206-212.

**Williams Antony J** Recent advances in NMR prediction and automated structure elucidation software [Journal] // Current Opinion in Drug Discovery & Development. - 2000. - Vol. 3. - pp. 298-305.

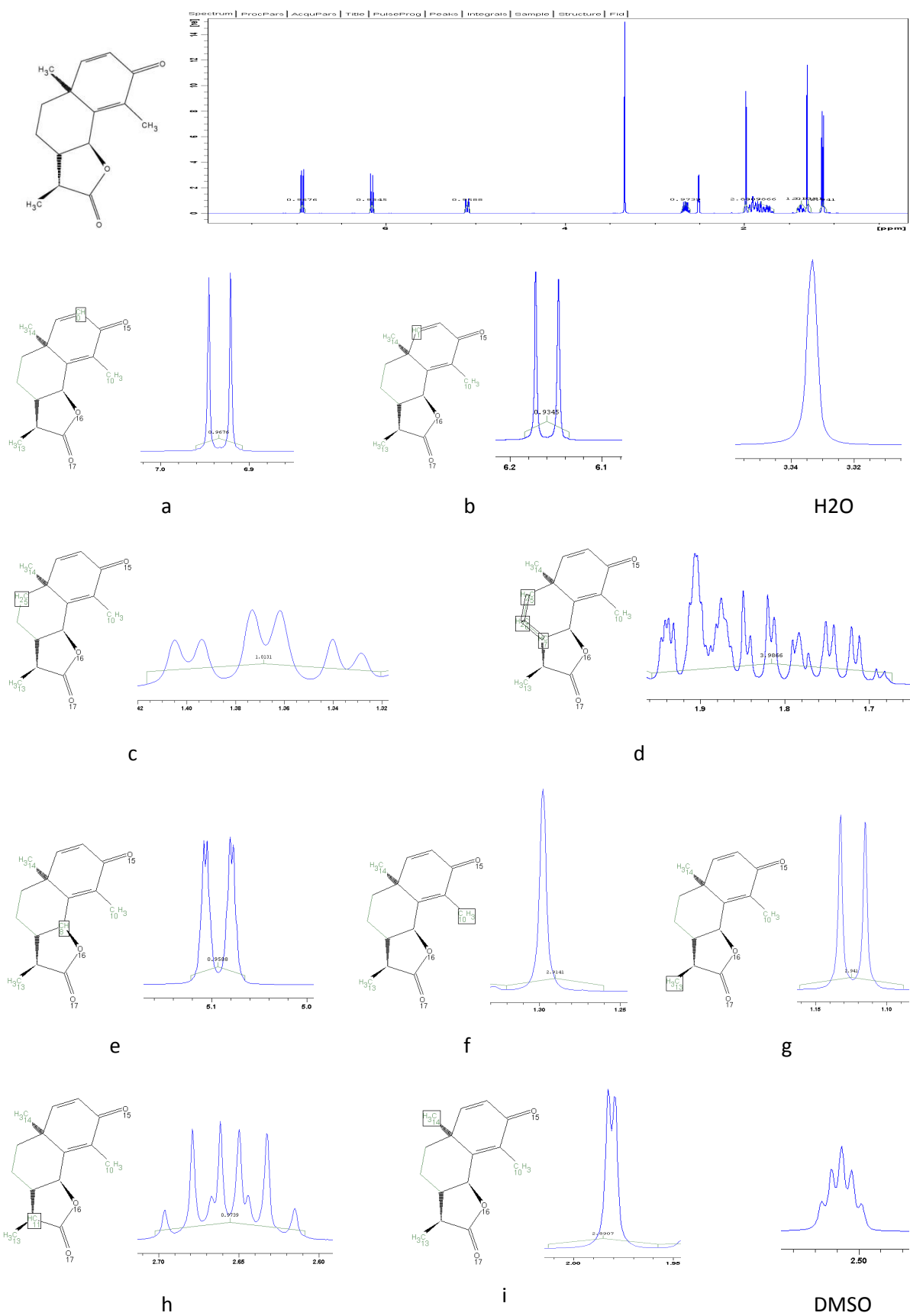
## **C. List of Detailed Assignments of 81 Spectrum-Structure Pairs**

## 1. +Pseudoephedrin

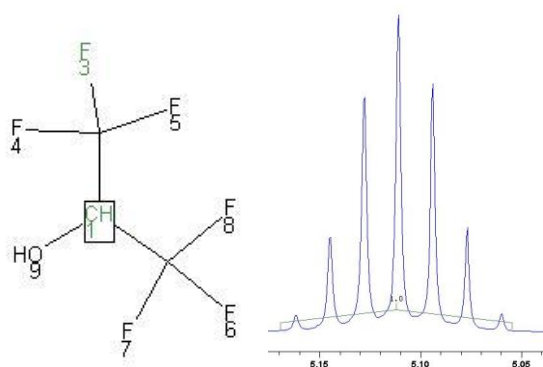
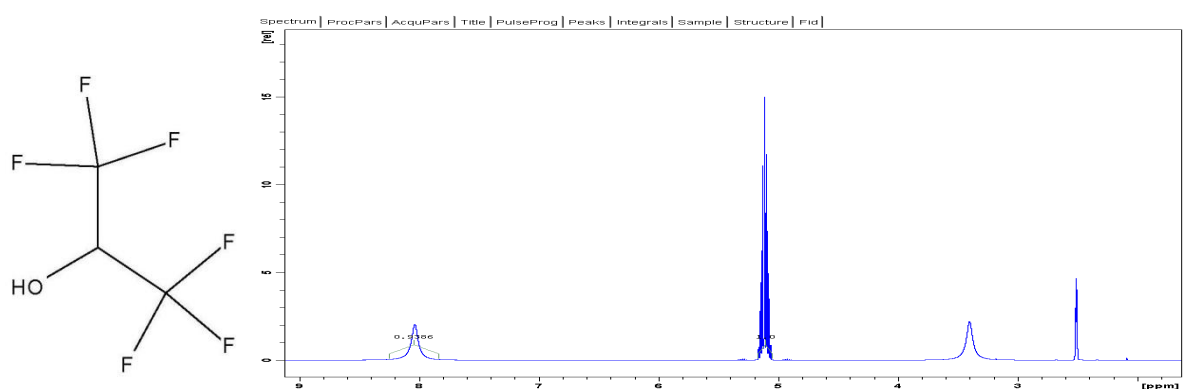




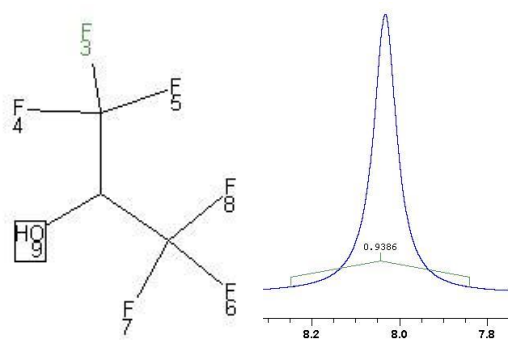
## 2. --alpha-Satonin



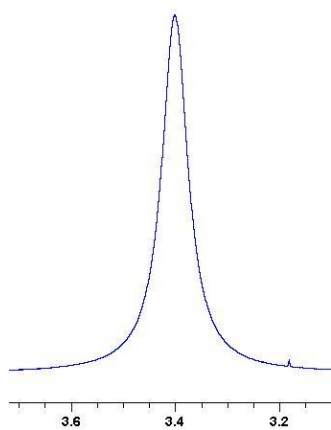
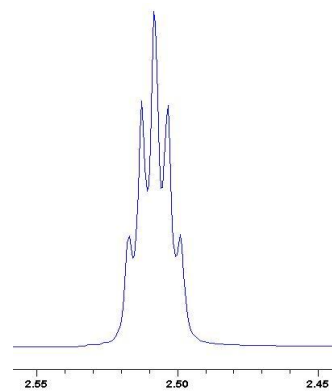
## 3. 1,1,1-3,3,3-Hexafluor-2-propanol



a

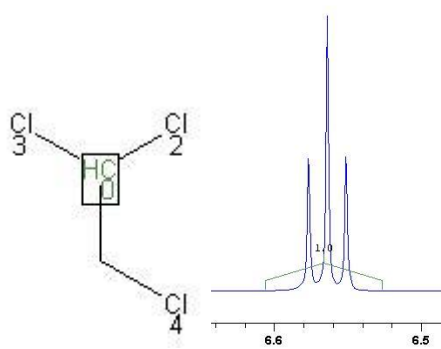
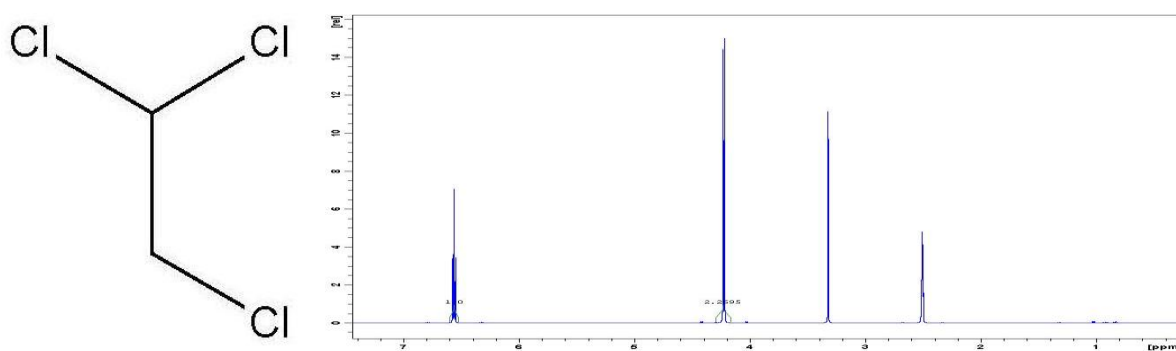


b

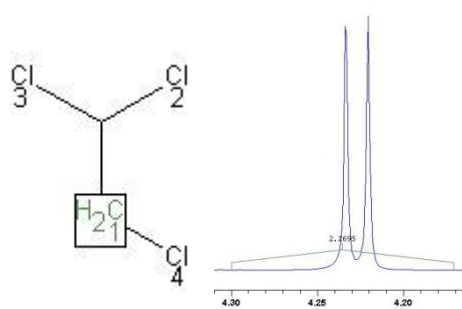
H<sub>2</sub>O

DMSO

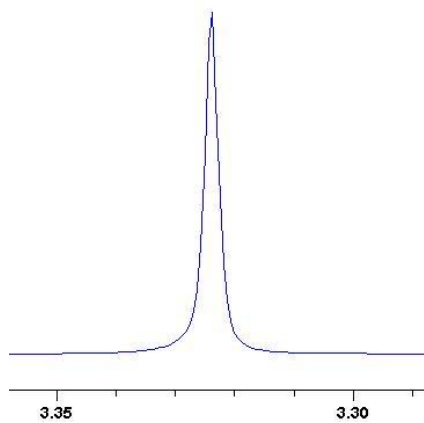
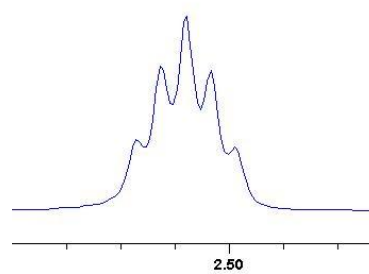
## 4. 1,1,2-Trichlorethan



a

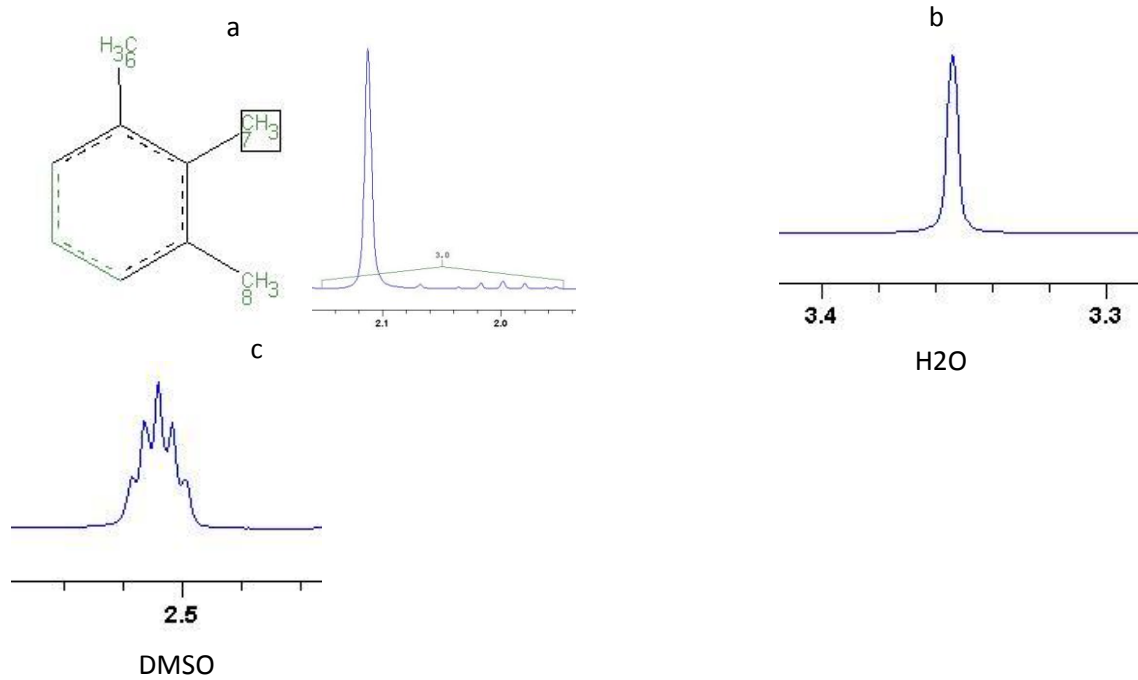
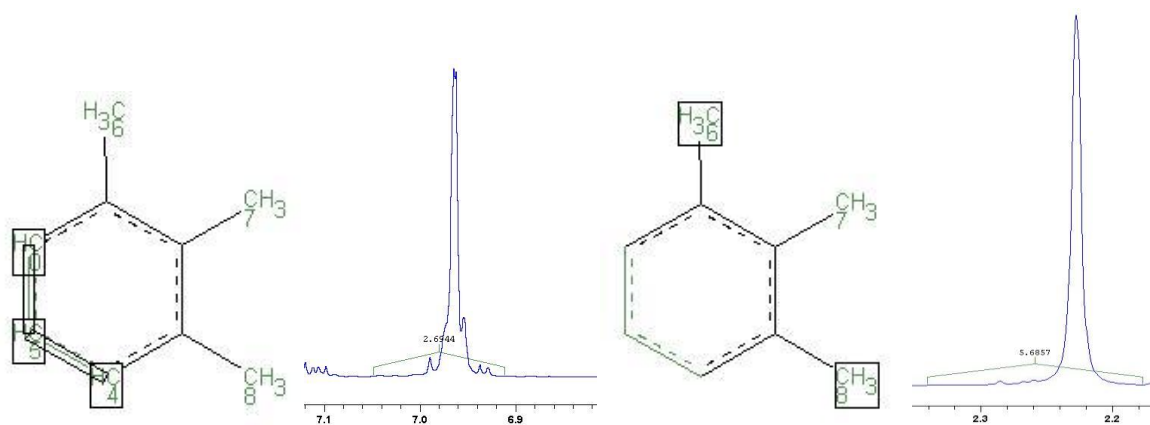
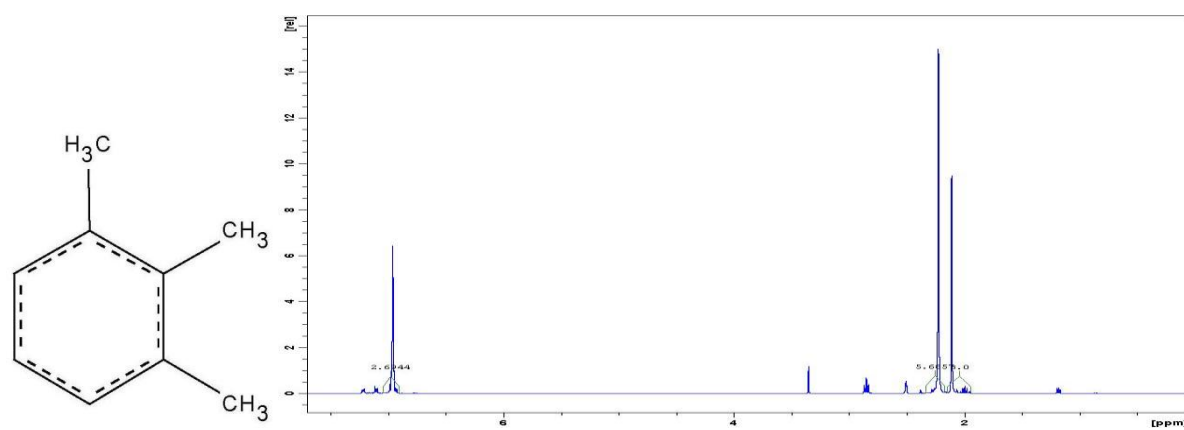


b

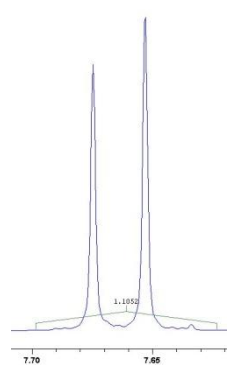
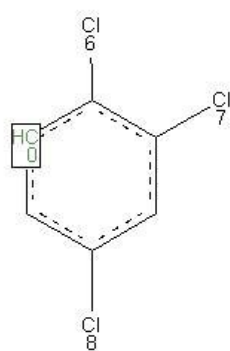
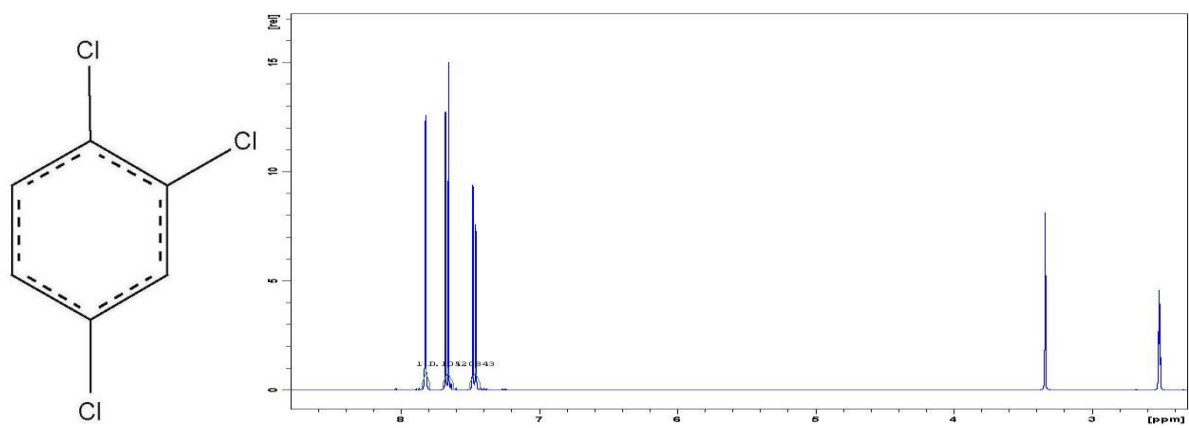
 $\text{H}_2\text{O}$ 

DMSO

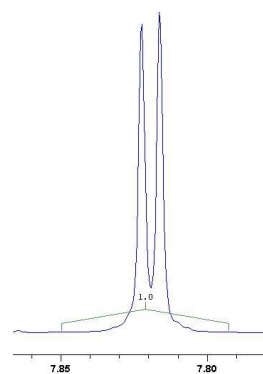
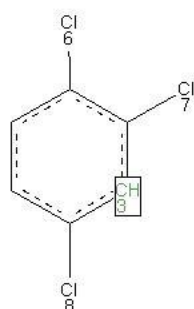
## 5. 1,2,3-Trimethylbenzol



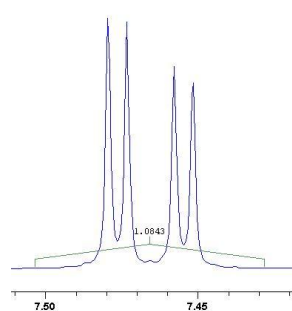
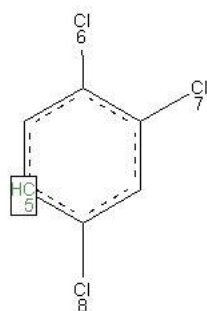
6. 1,2,4-Trichlorbenzol



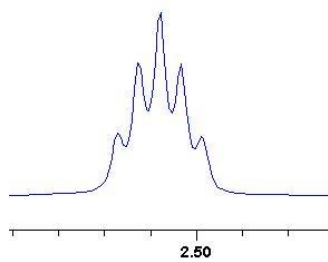
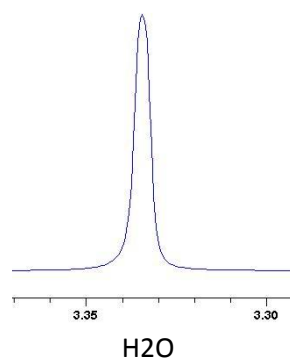
a



b

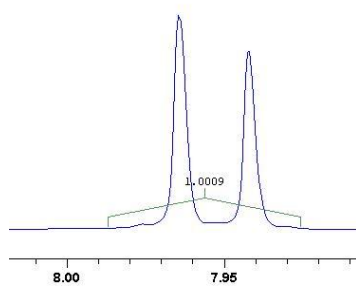
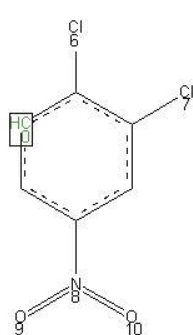
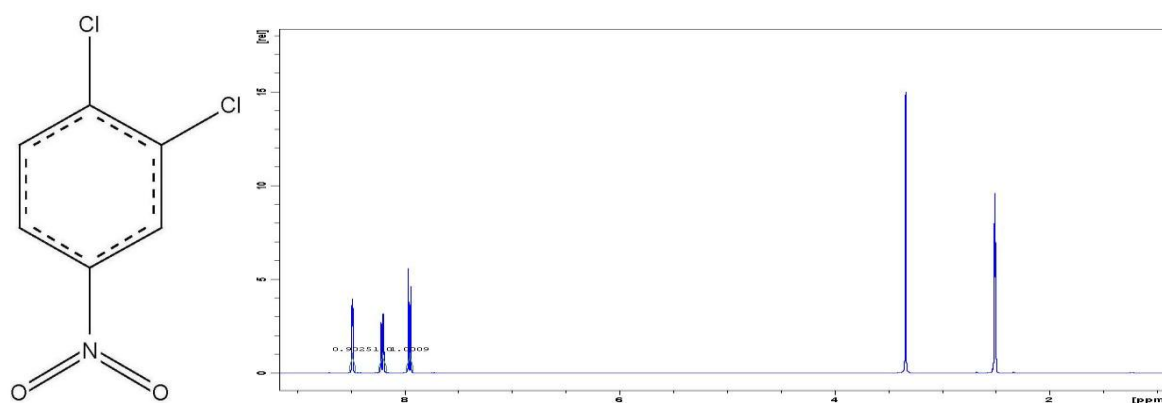


c

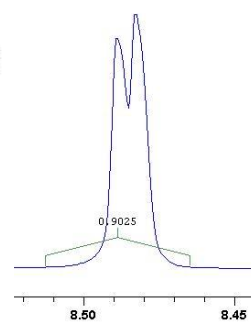
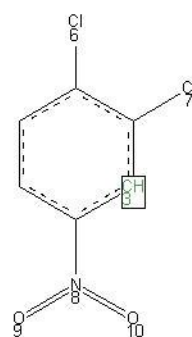


DMSO

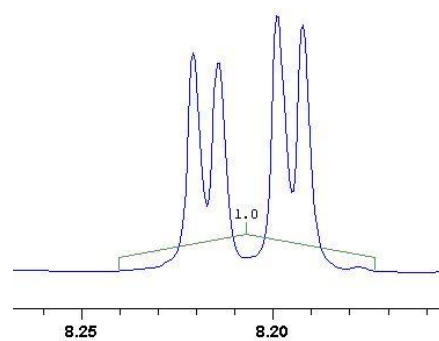
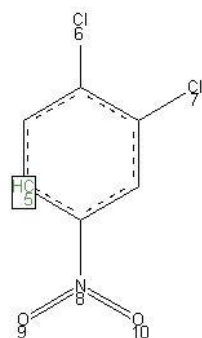
## 7. 1,2-Dichlor-4-nitrobenzol



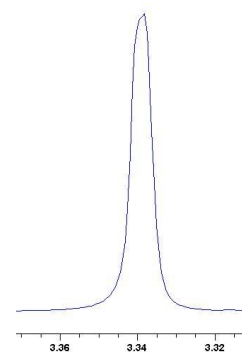
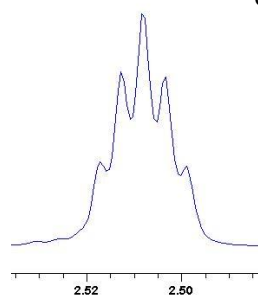
a



b

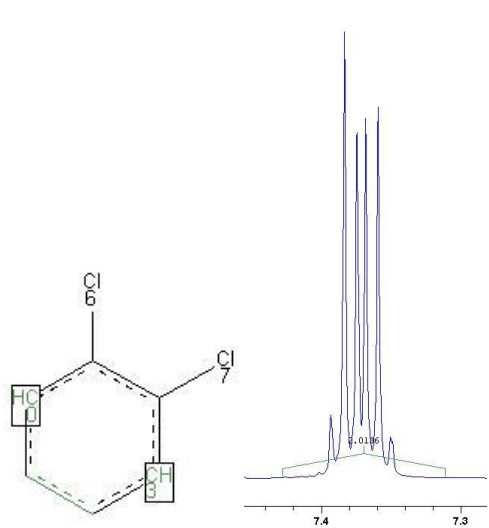
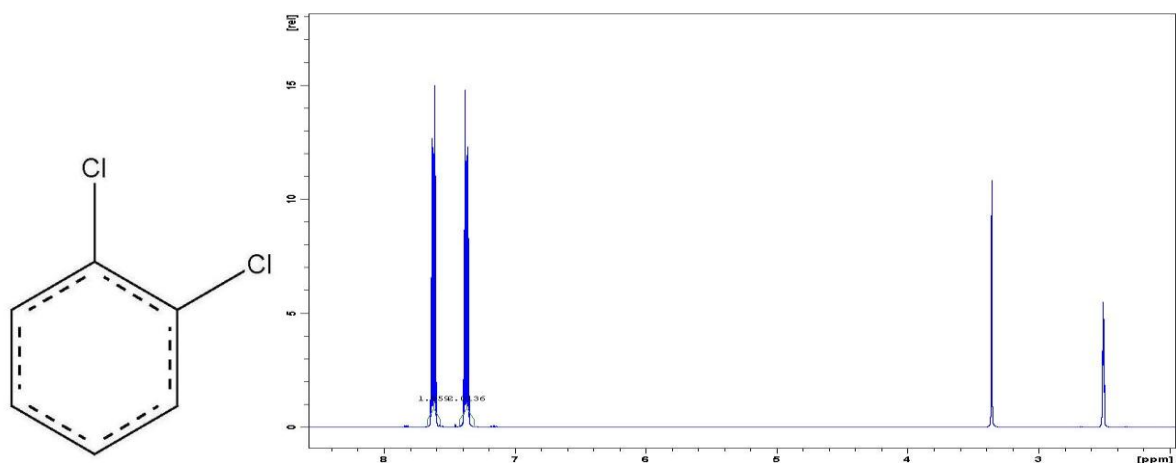


c

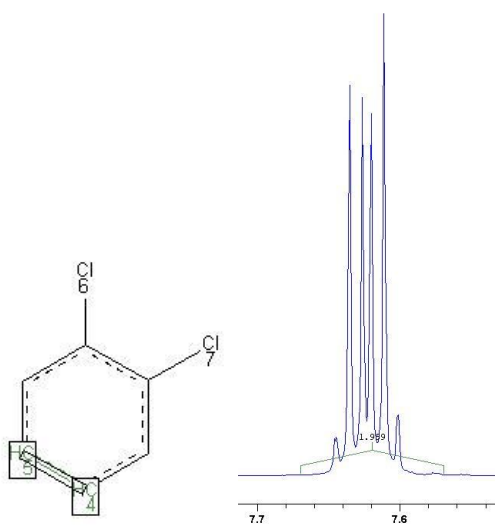
H<sub>2</sub>O

DMSO

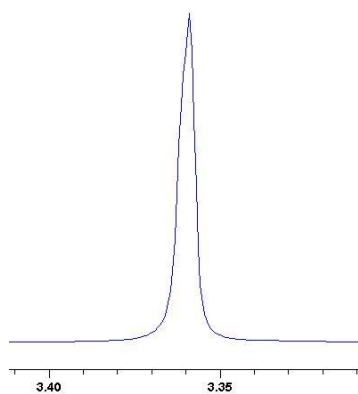
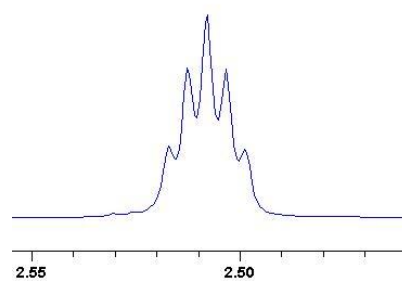
## 8. 1,2-Dichlorbenzol



a

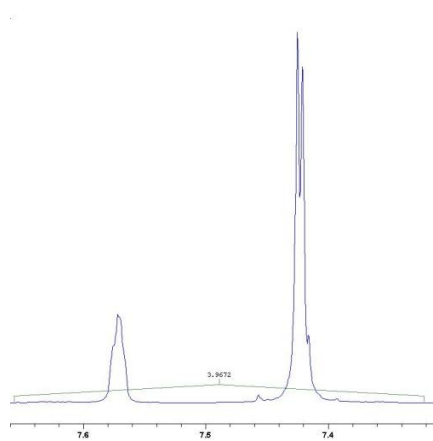
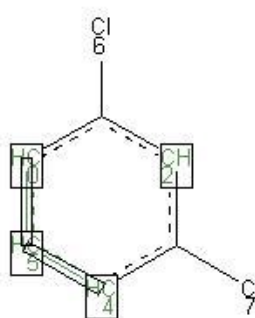
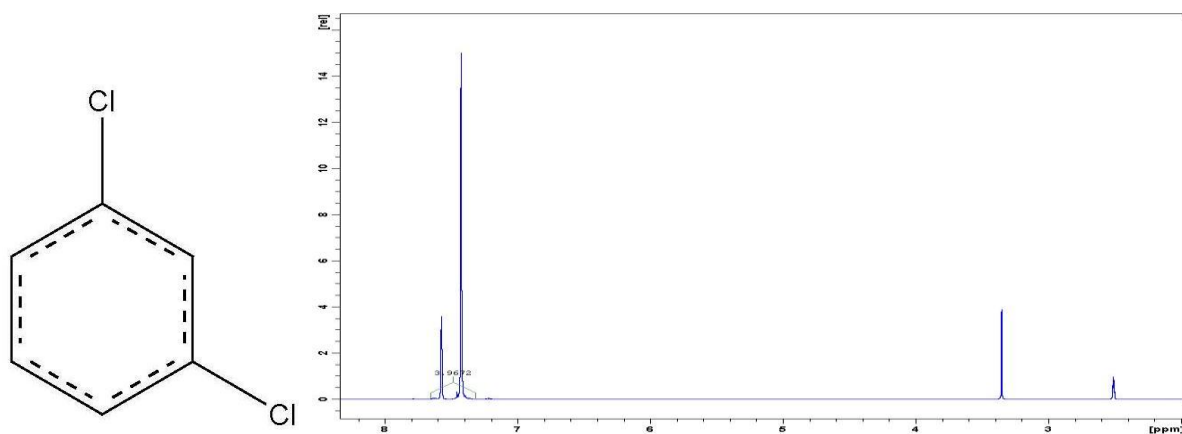


b

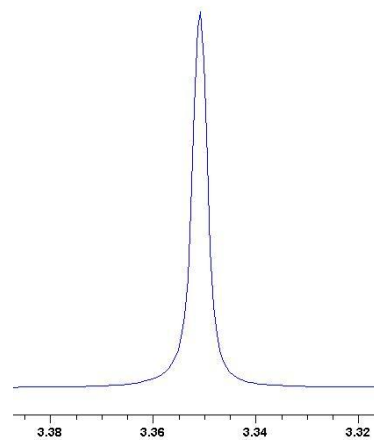
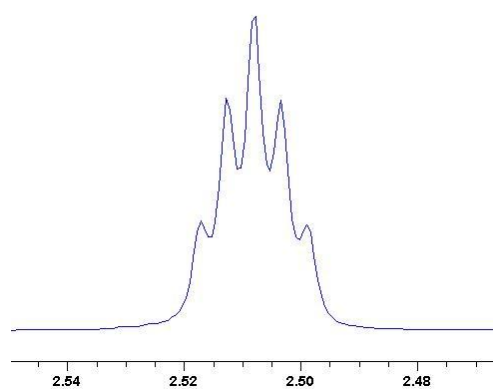
 $\text{H}_2\text{O}$ 

DMSO

## 9. 1,3-Dichlorbenzol



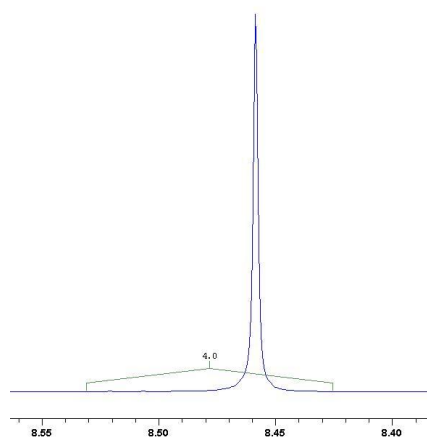
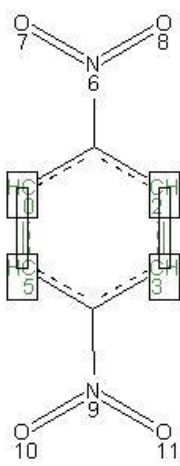
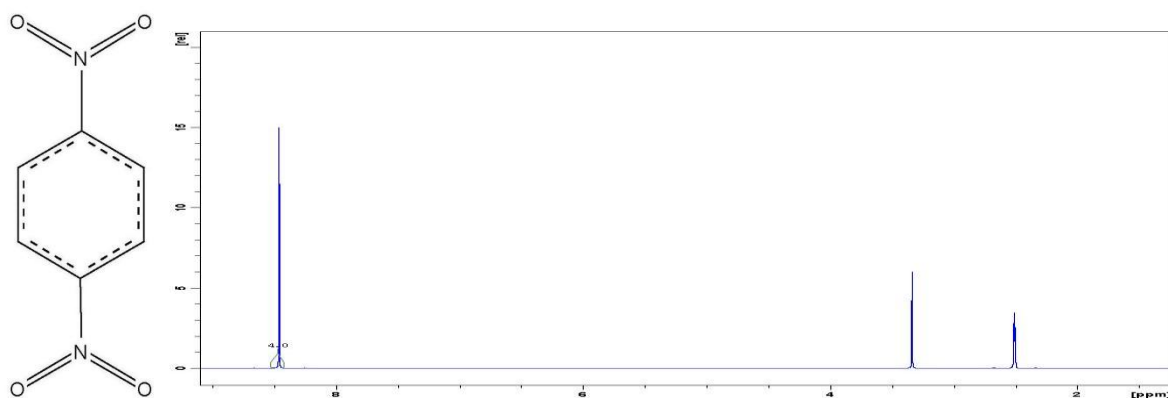
a

 $\text{H}_2\text{O}$ 

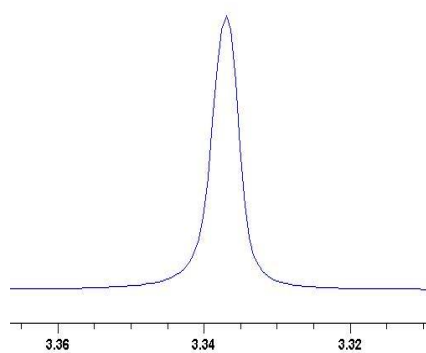
DMSO



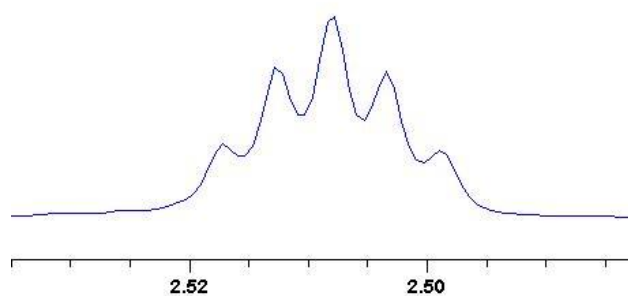
10. 1,4-Dinitrobenzol



a

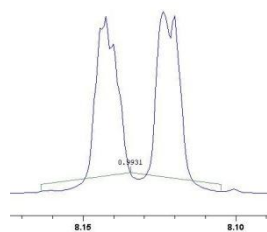
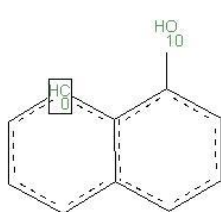
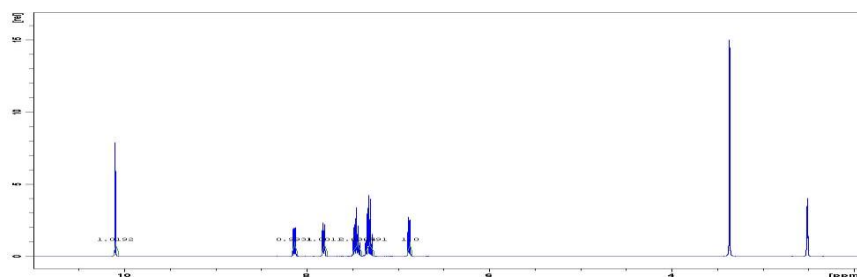
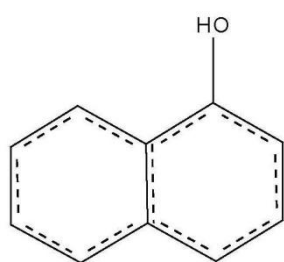


H<sub>2</sub>O

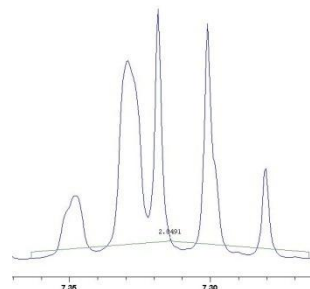
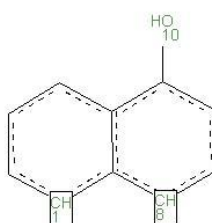


DMSO

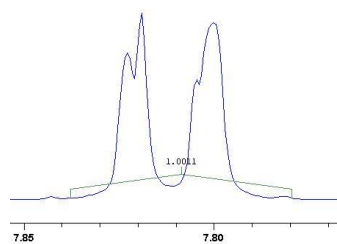
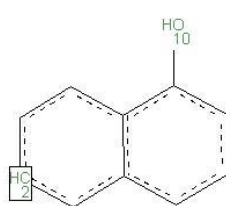
## 11. 1-Naphthol



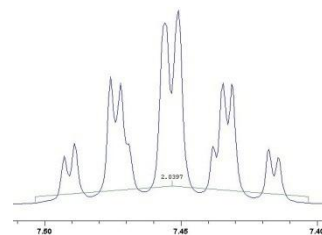
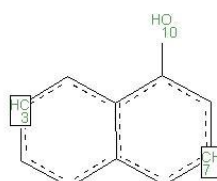
a



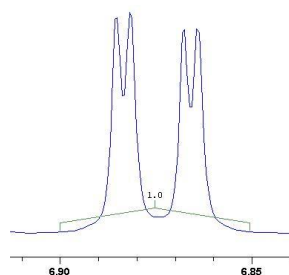
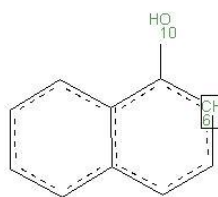
b



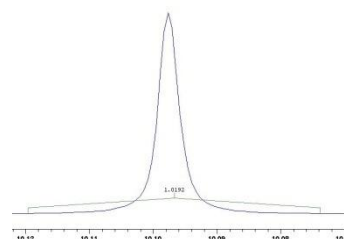
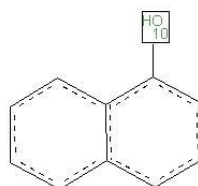
C



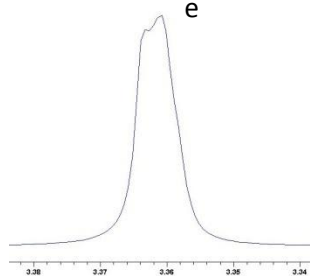
d



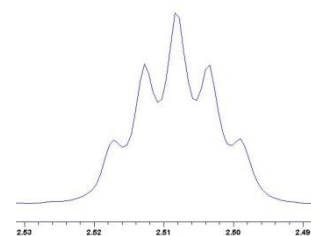
e



**f**

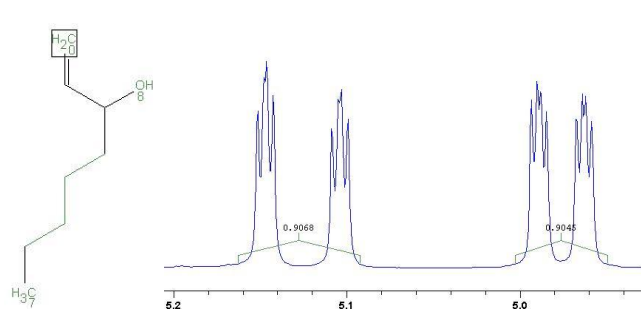
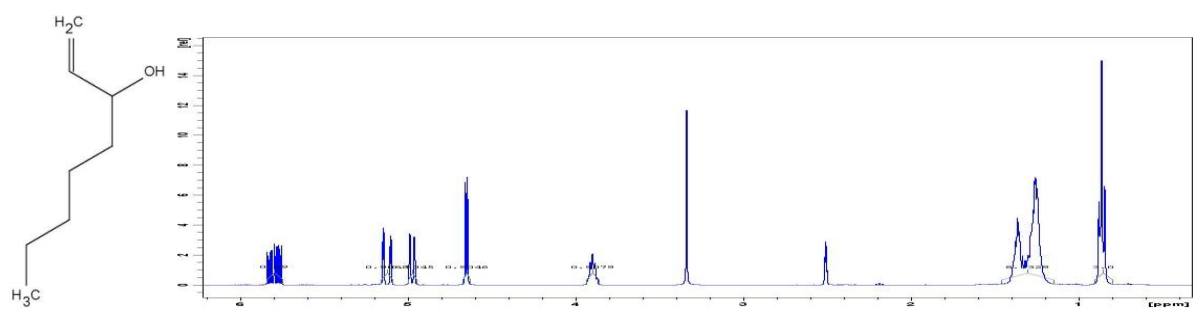


H<sub>2</sub>O

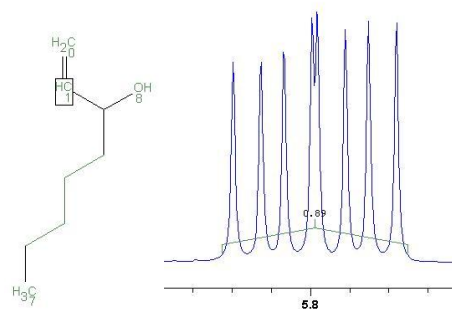


DMSO

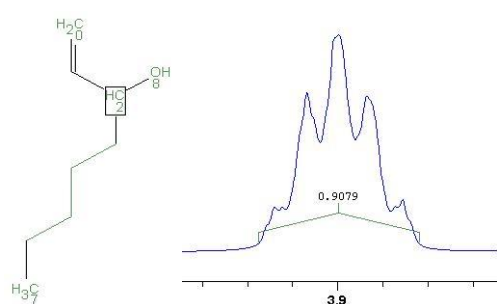
## 12. 1-Octen-3-ol



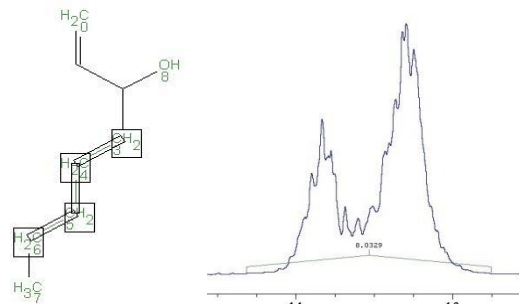
a



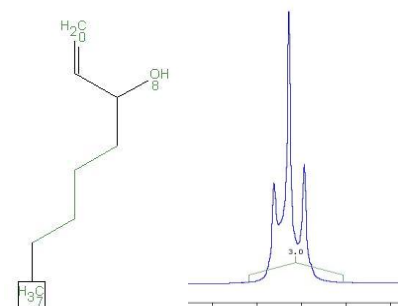
b



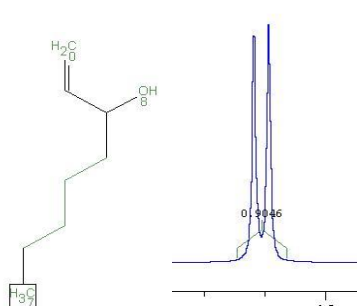
c



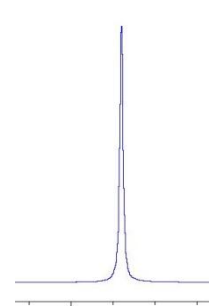
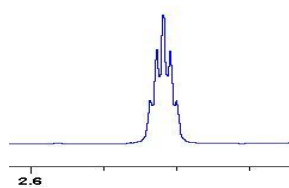
d



e

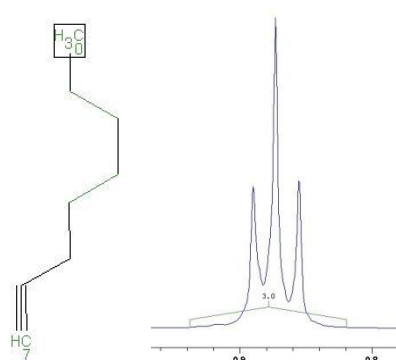
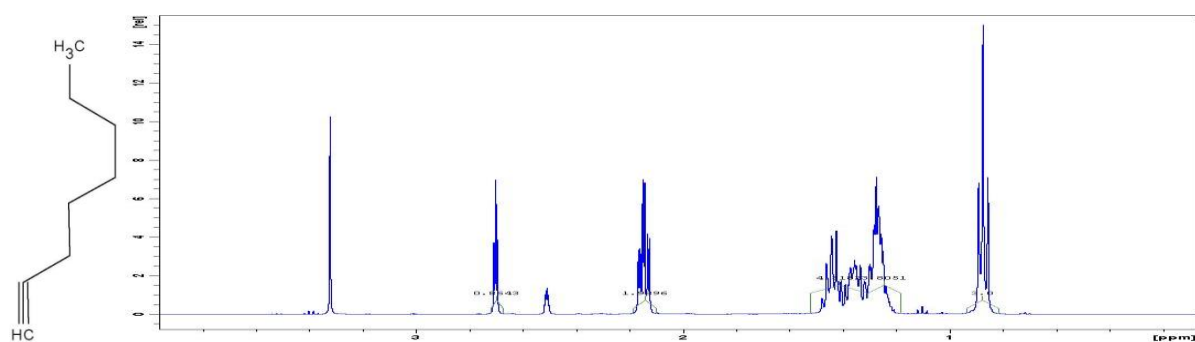


f

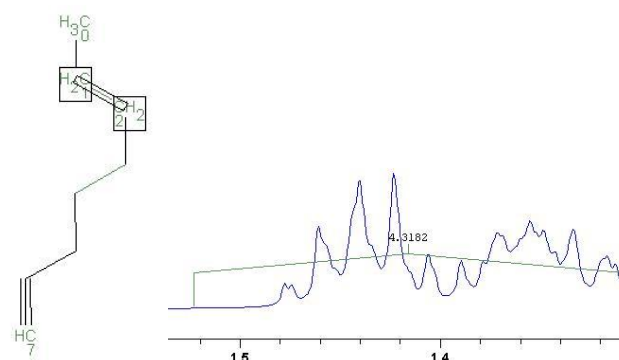
H<sub>2</sub>O

DMSO

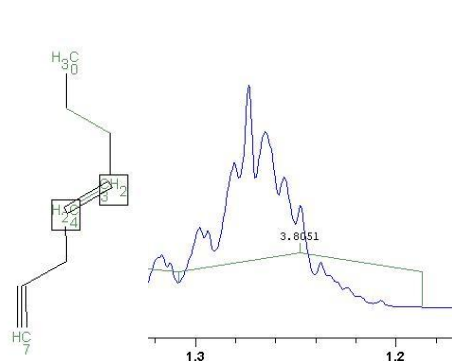
## 13. 1-Octyne



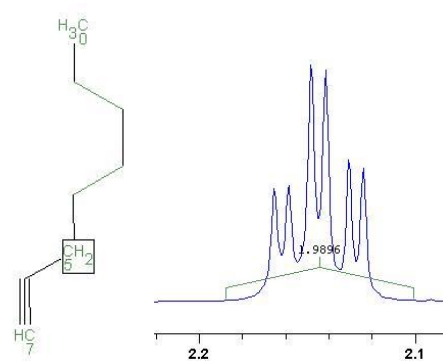
a



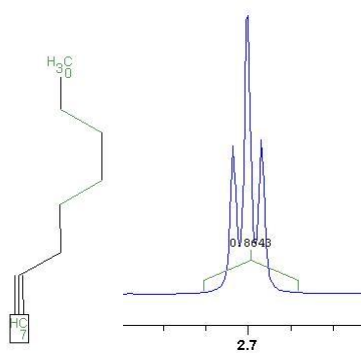
b



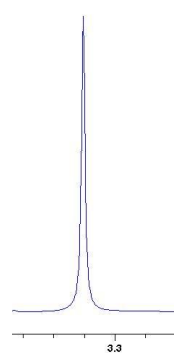
c



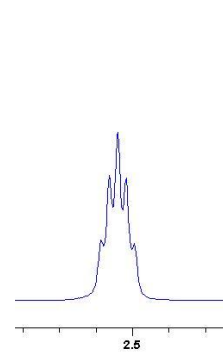
d



e

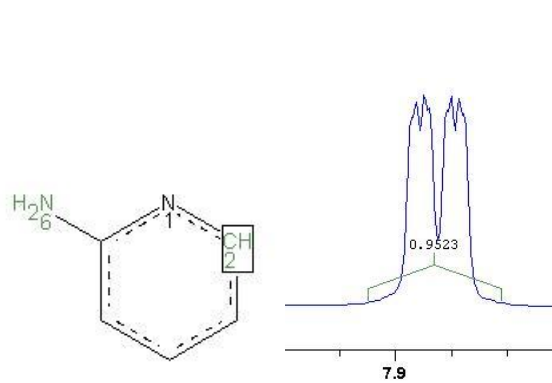
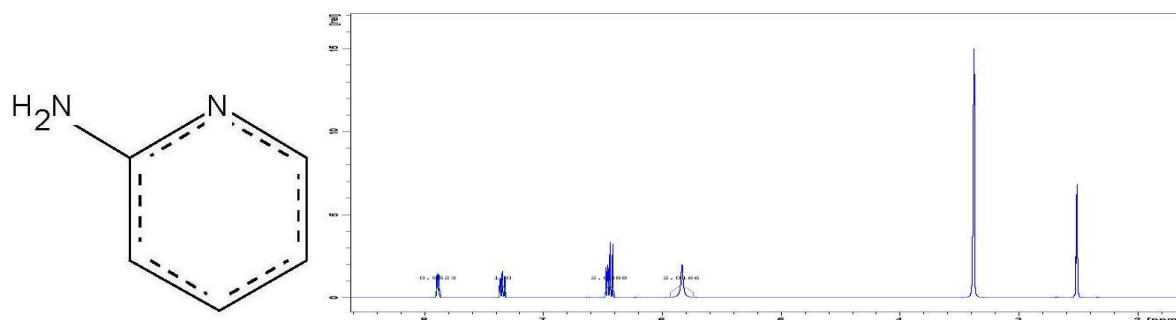


H2O

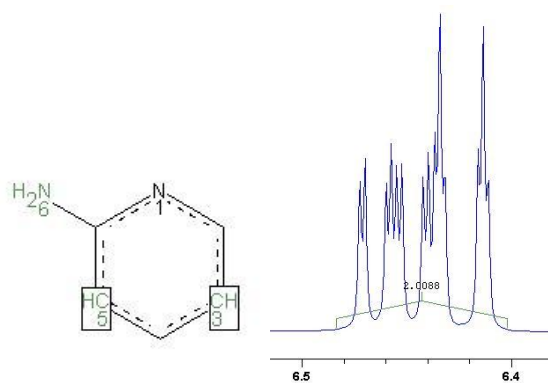


DMSO

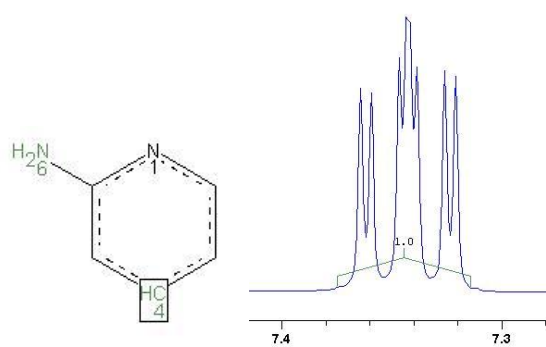
## 14. 2-Aminopyridin



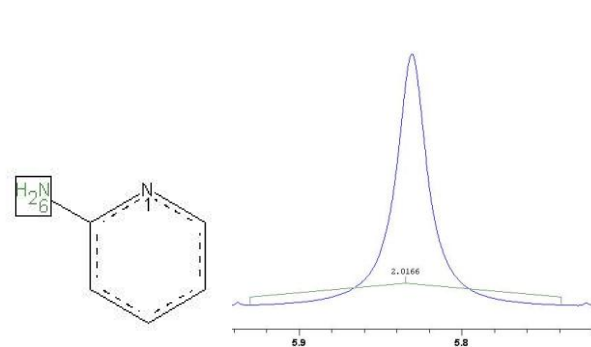
a



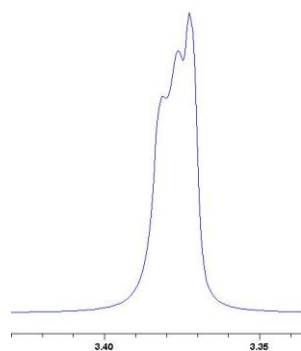
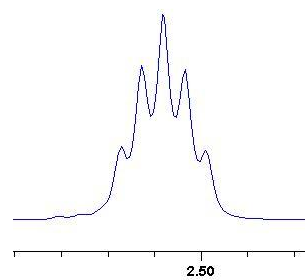
b



c

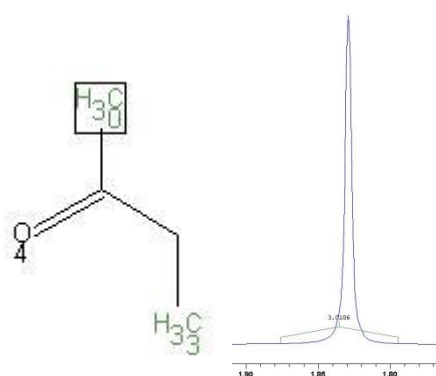
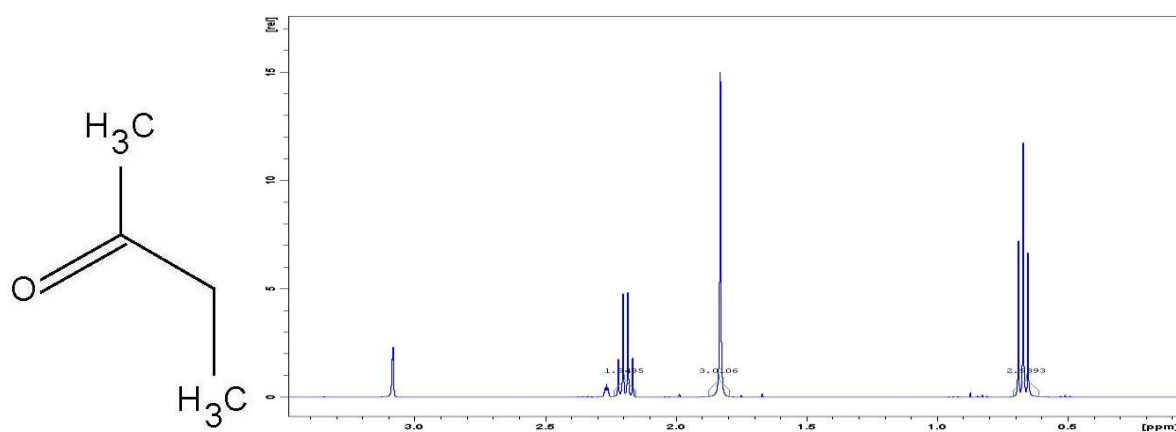


d

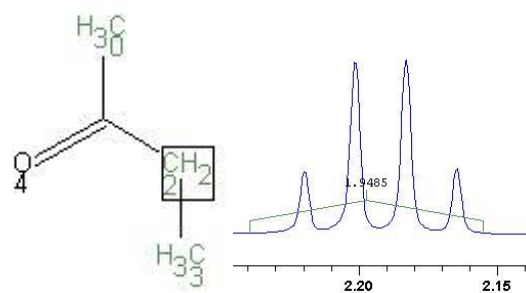
H<sub>2</sub>O

DMSO

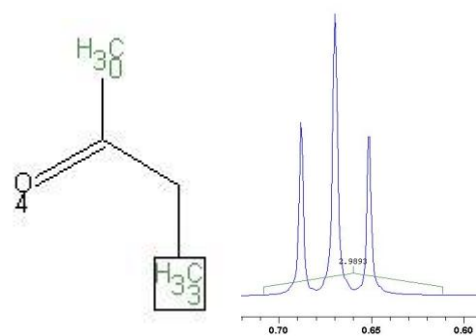
## 15. 2-Butanon



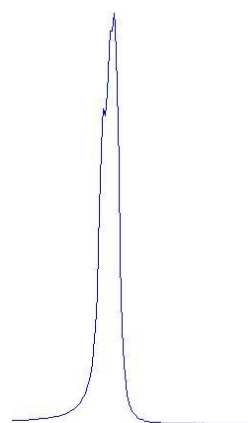
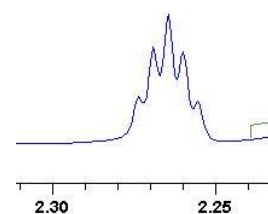
a



b

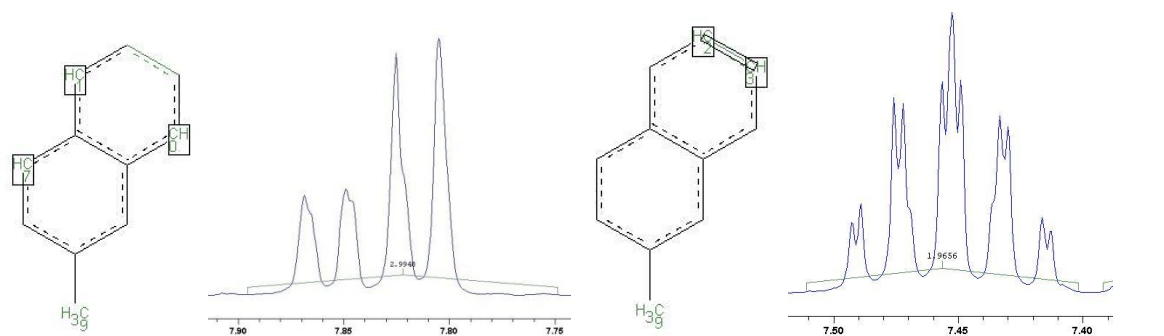
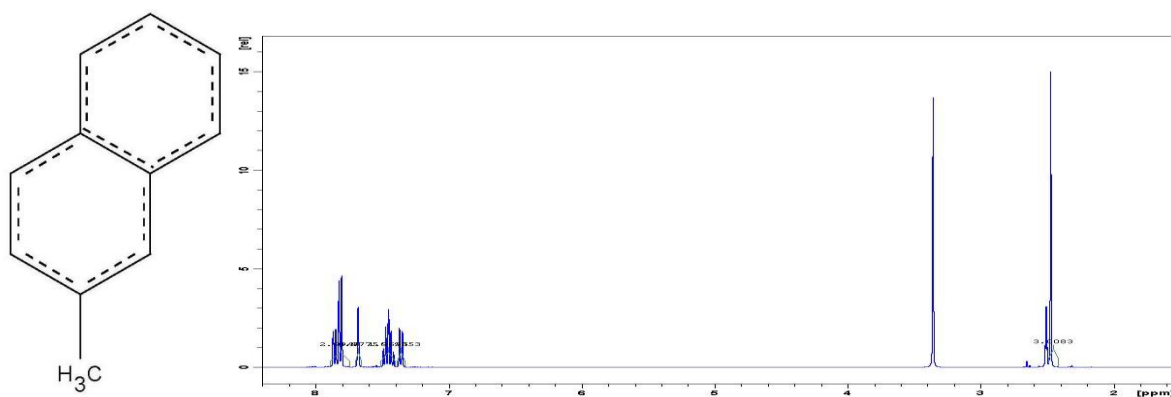


c

H<sub>2</sub>O

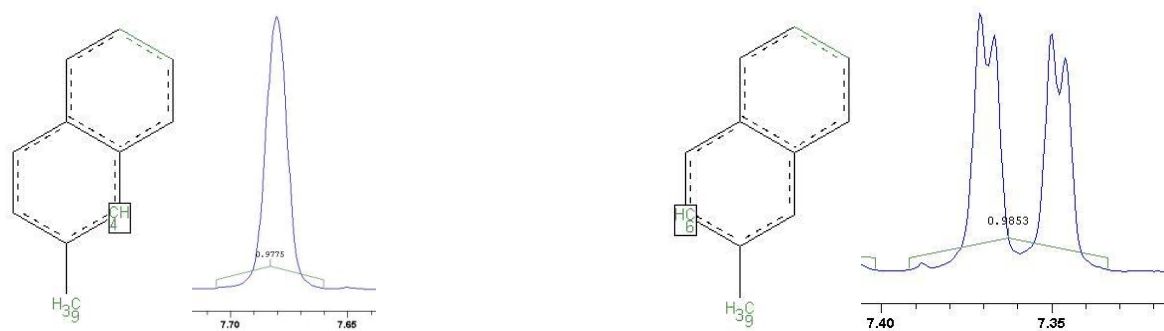
DMSO

## 16. 2-Methyl-naphthalin



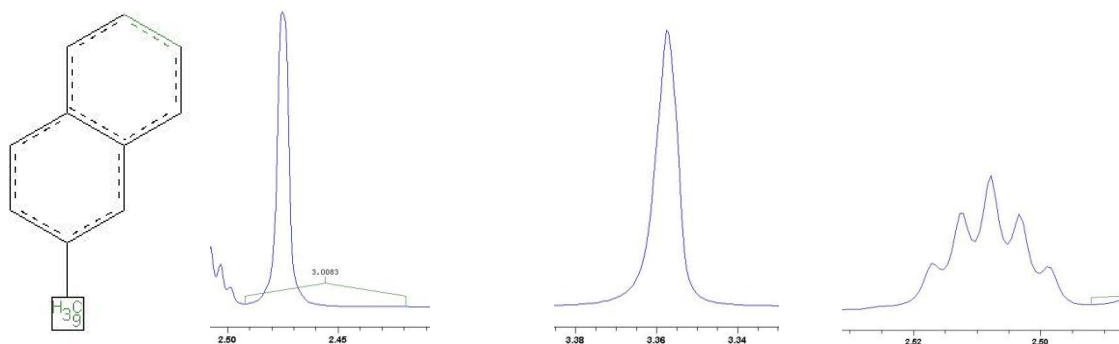
a

b



c

d

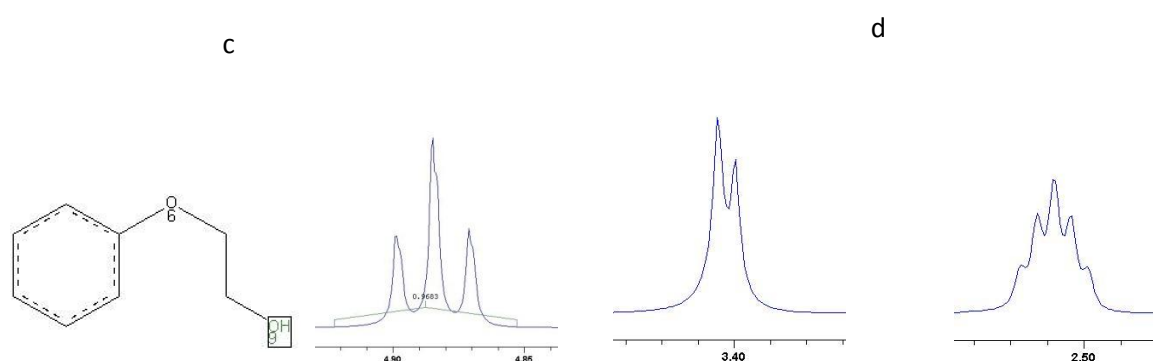
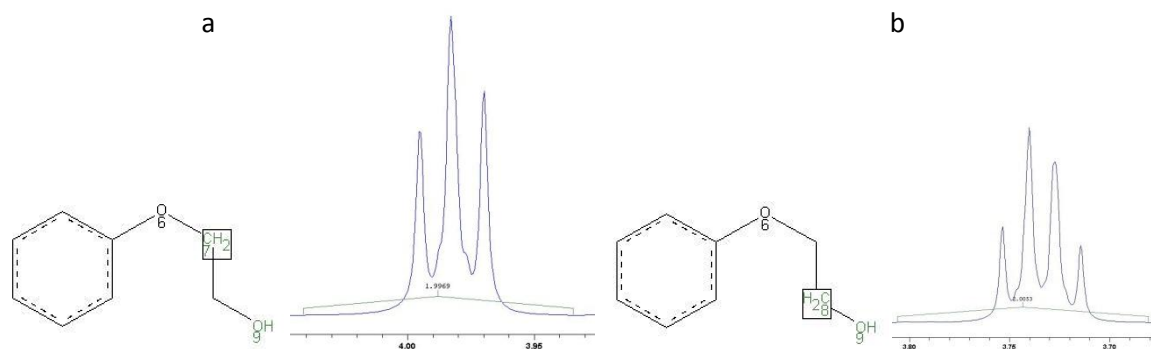
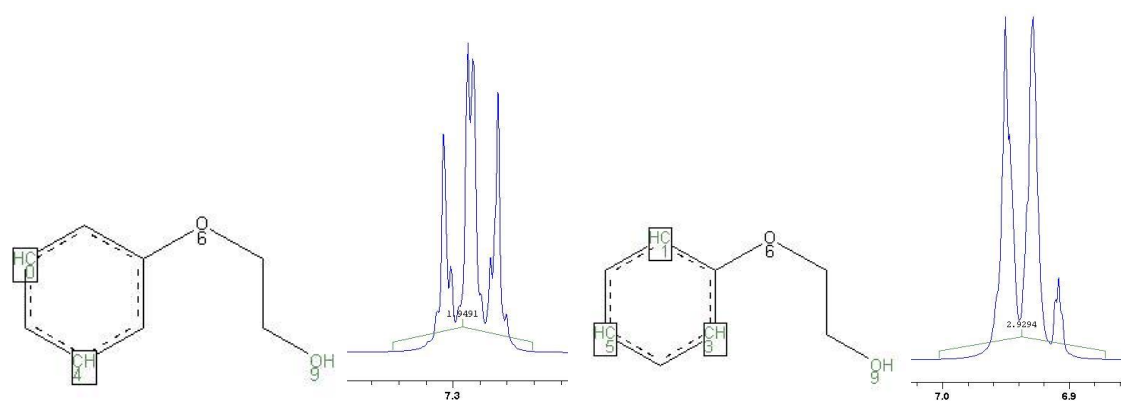
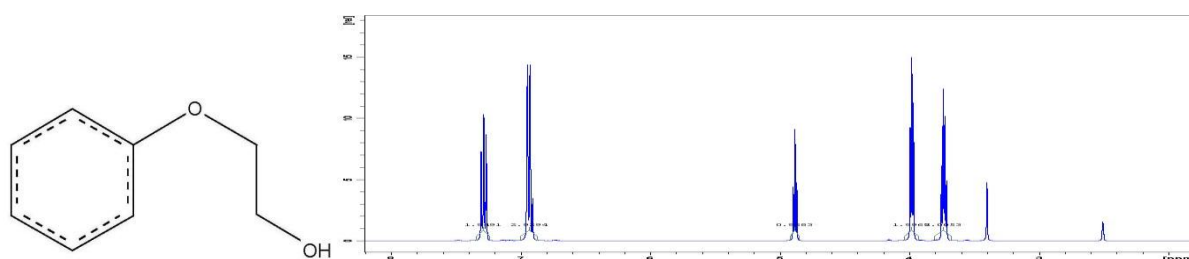


e

H<sub>2</sub>O

DMSO

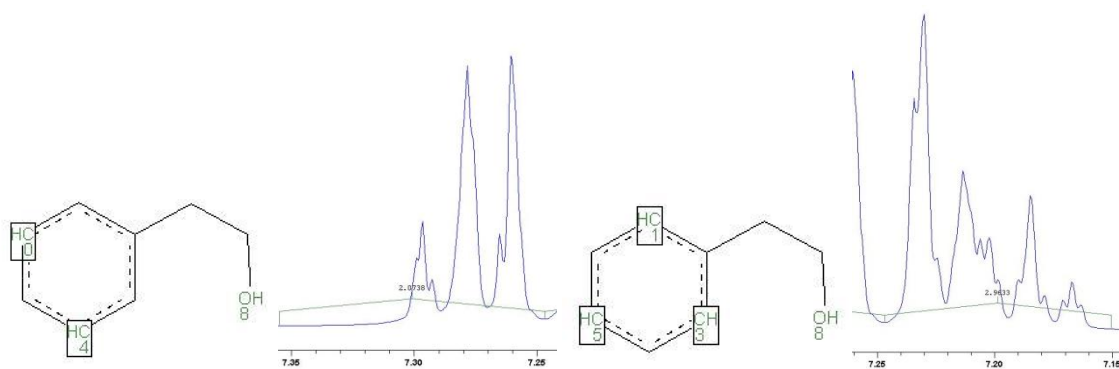
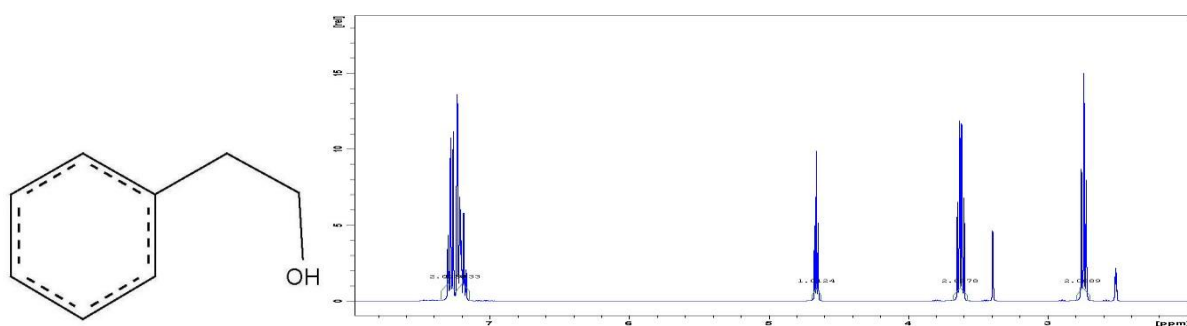
## 17. 2-Phenoxyethanol

H<sub>2</sub>O

DMSO

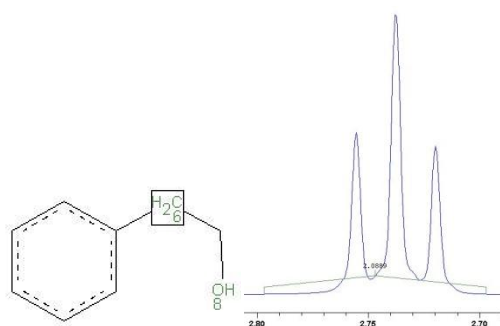


## 18. 2-phenyl-ethylalcohol

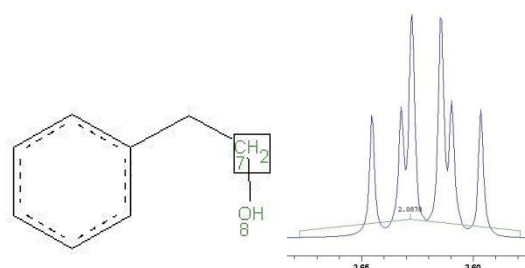


a

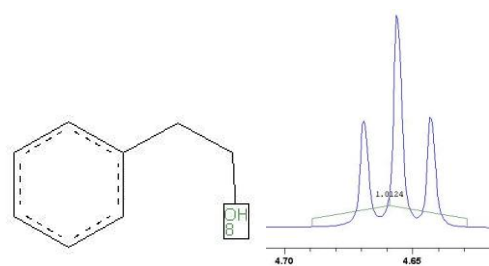
b



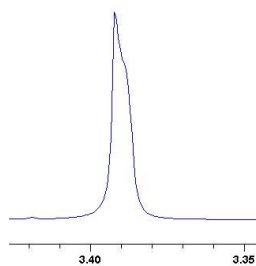
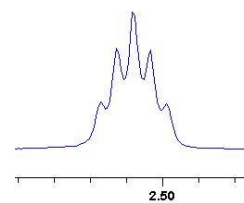
c



d

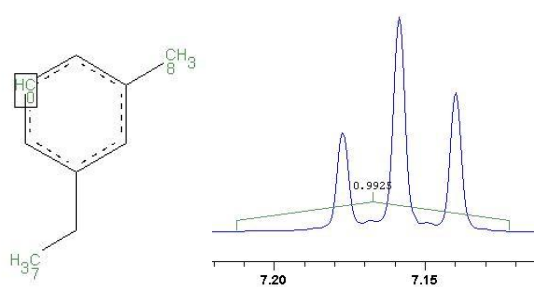
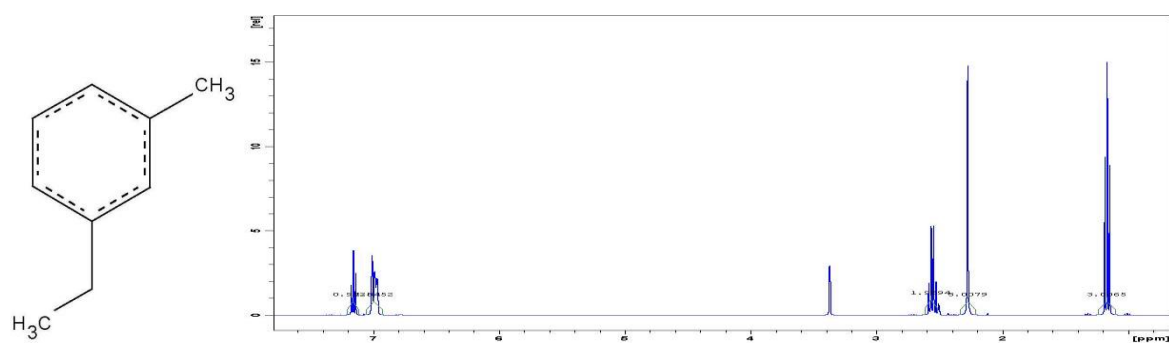


e

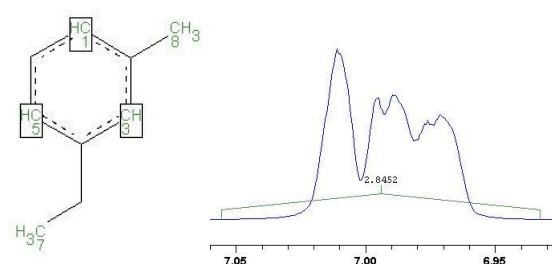
H<sub>2</sub>O

DMSO

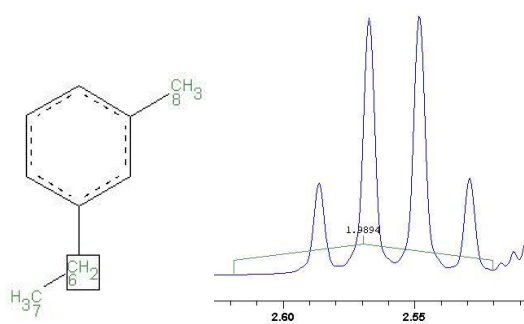
## 19. 3-Ethyltoluol



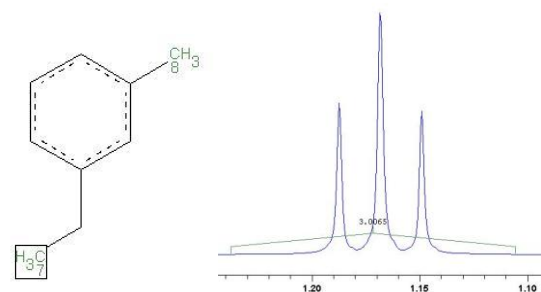
a



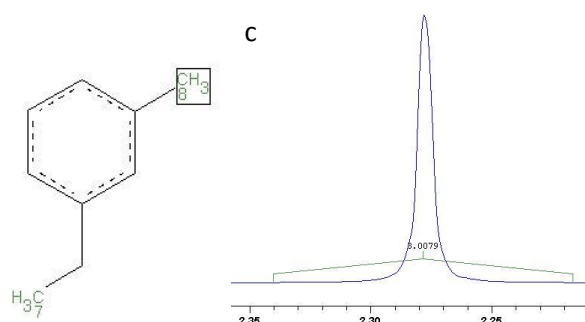
b



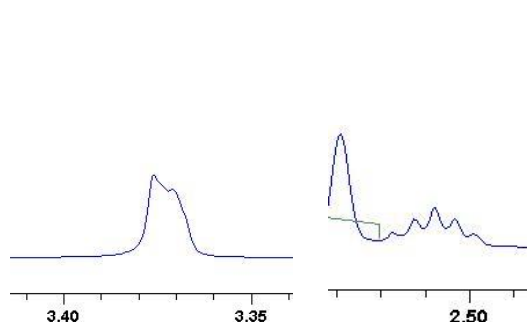
c



d

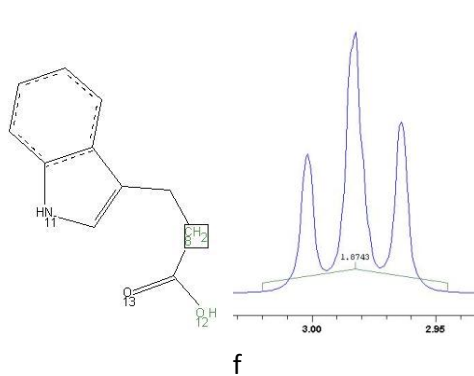
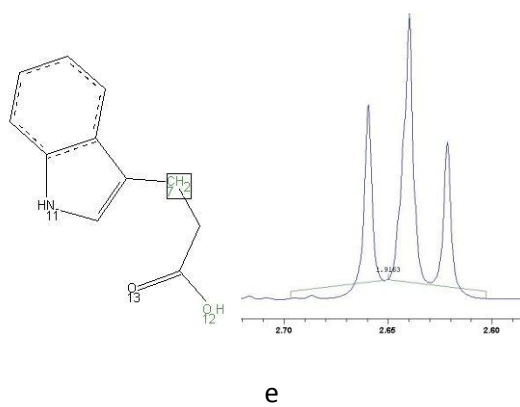
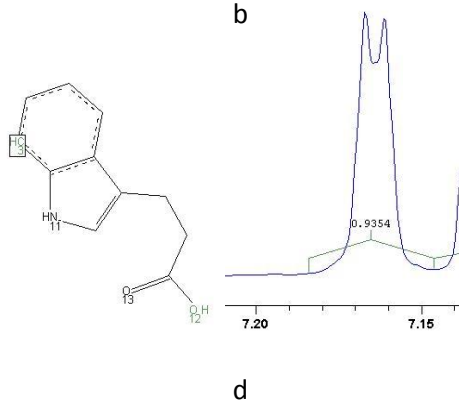
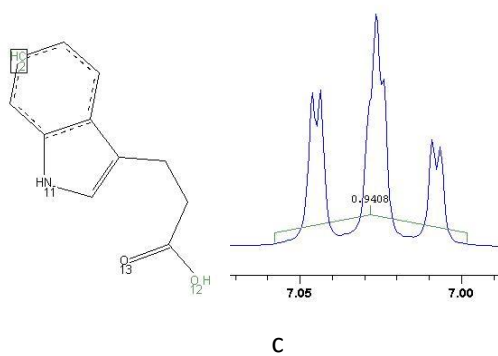
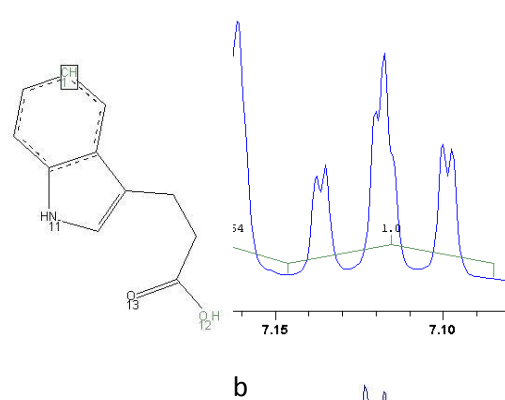
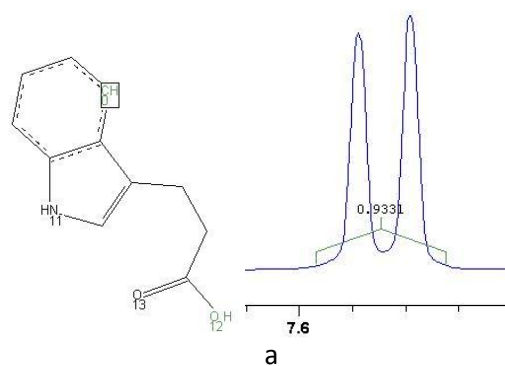
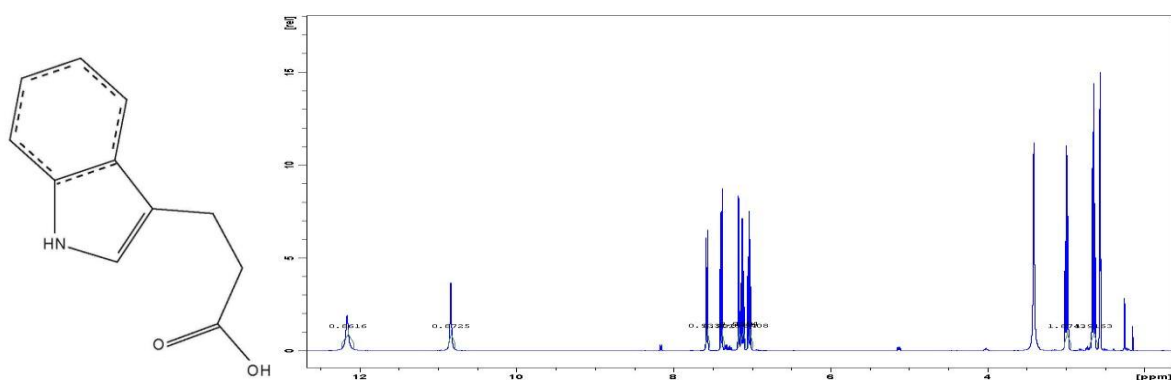


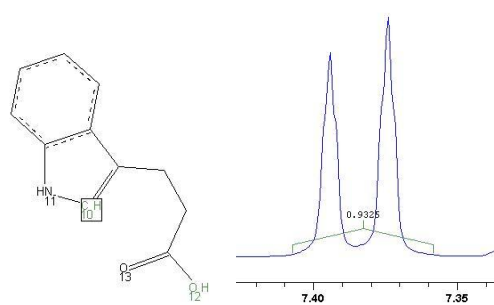
e

H<sub>2</sub>O

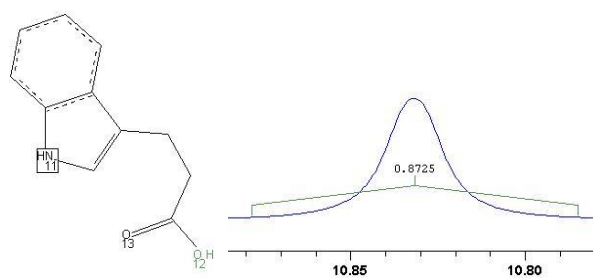
DMSO

## 20. 3-Indolepropionic acid

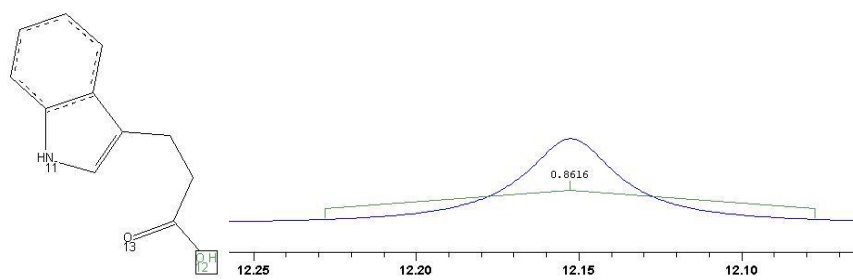




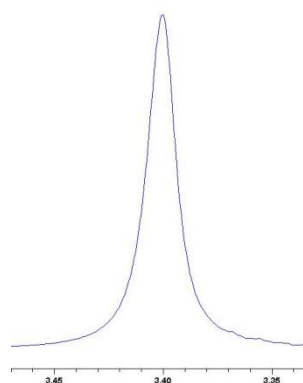
g



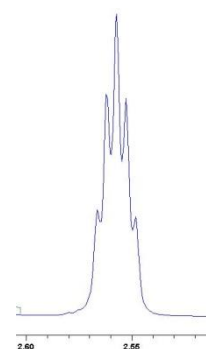
h



i

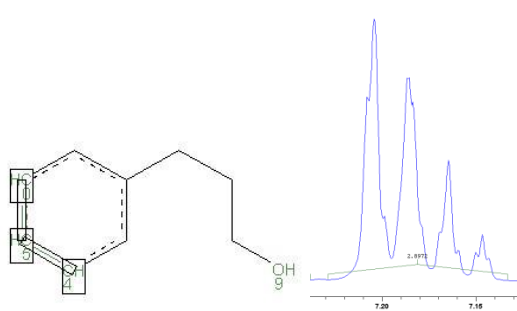
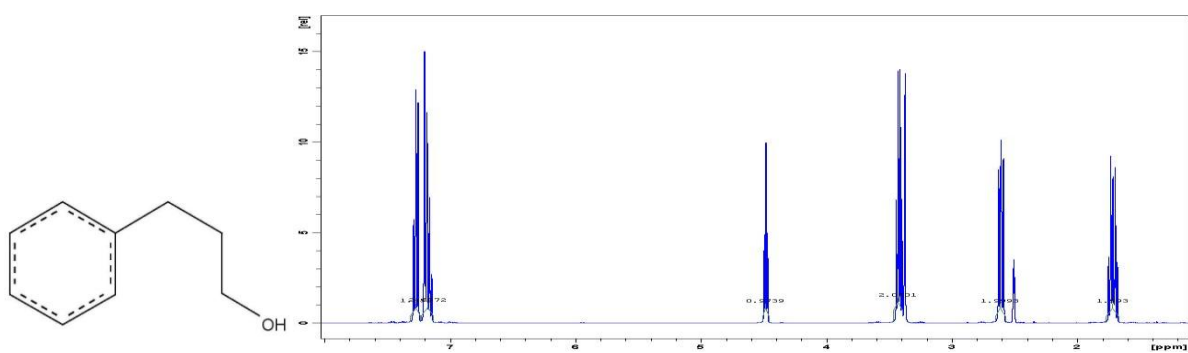


H2O

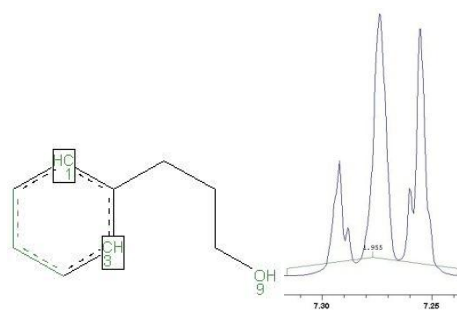


DMSO

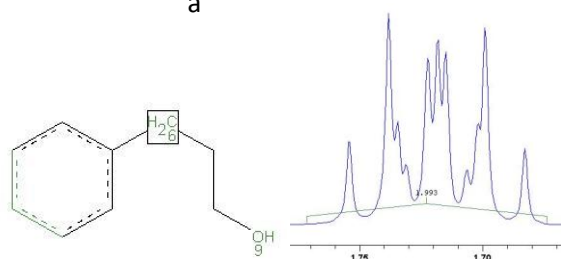
## 21. 3-Phenyl-propylalcohol



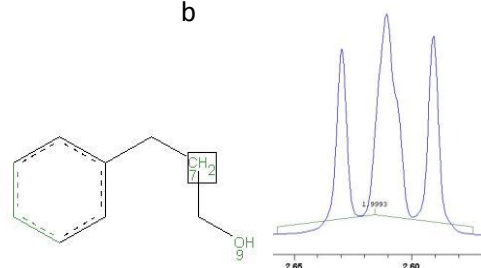
a



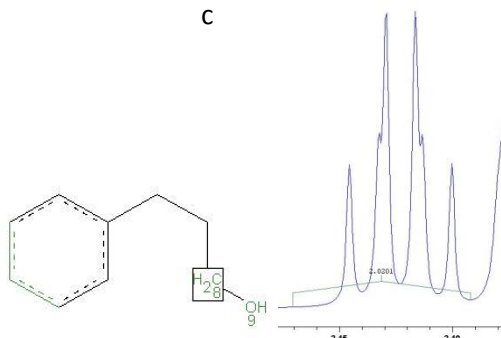
b



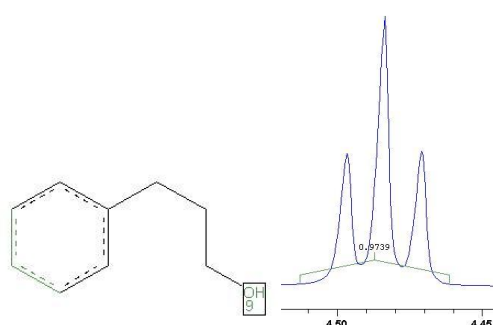
c



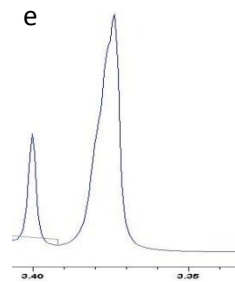
d



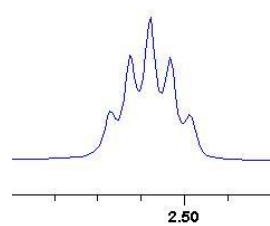
e



f

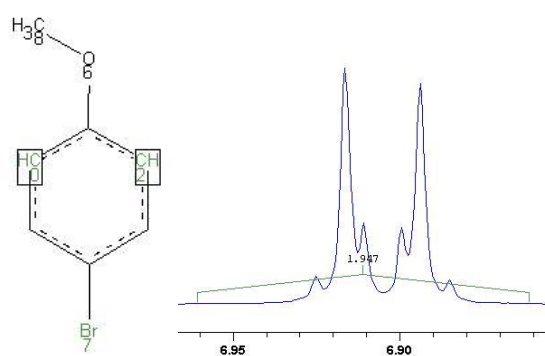
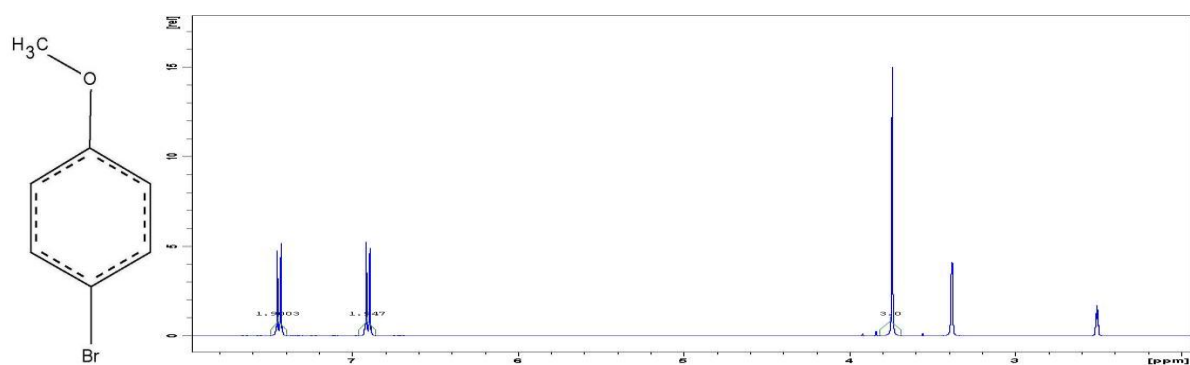


H2O

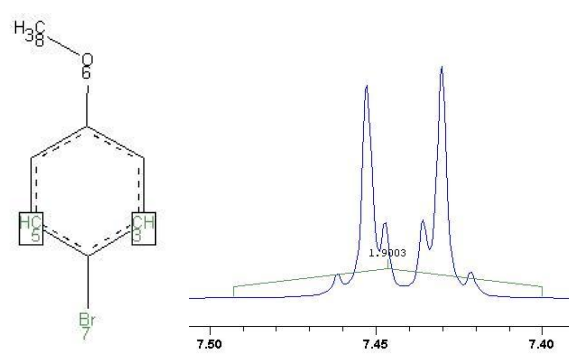


DMSO

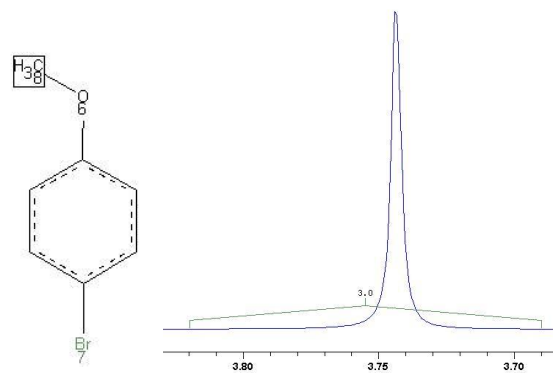
## 22. 4-Bromanisol



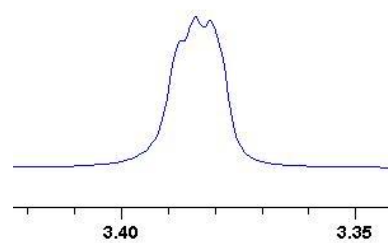
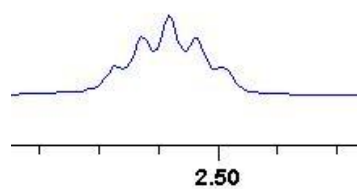
a



b

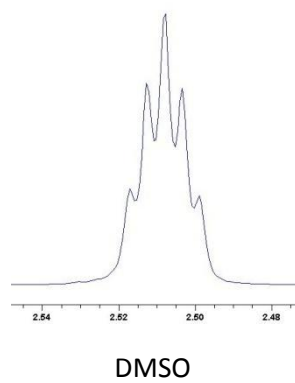
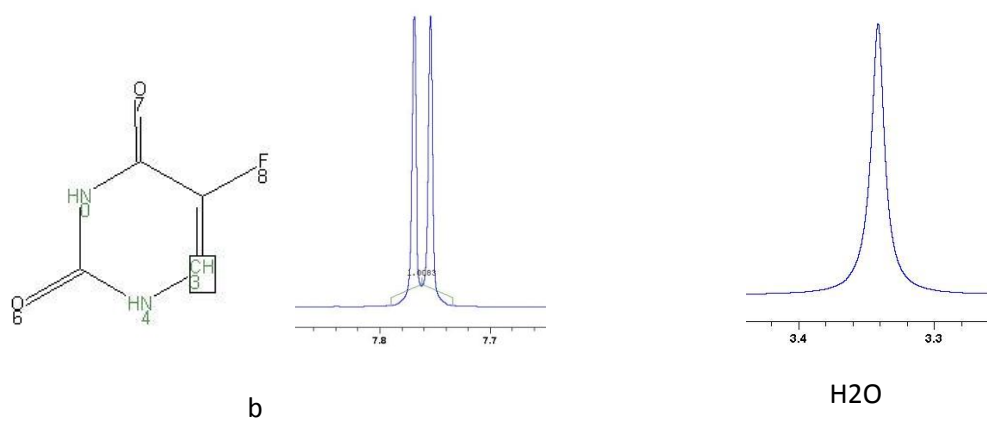
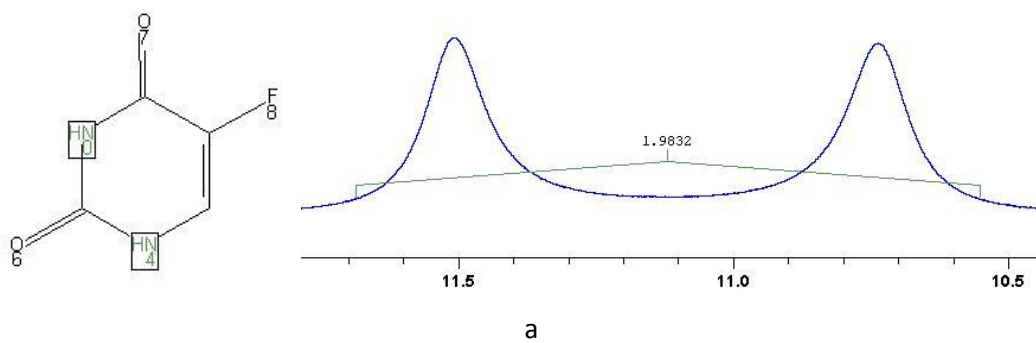
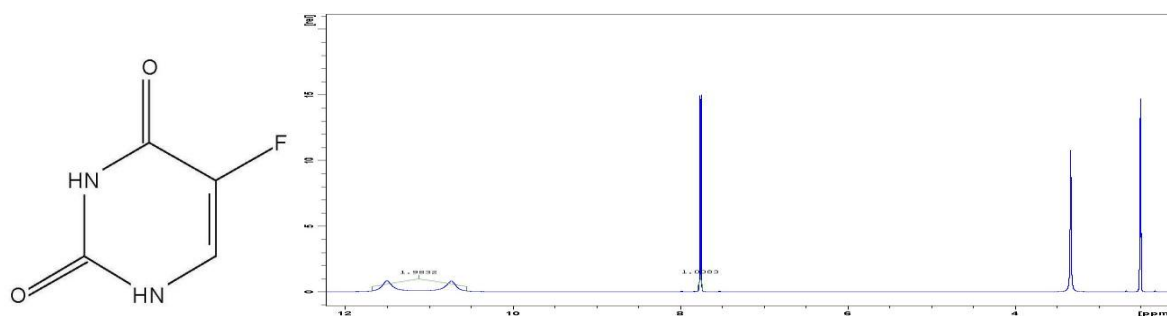


c

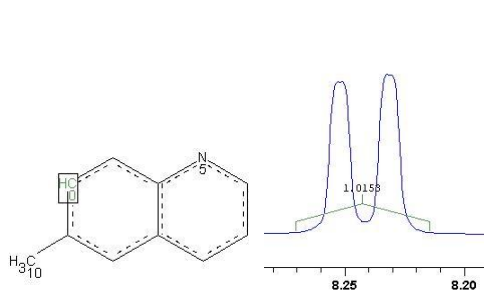
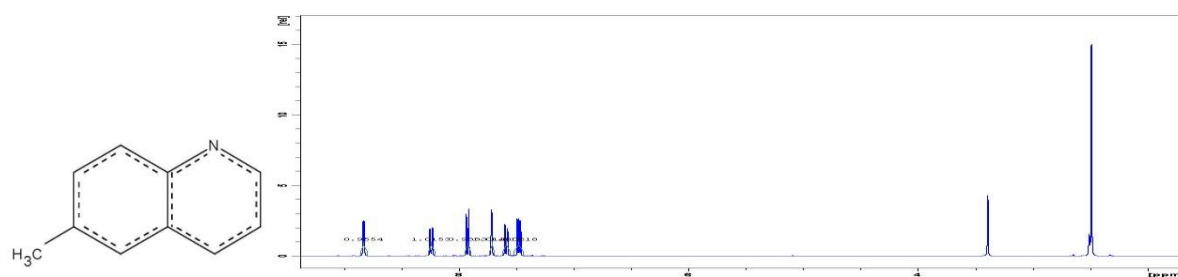
H<sub>2</sub>O

DMSO

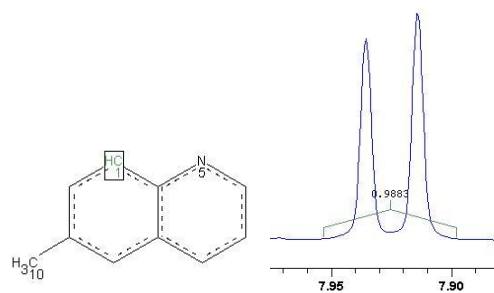
## 23. 5-Fluorouracil



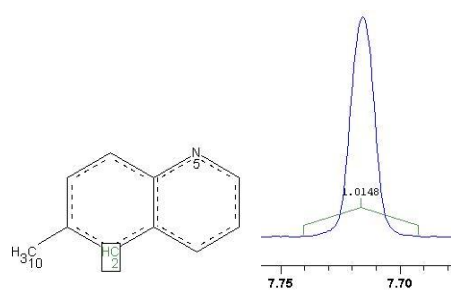
## 24. 6-Methyl-chinolin



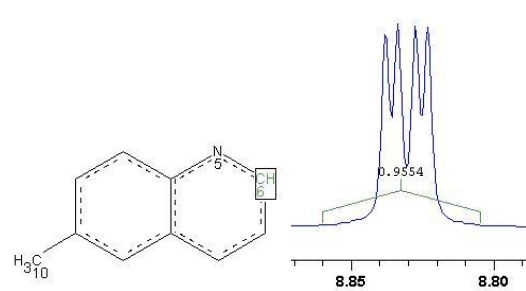
a



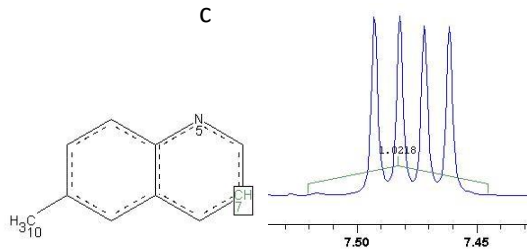
b



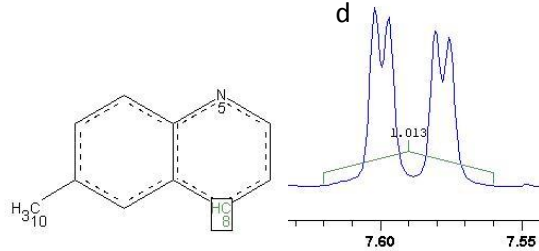
c



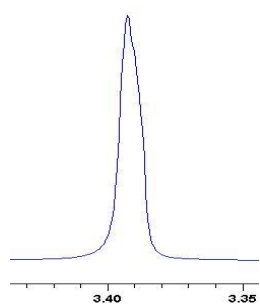
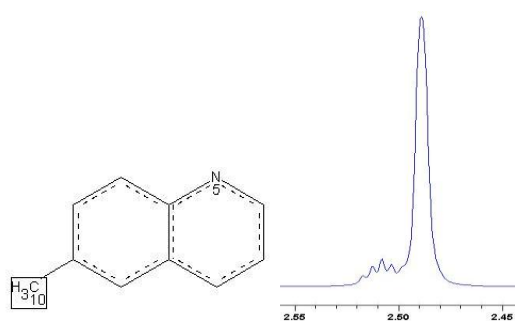
d



e



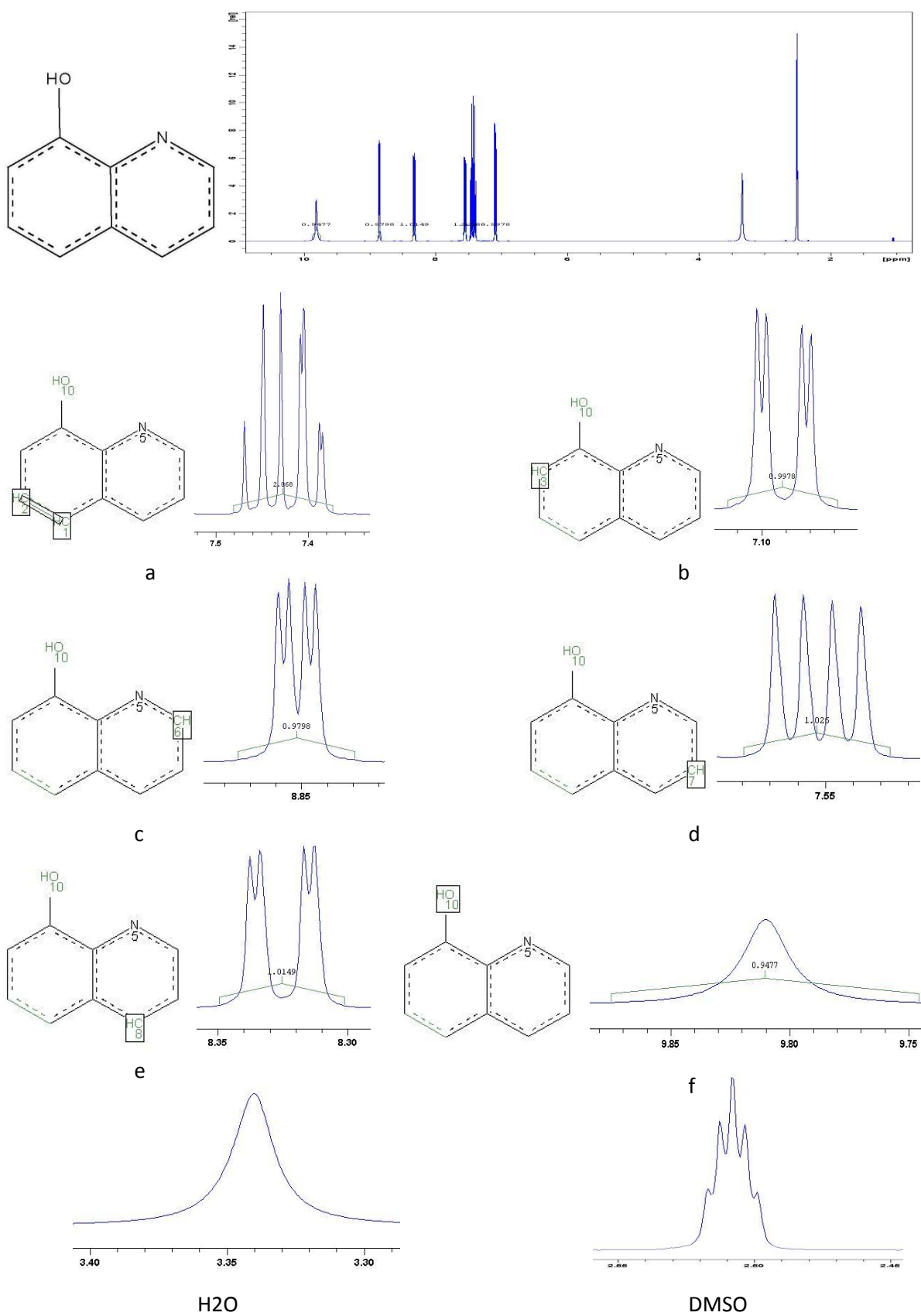
f

H<sub>2</sub>O

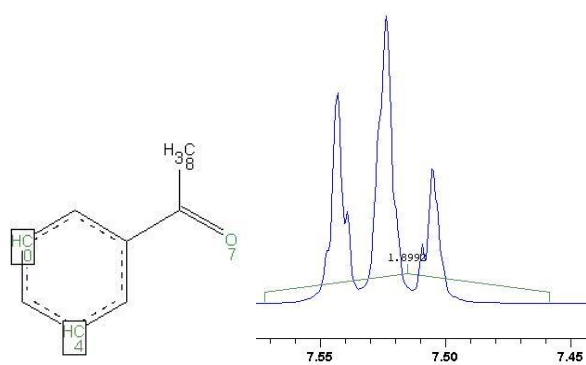
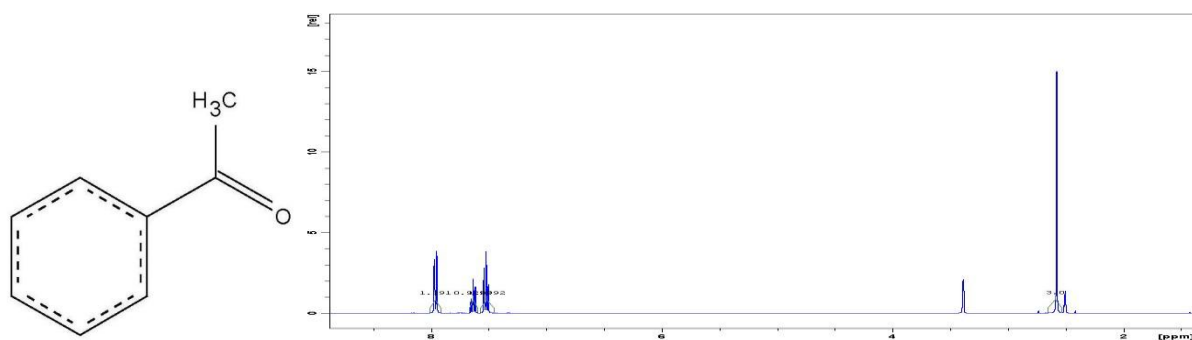
g and DMSO



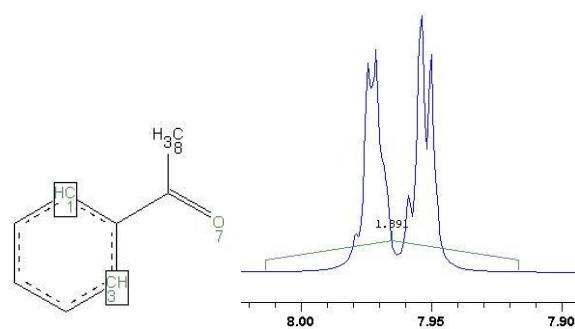
## 25. 8-Hydroxy-chinolin



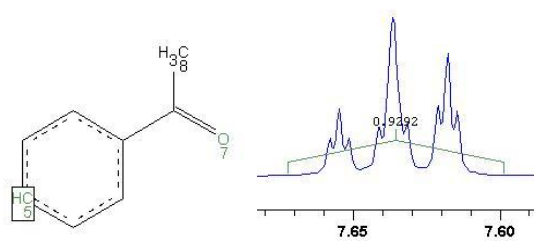
## 26. Acetophenon



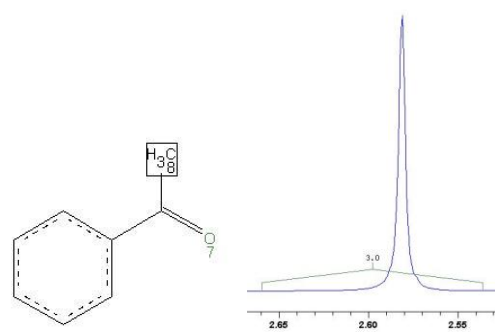
a



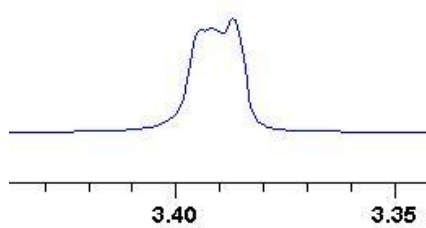
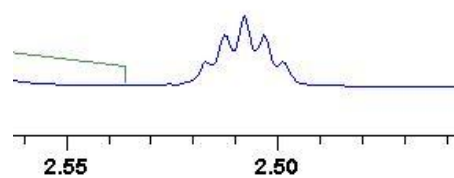
b



c

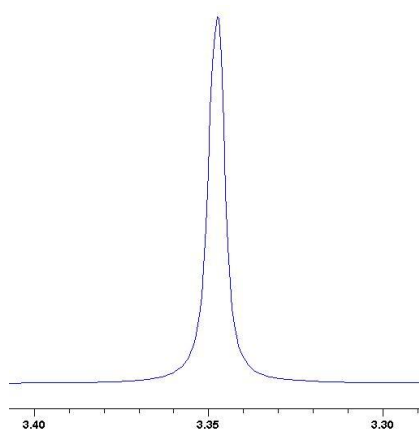
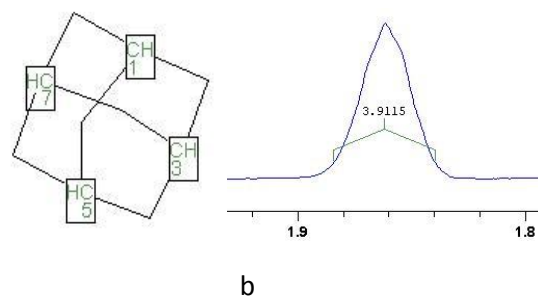
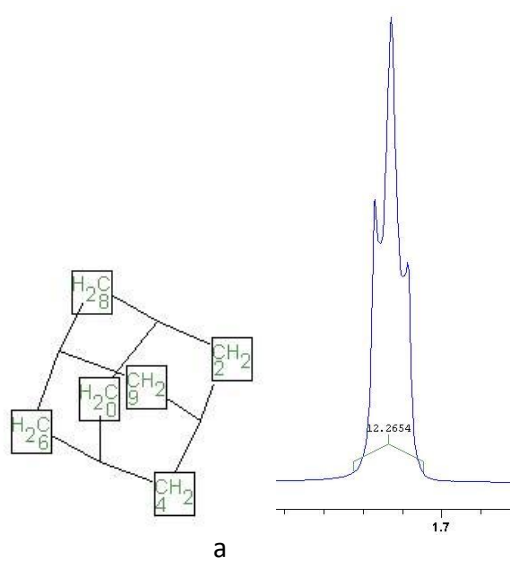
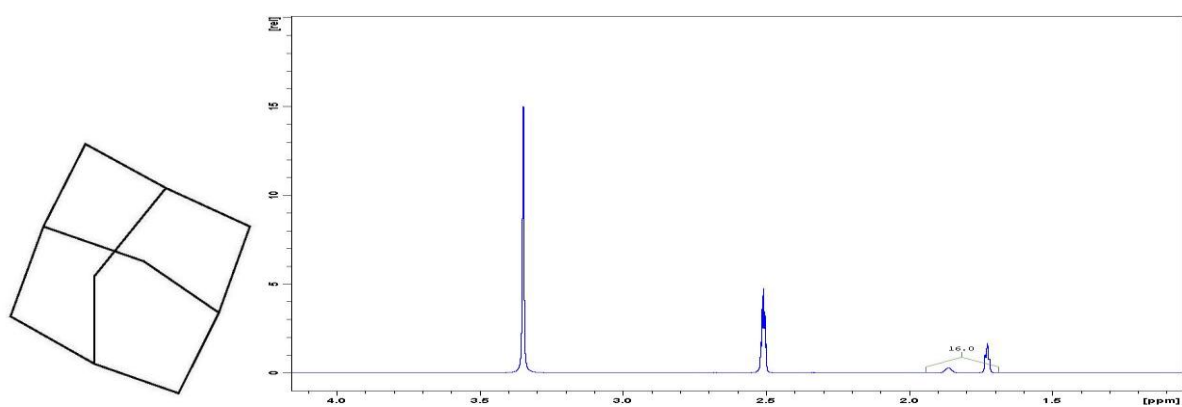
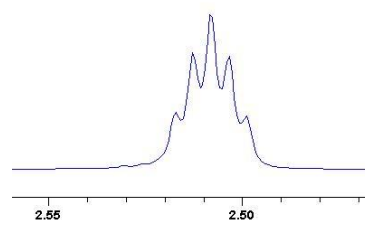


d

 $\text{H}_2\text{O}$ 

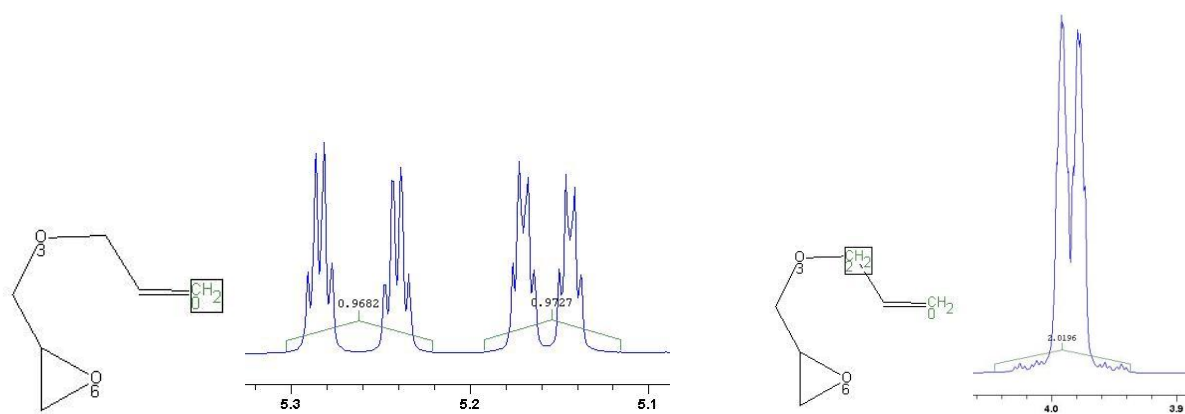
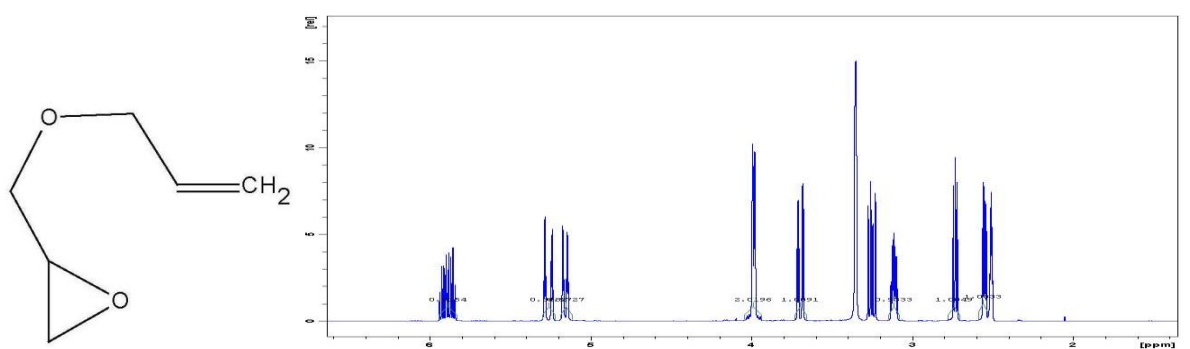
DMSO

## 27. Adamantan

H<sub>2</sub>O

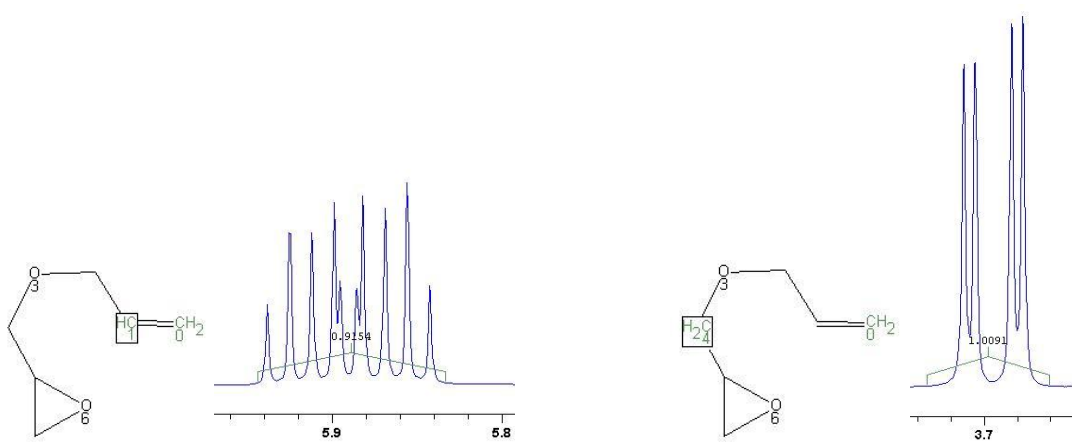
DMSO

## 28. Allylglycidether



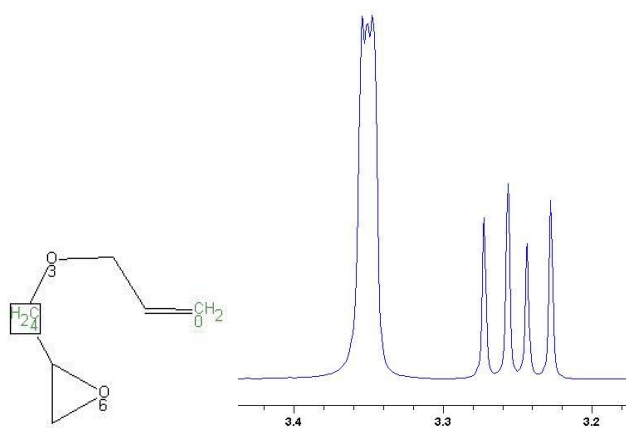
a

b

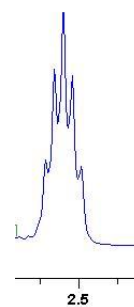


c

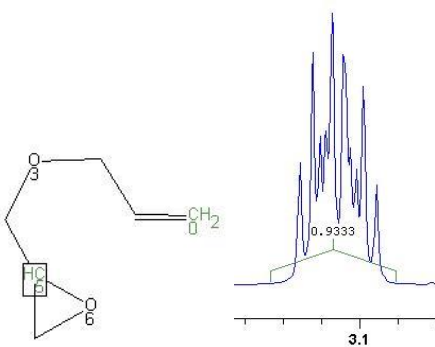
d



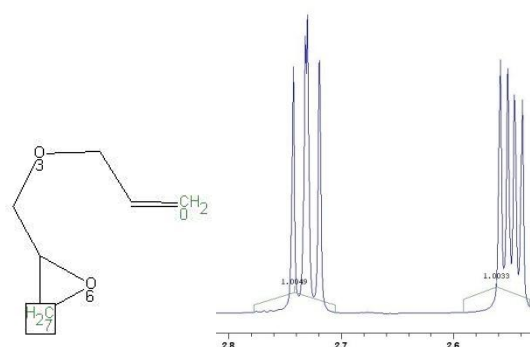
**e** and H<sub>2</sub>O



DMSO

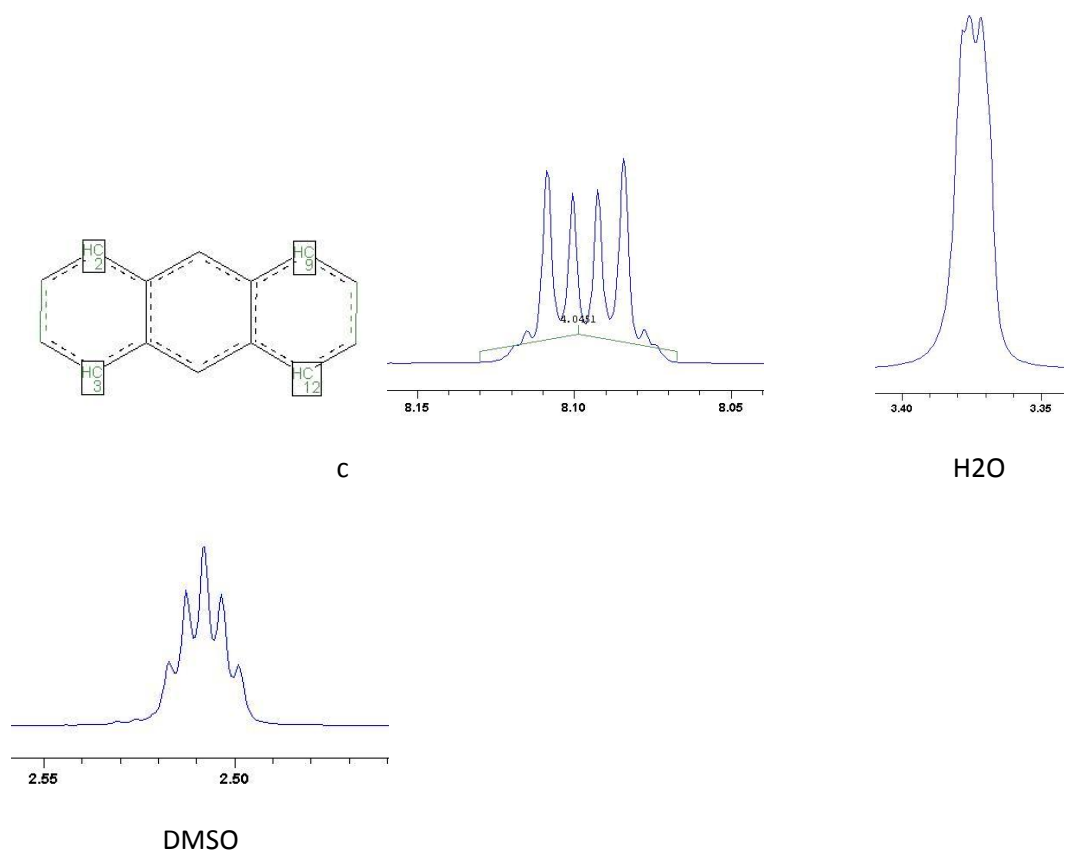
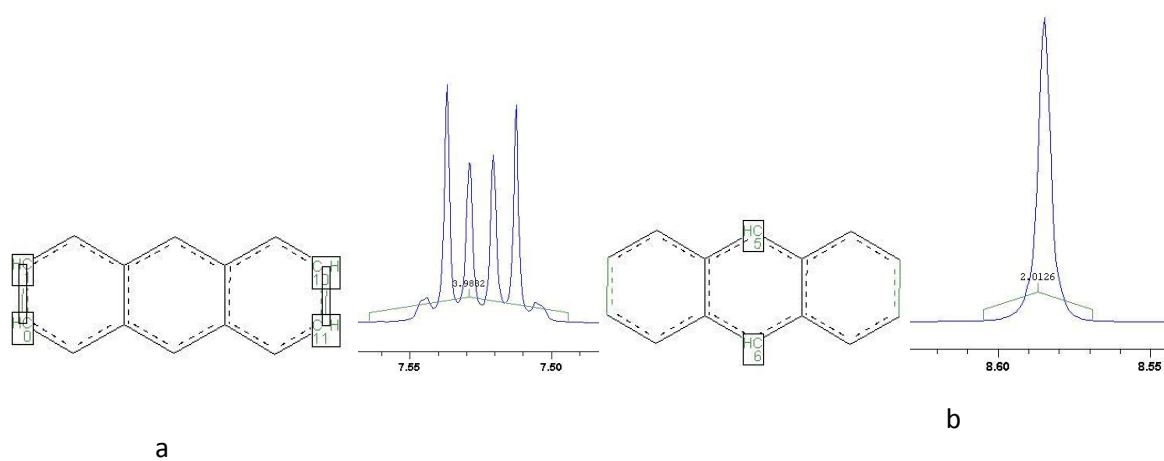
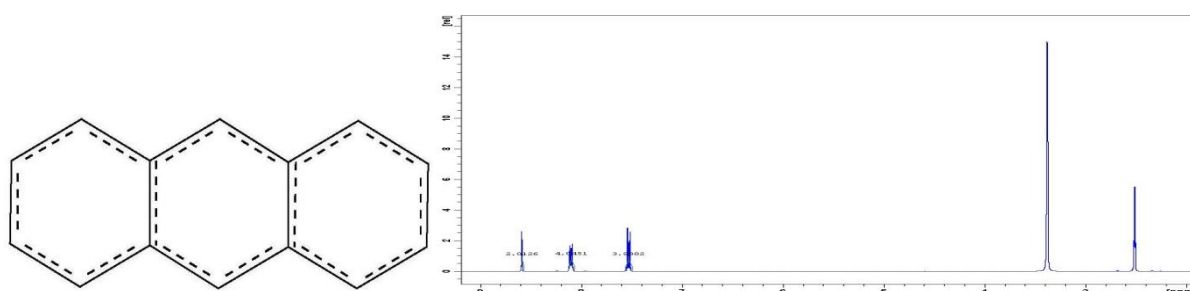


**f**

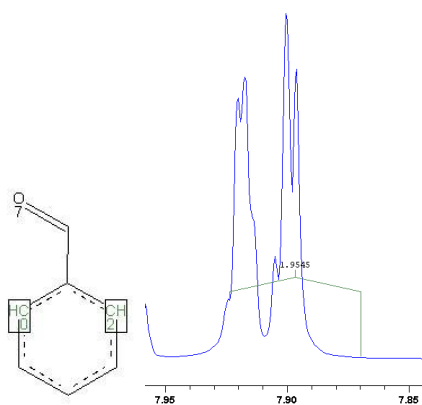
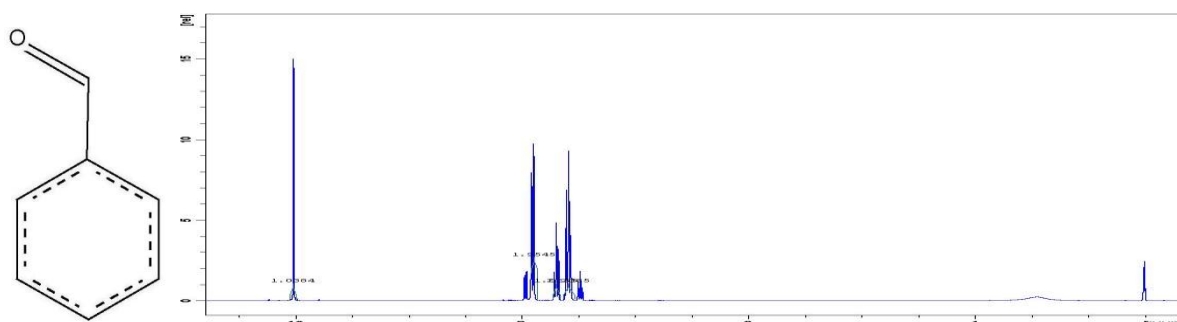


**g**

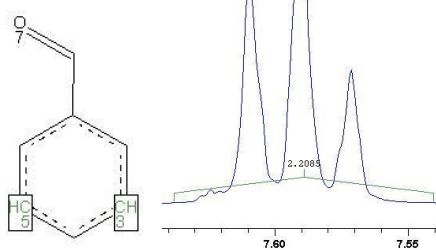
## 29. Anthracen



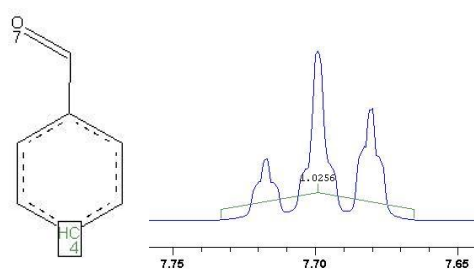
## 30. Benzaldehyd



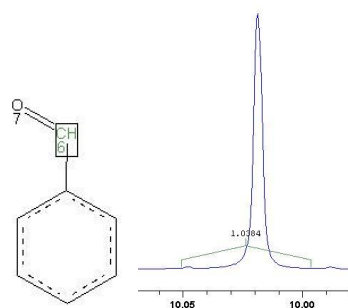
a



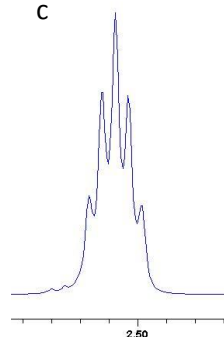
b



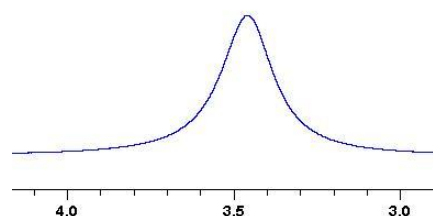
c



d

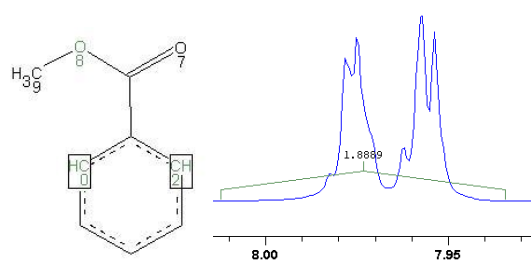
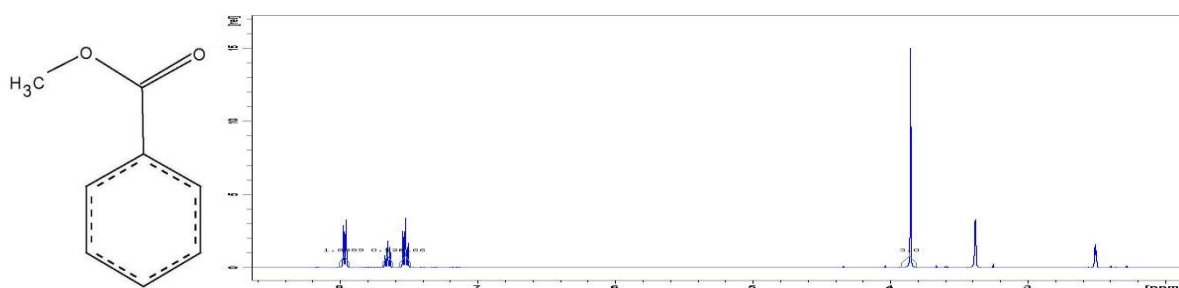


DMSO

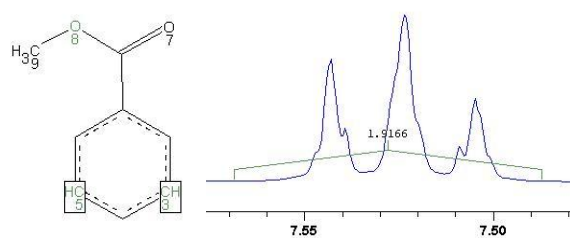


H2O

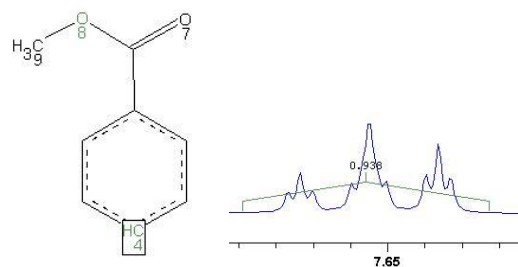
## 31. Benzoesauremethylester



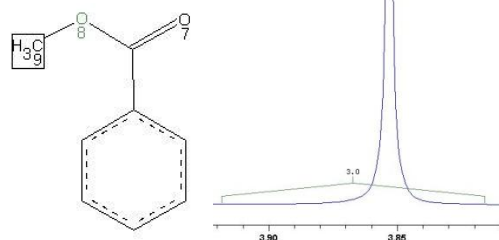
a



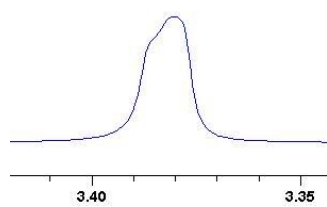
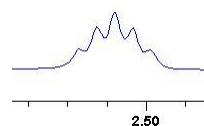
b



c



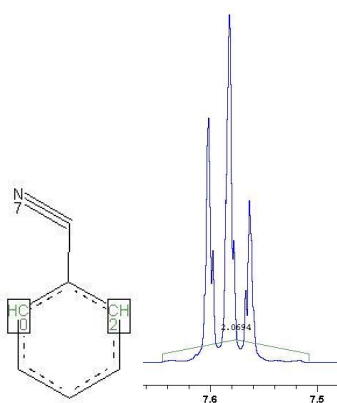
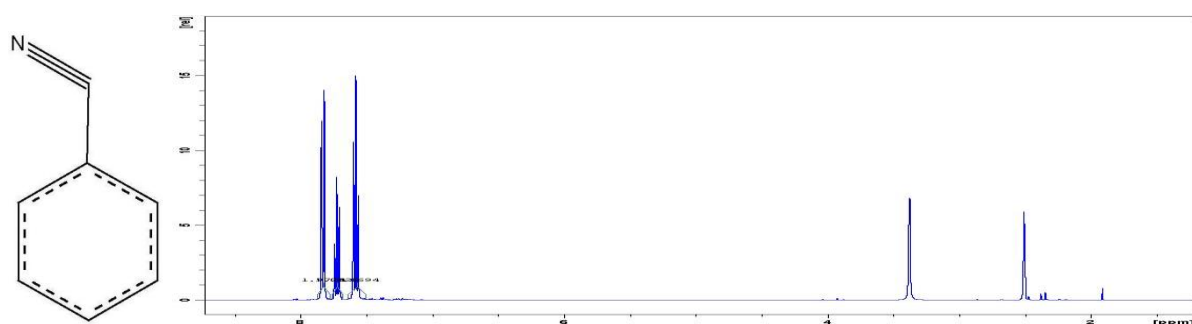
d

 $\text{H}_2\text{O}$ 

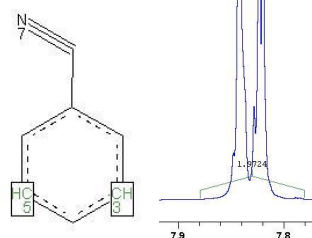
DMSO



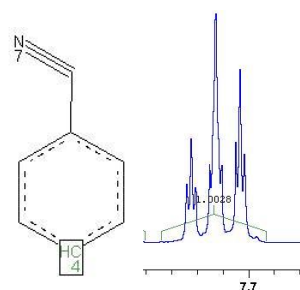
## 32. Benzonitril



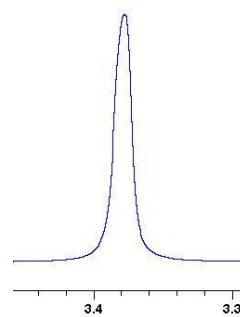
a



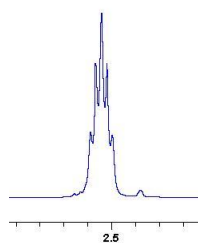
b



c

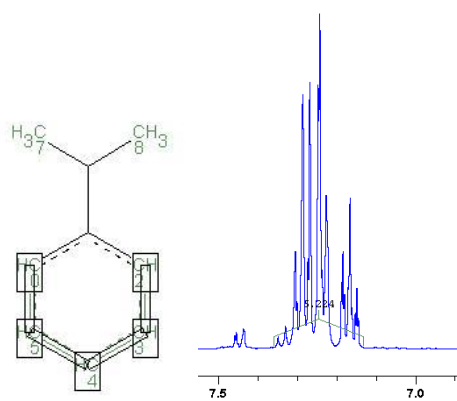
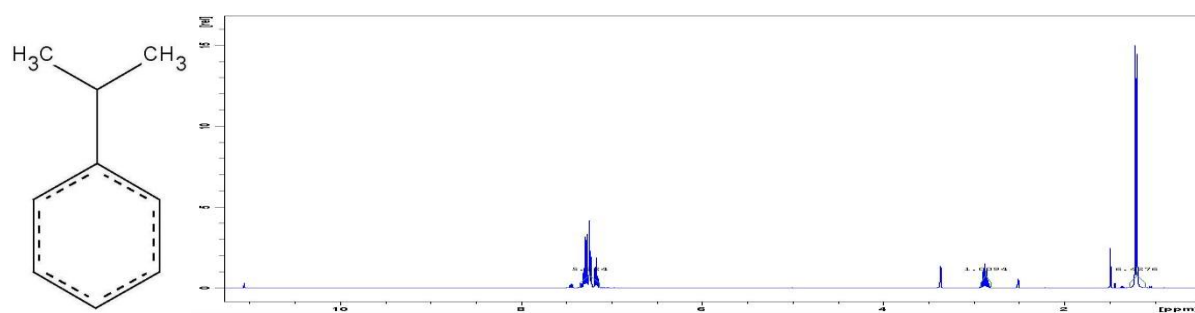


H2O

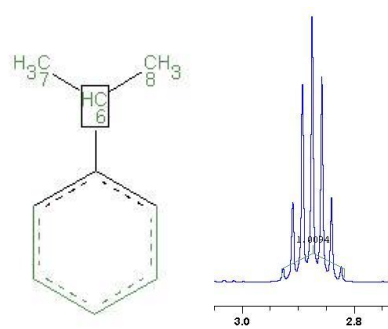


DMSO

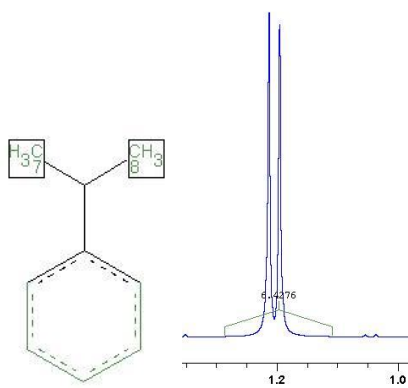
## 33. Cumol



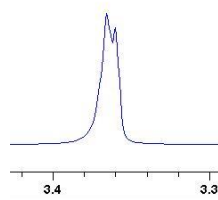
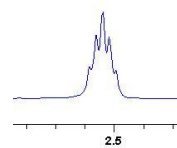
a



b

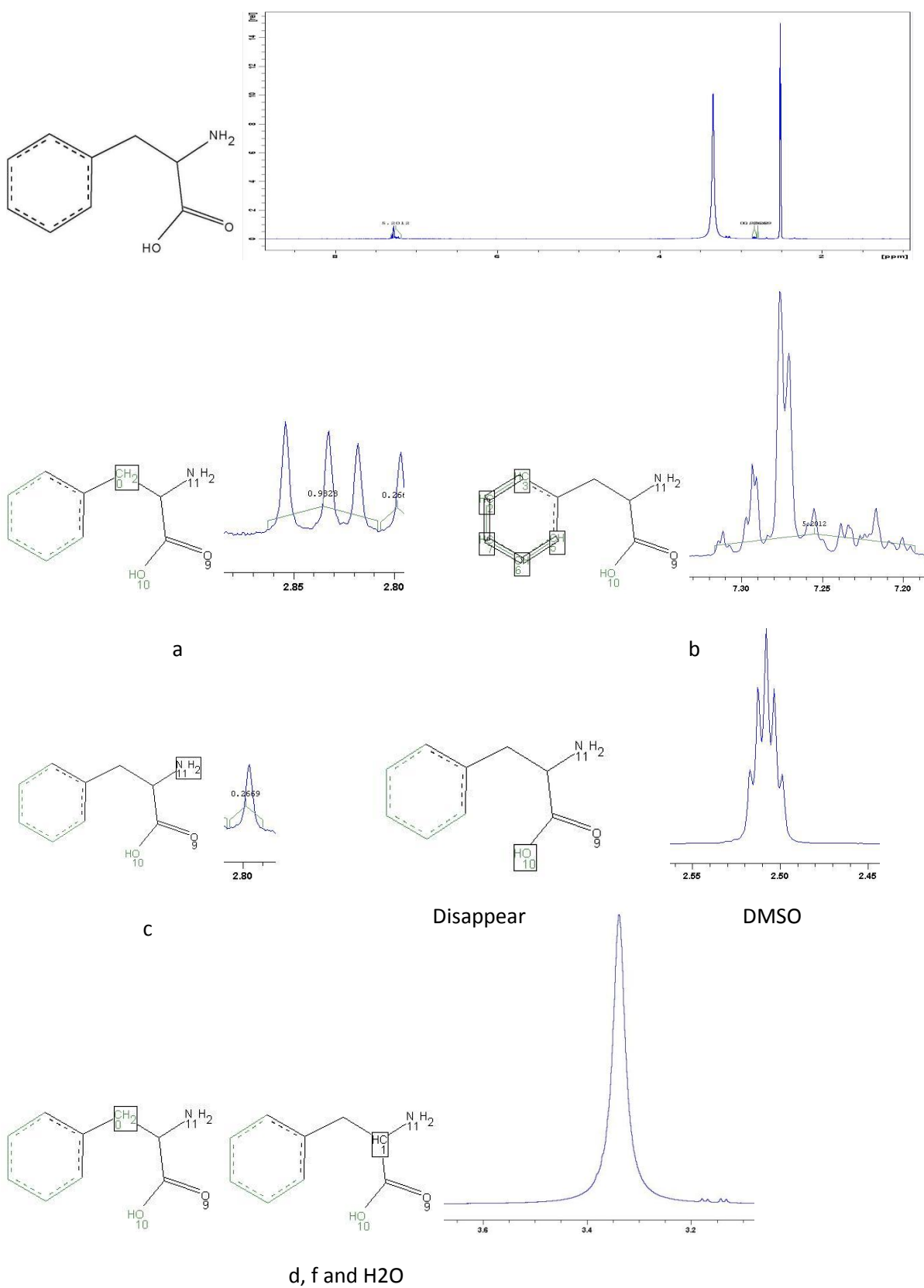


c

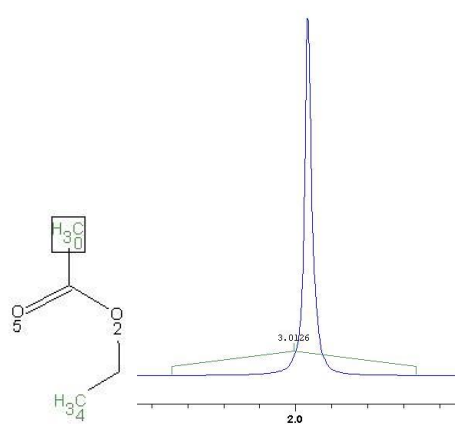
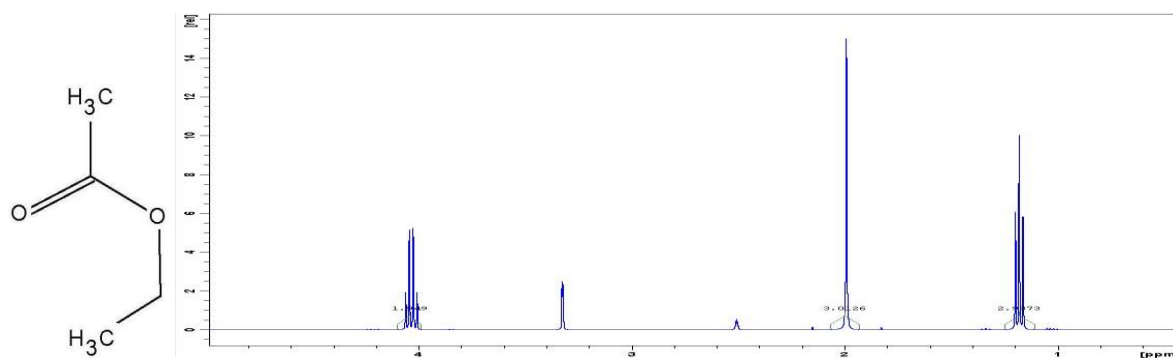
H<sub>2</sub>O

DMSO

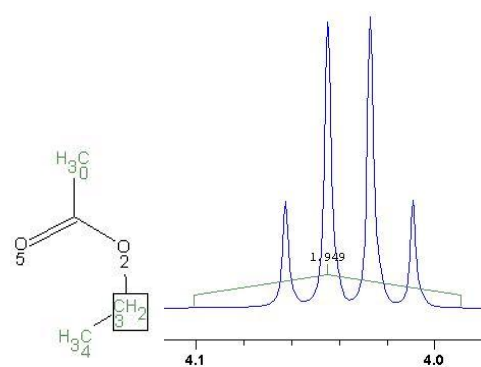
## 34. D,L-Phenylalanine



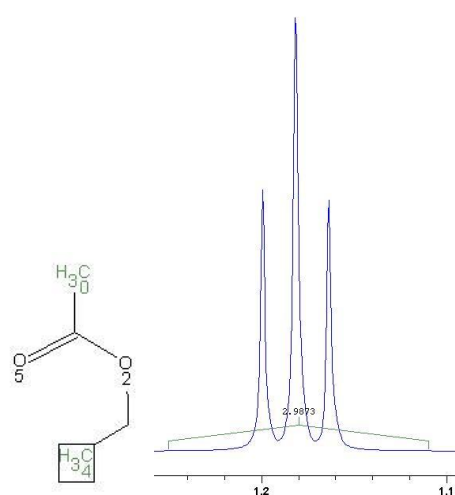
## 35. Essigester



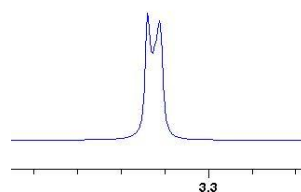
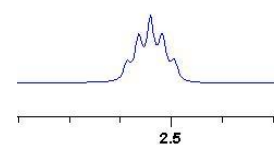
a



b

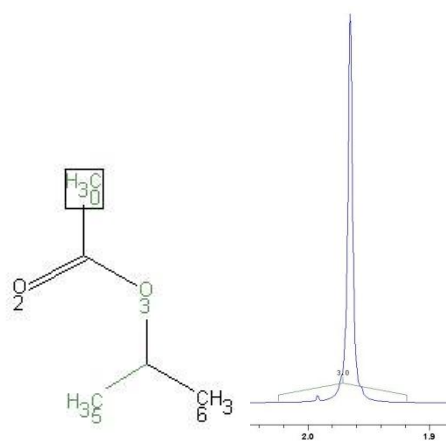
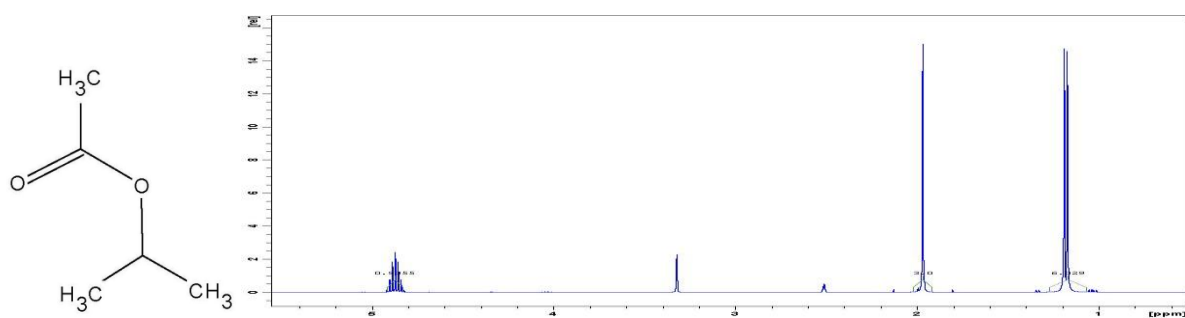


c

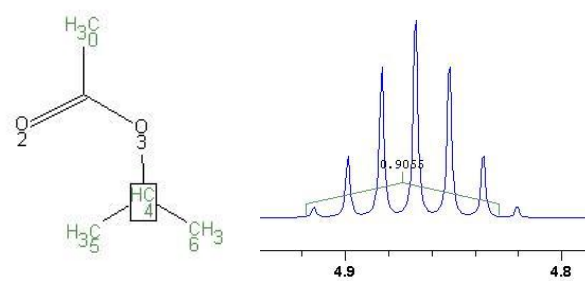
H<sub>2</sub>O

DMSO

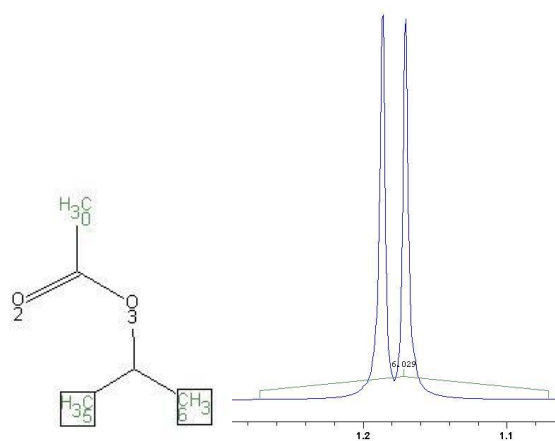
## 36. Essigsaeure-isopropyl-ester



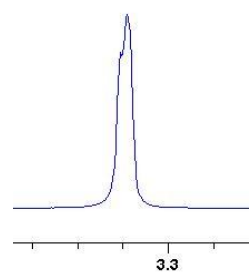
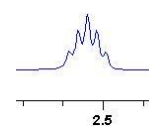
a



b

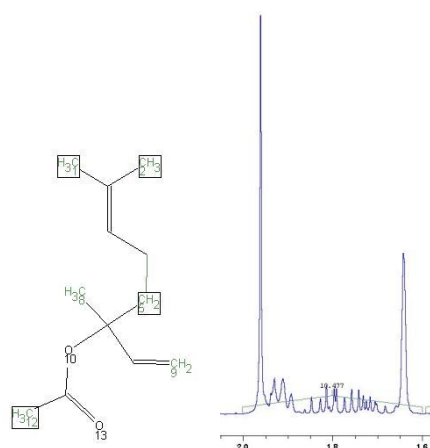
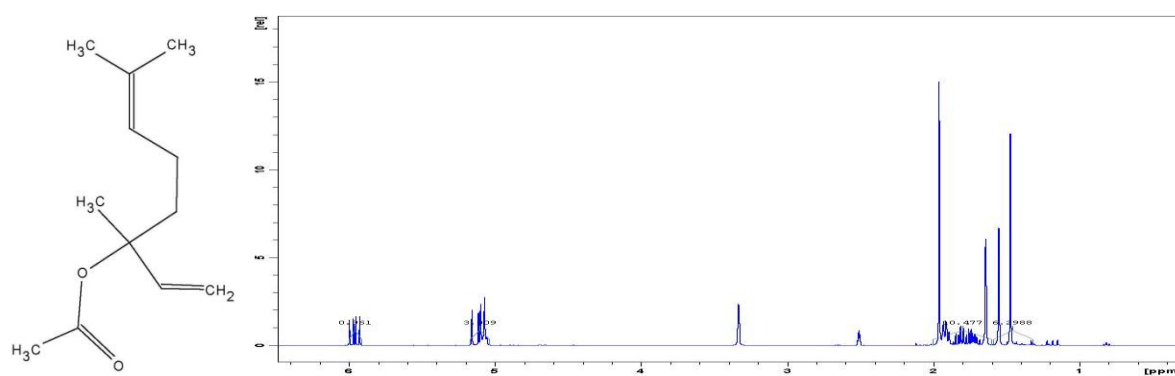


c

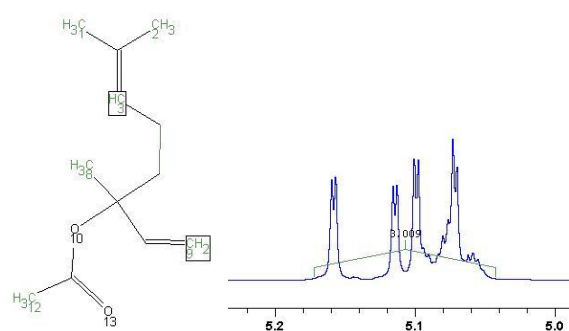
H<sub>2</sub>O

DMSO

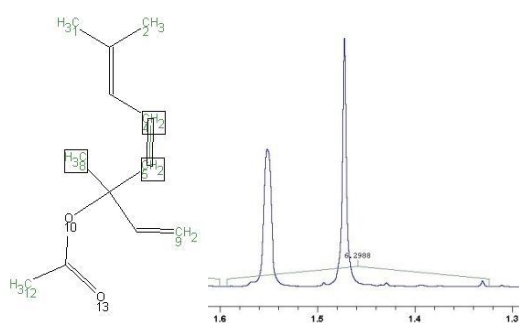
## 37. Essigsaeurelinallylester



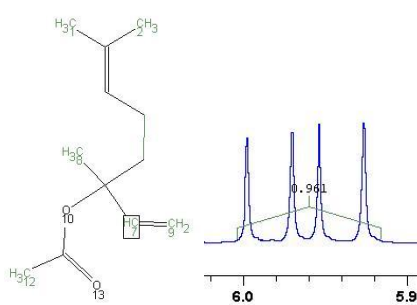
a



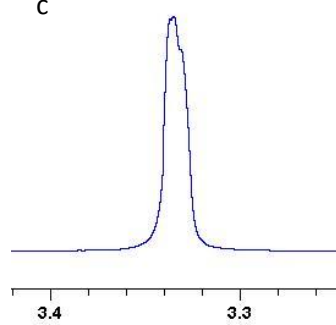
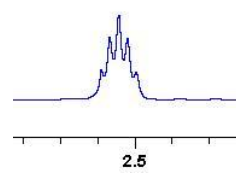
b



c

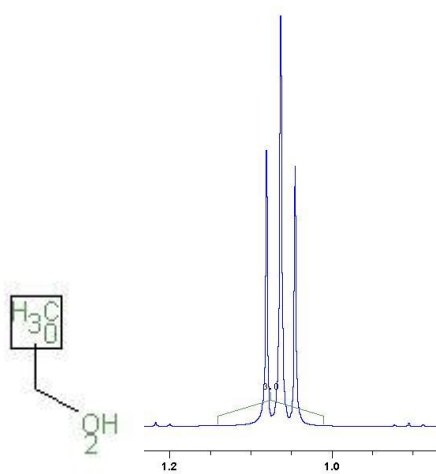
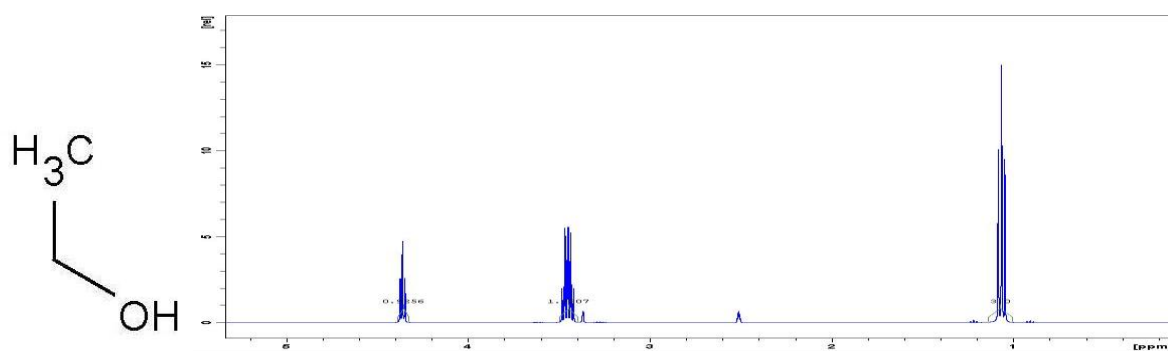


d

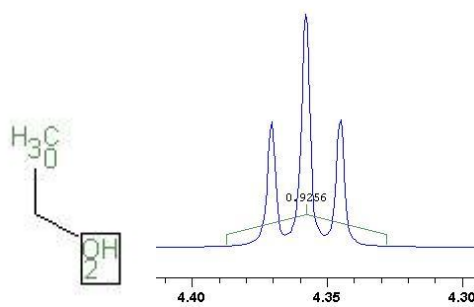
H<sub>2</sub>O

DMSO

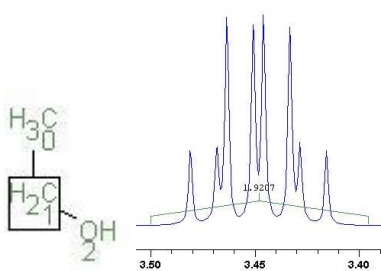
## 38. Ethanol



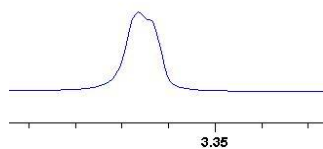
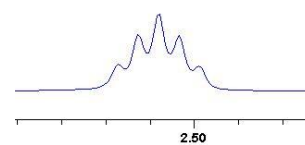
a



b

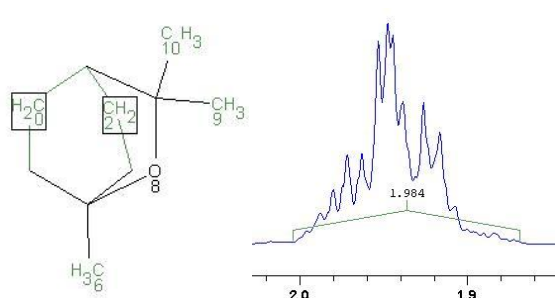
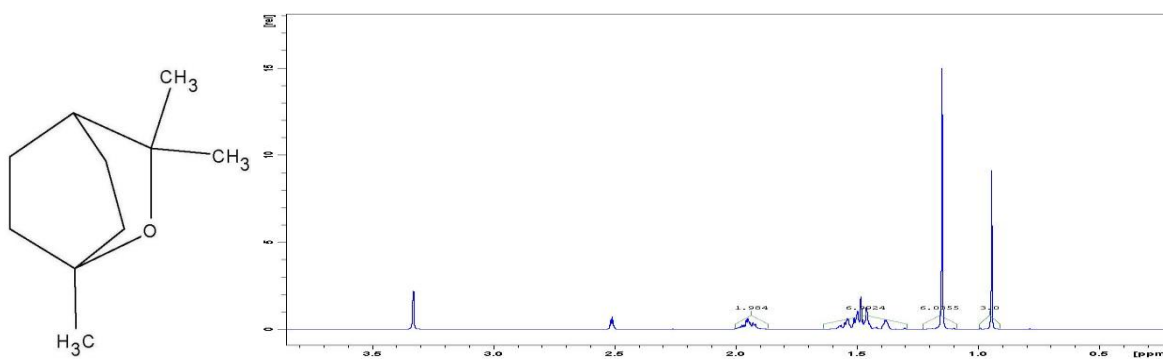


c

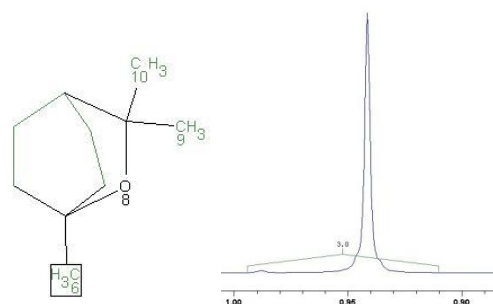
H<sub>2</sub>O

DMSO

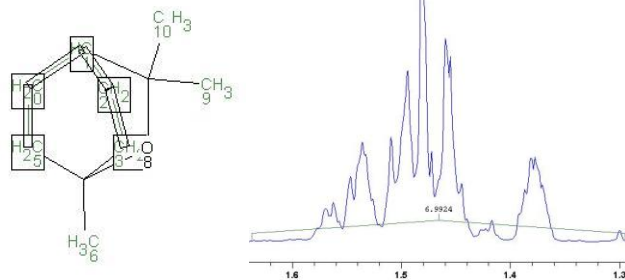
## 39. Eucalyptol



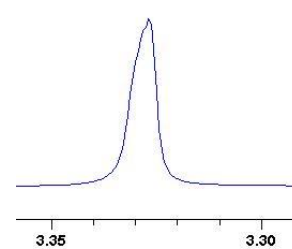
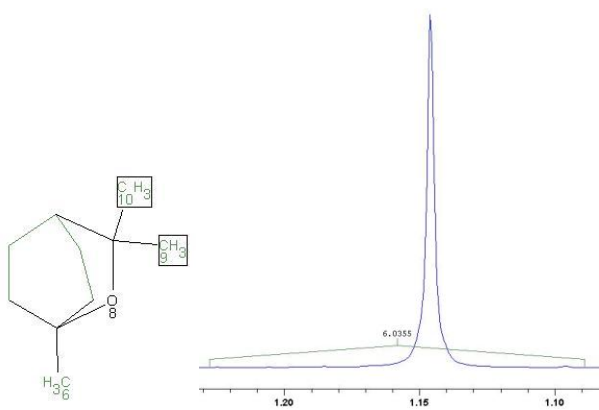
a



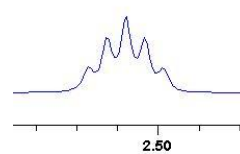
b



c

H<sub>2</sub>O

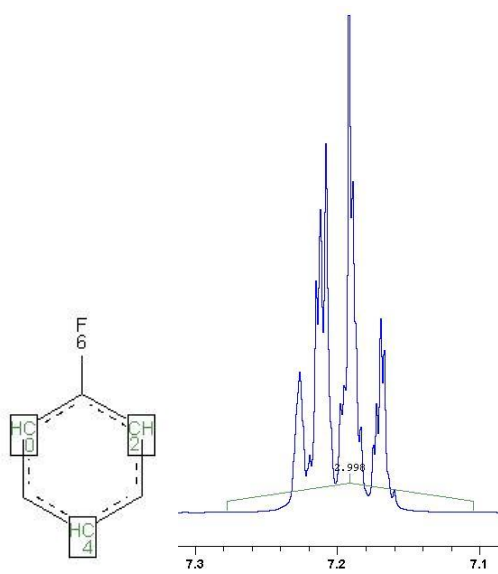
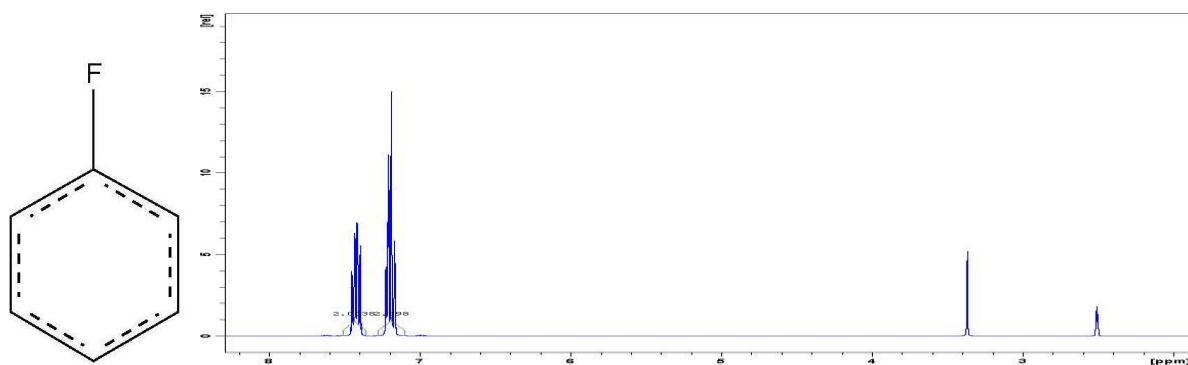
d



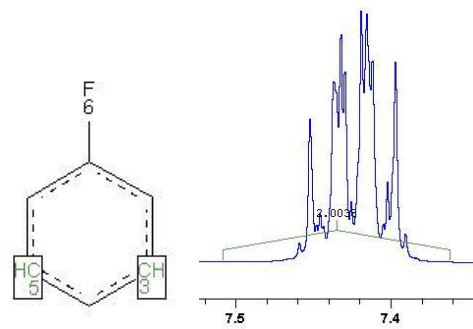
DMSO



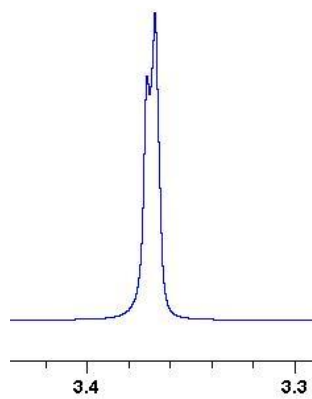
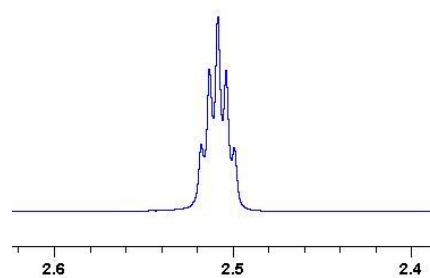
## 40. Fluorbenzol



a

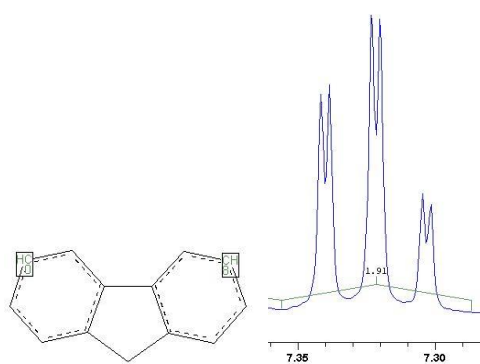
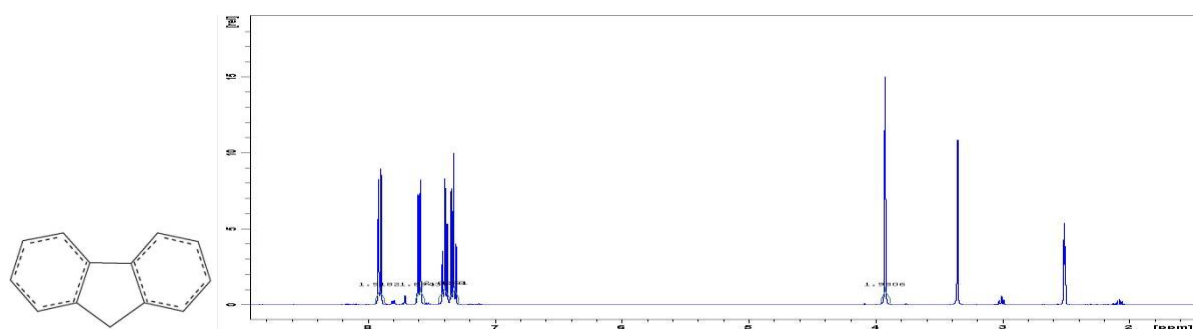


b

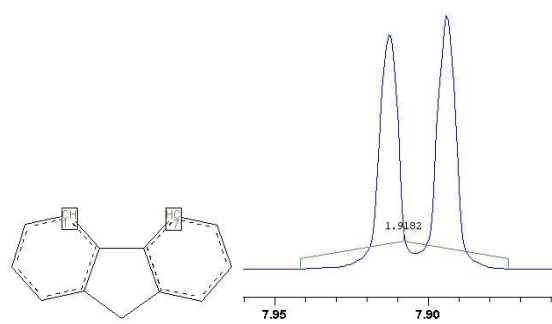
H<sub>2</sub>O

DMSO

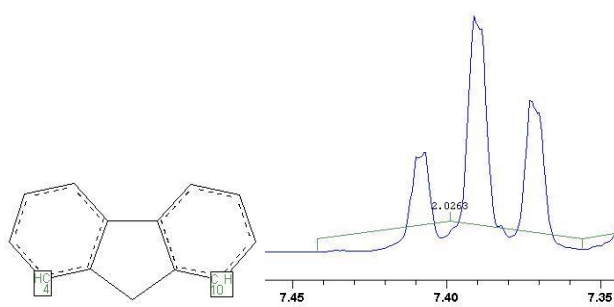
## 41. Fluoren



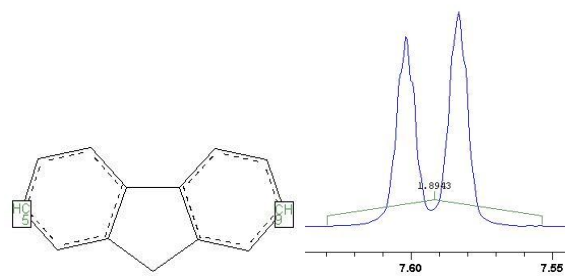
a



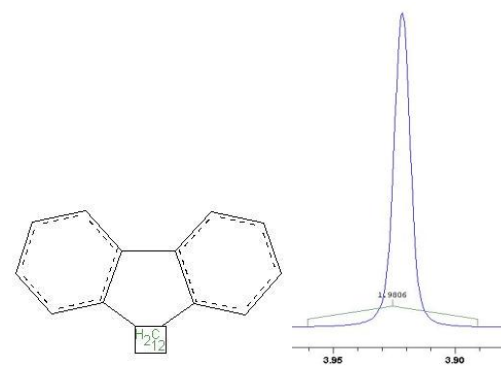
b



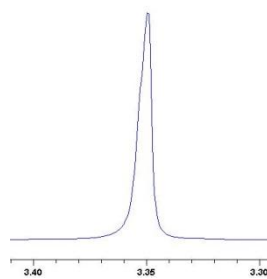
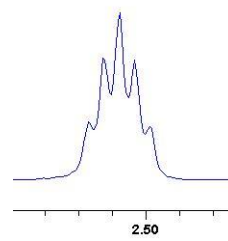
c



d

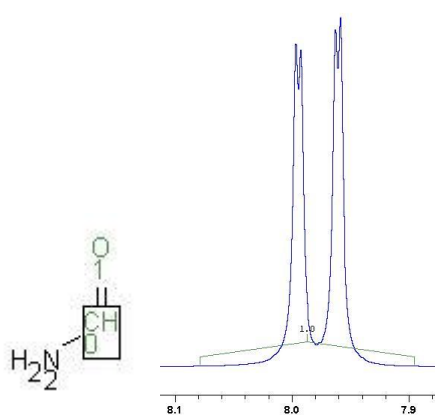
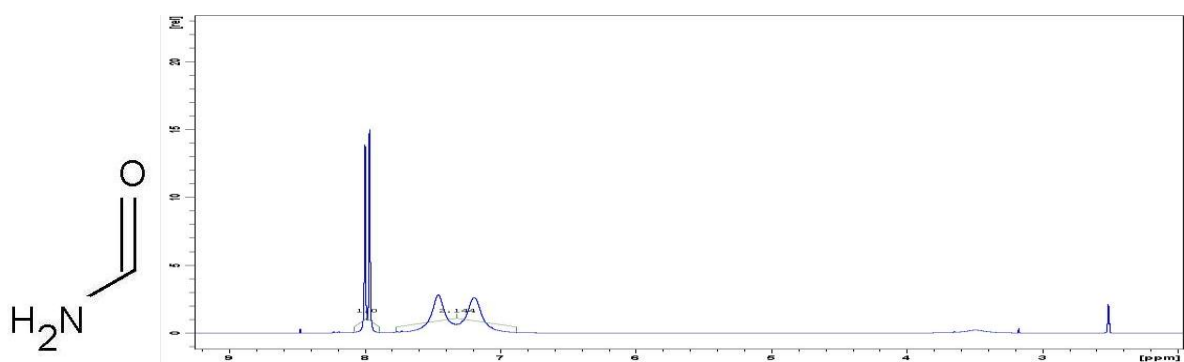


e

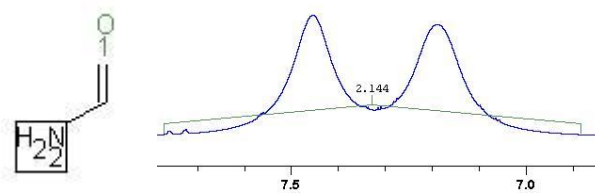
H<sub>2</sub>O

DMSO

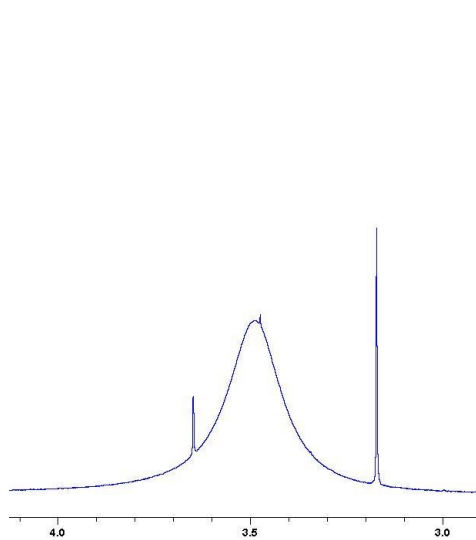
## 42. Formamid



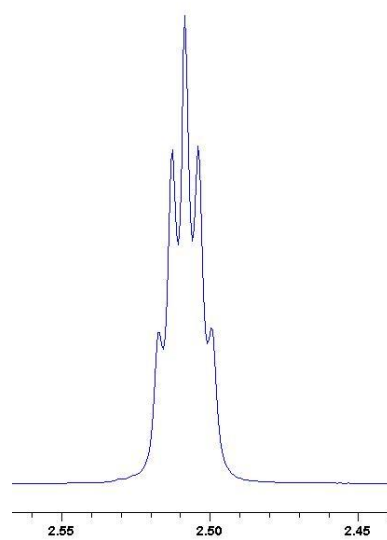
a



b

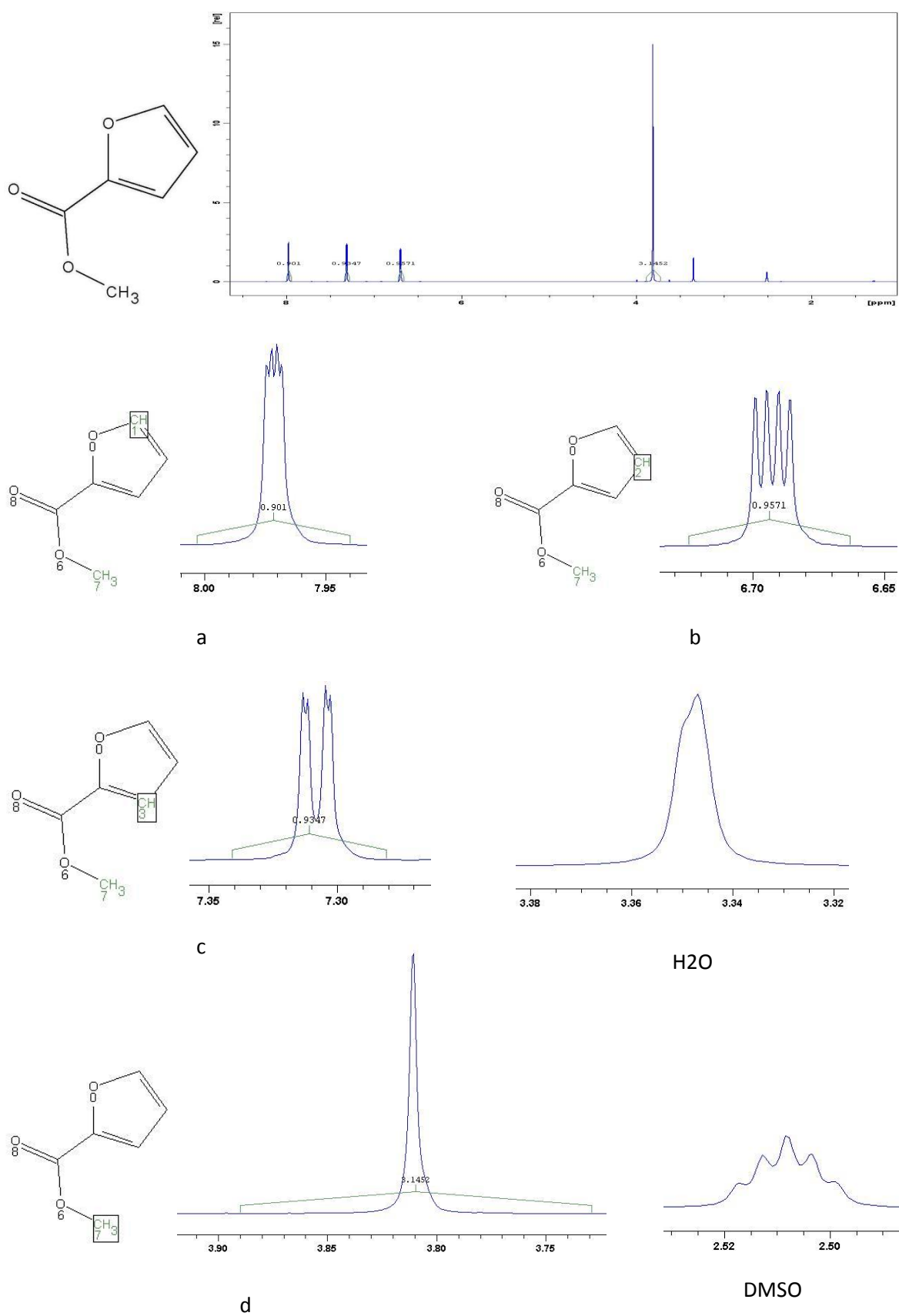


H2O

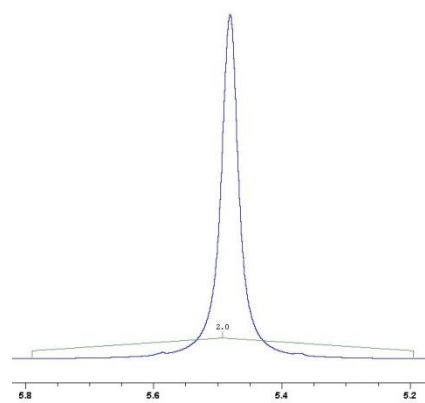
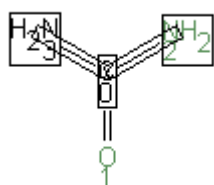
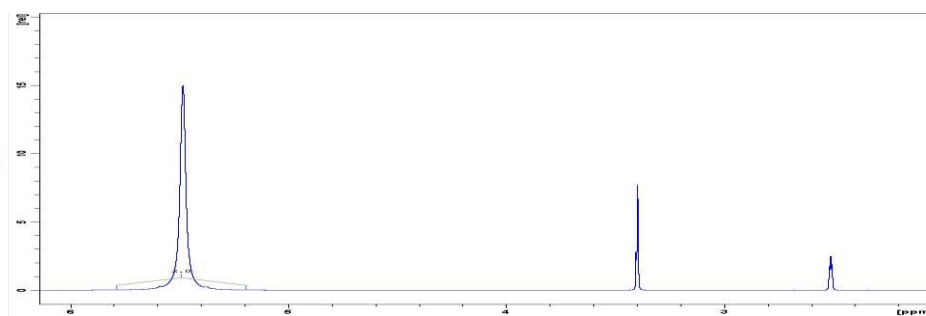
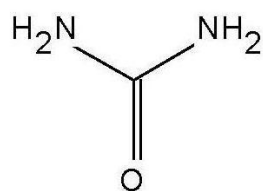


DMSO

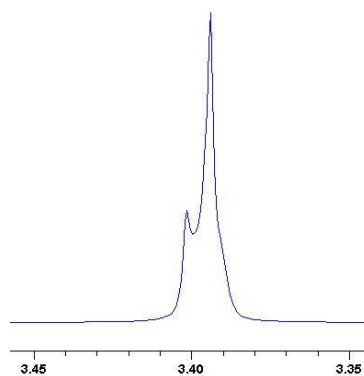
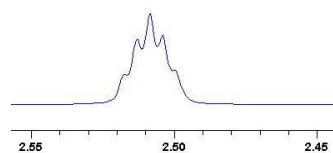
## 43. Furan-2-carbonsaeuremethylest



## 44. Harnstoff

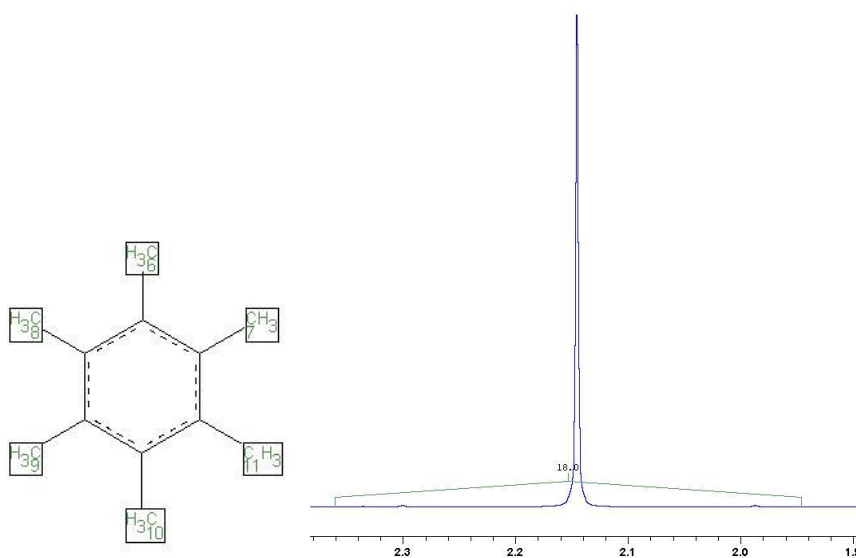
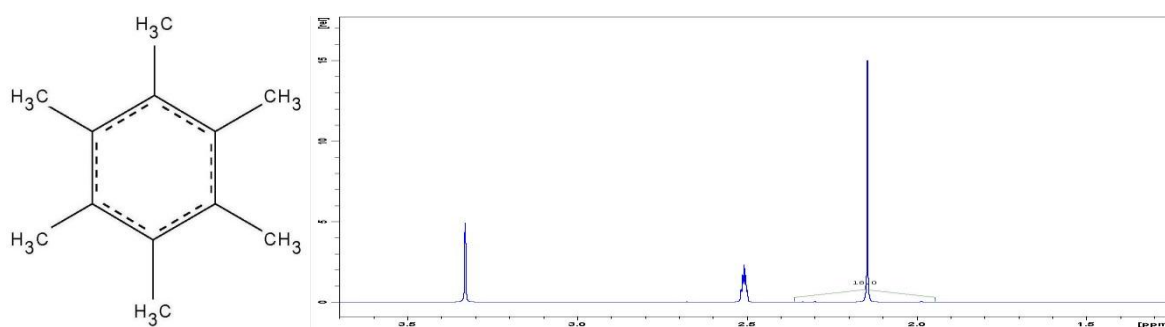


a

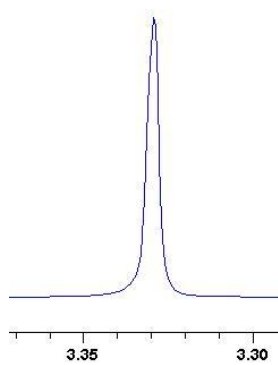
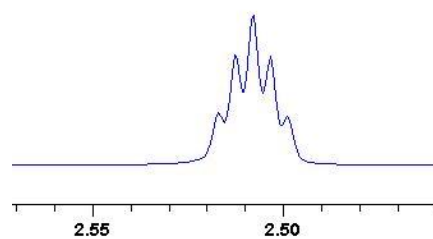
H<sub>2</sub>O

DMSO

## 45. Hexamethylbenzol

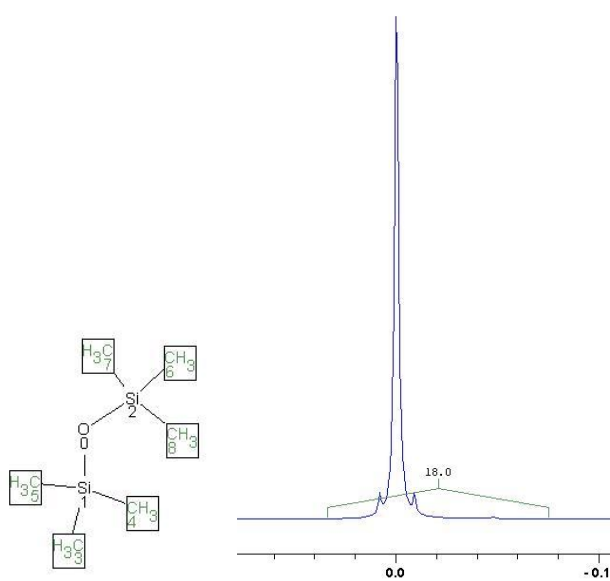
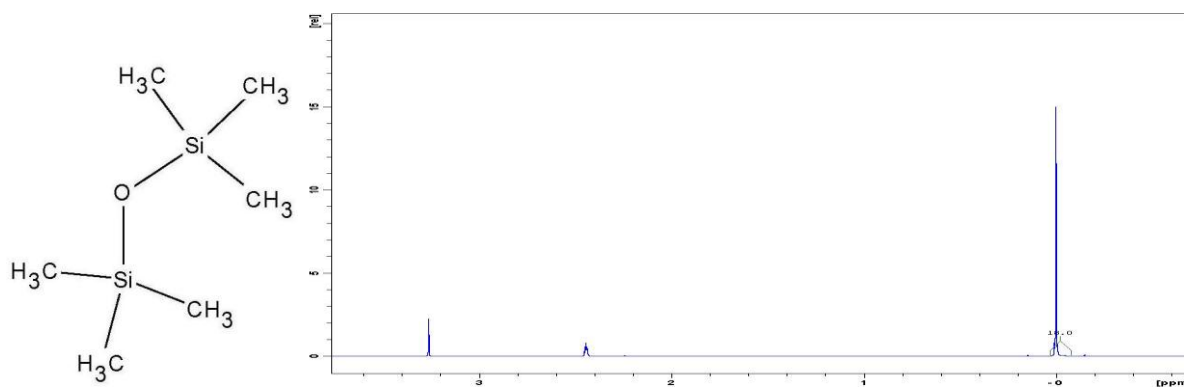


a

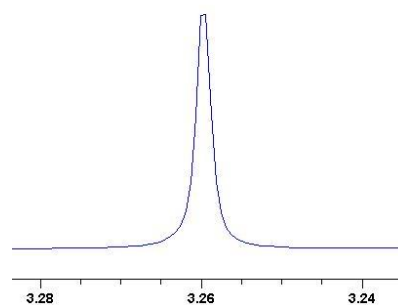
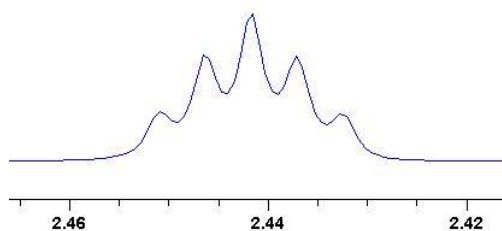
H<sub>2</sub>O

DMSO

## 46. Hexamethyldisiloxan

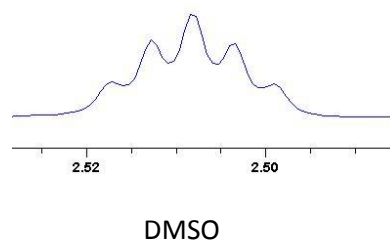
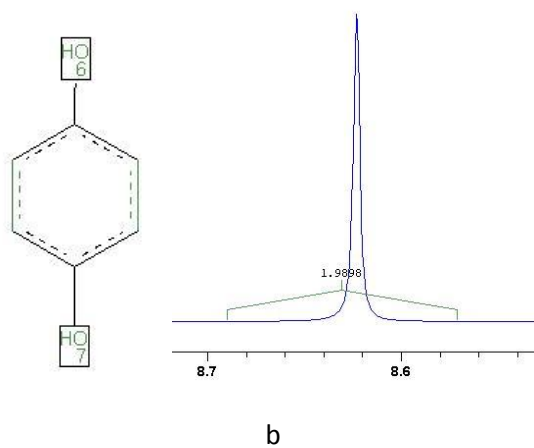
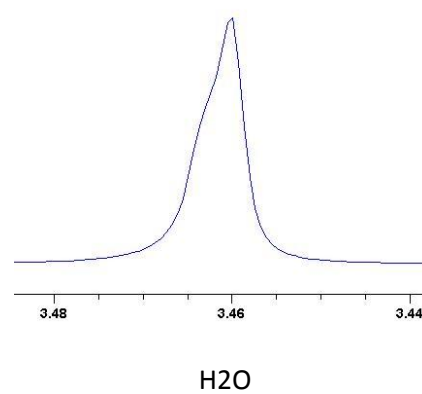
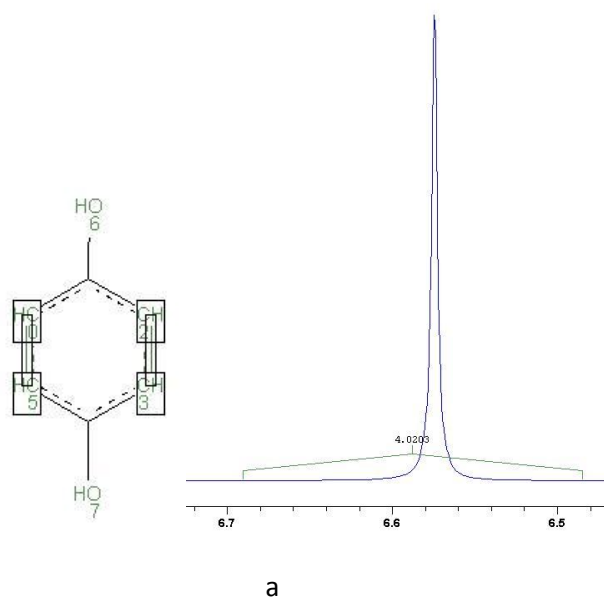
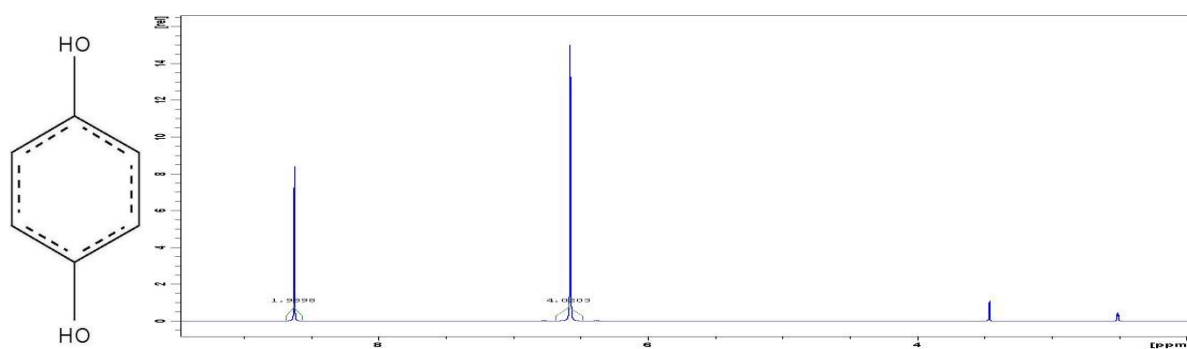


a

H<sub>2</sub>O

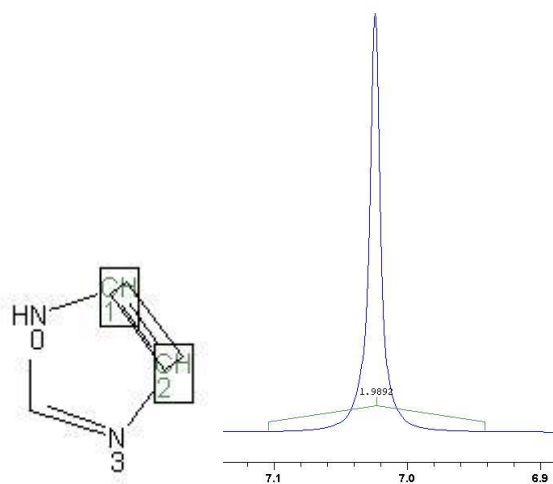
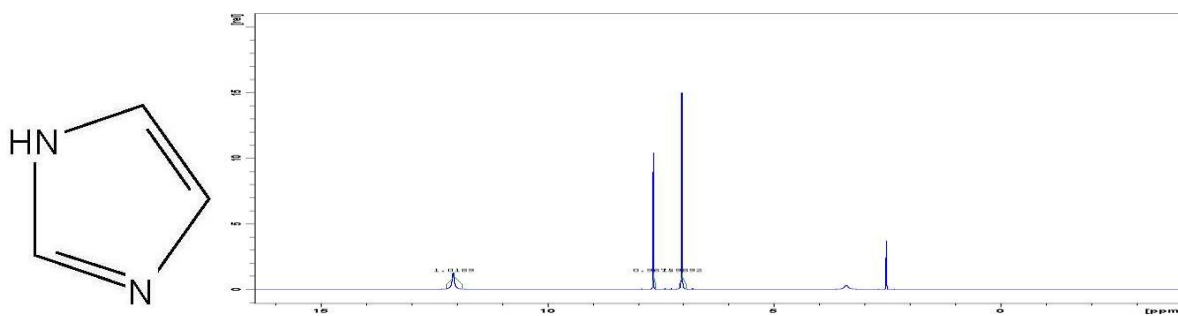
DMSO

## 47. Hydrochinon

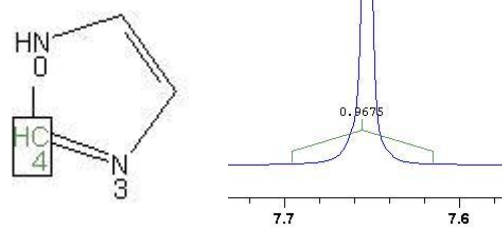




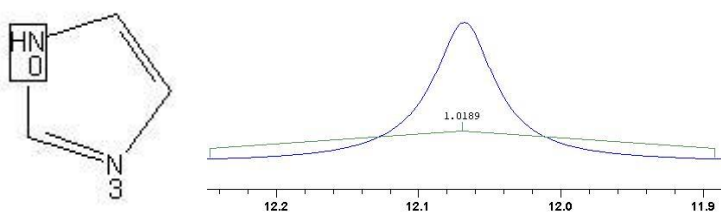
## 48. Imidazol



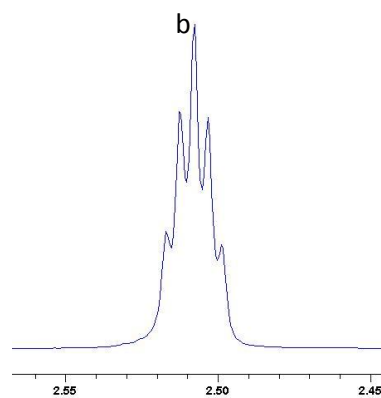
a



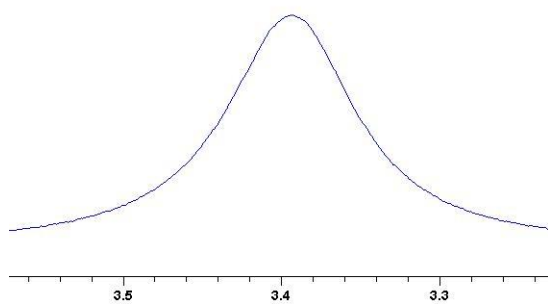
b



c

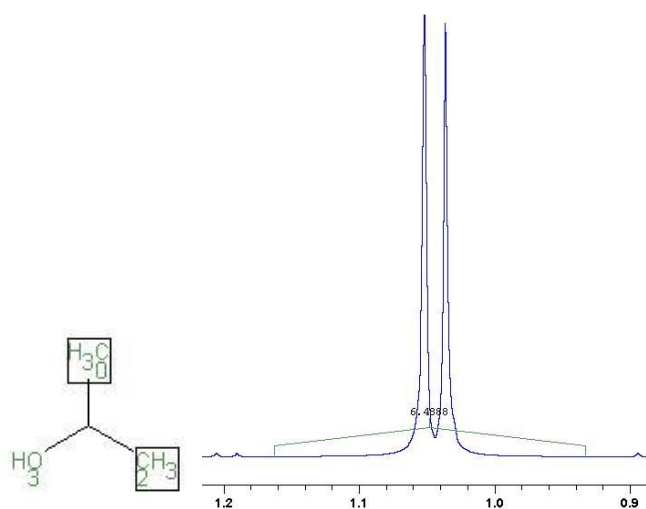
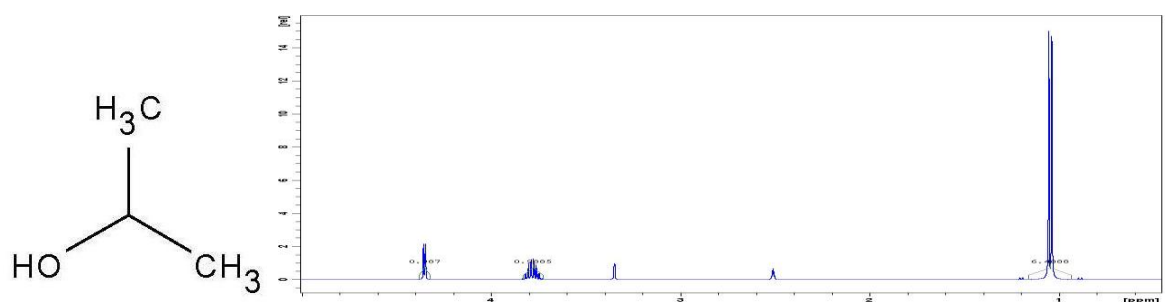


DMSO

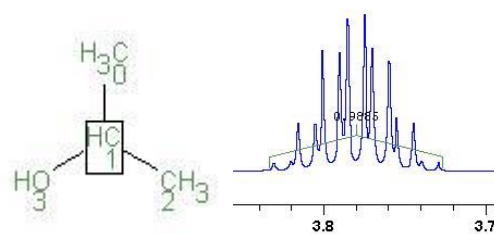


H2O

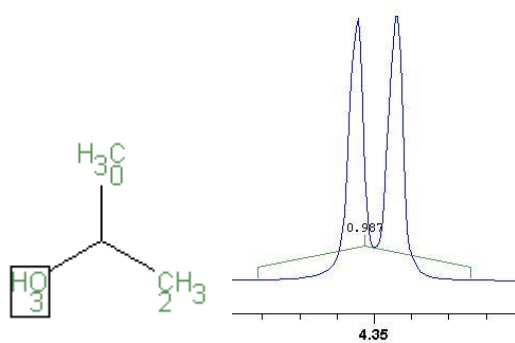
## 49. Isopropanol



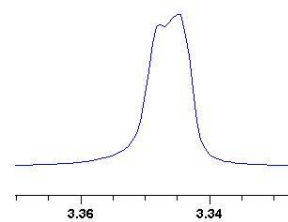
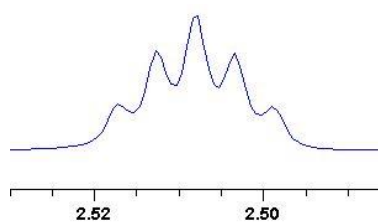
a



b

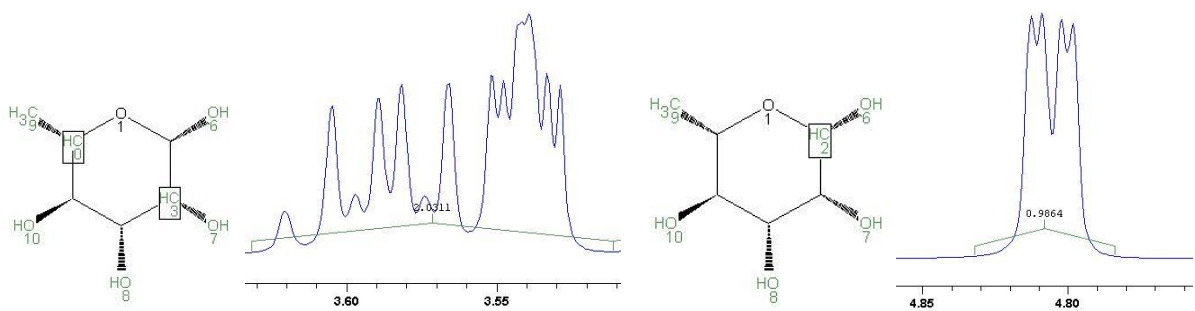
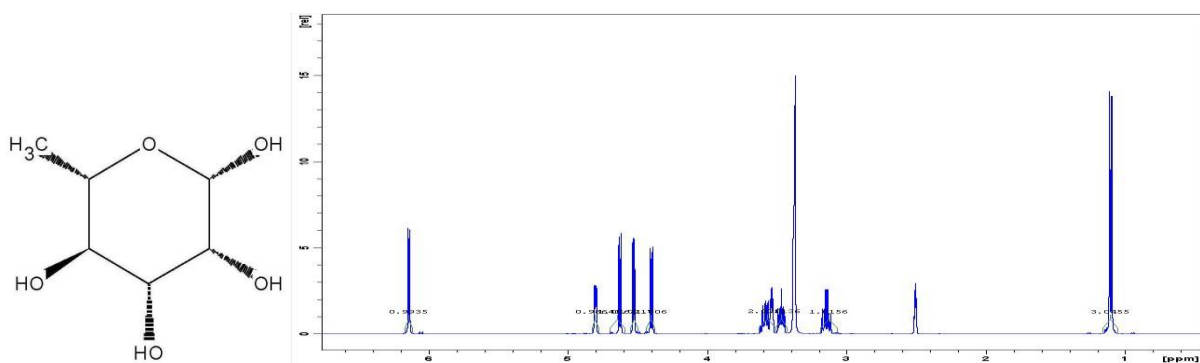


c

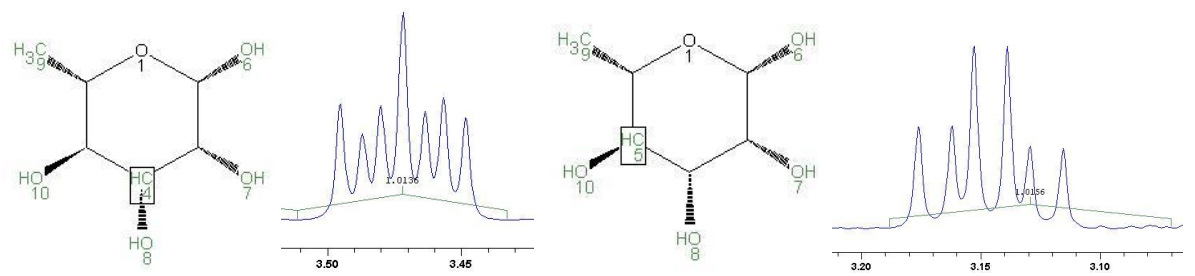
H<sub>2</sub>O

DMSO

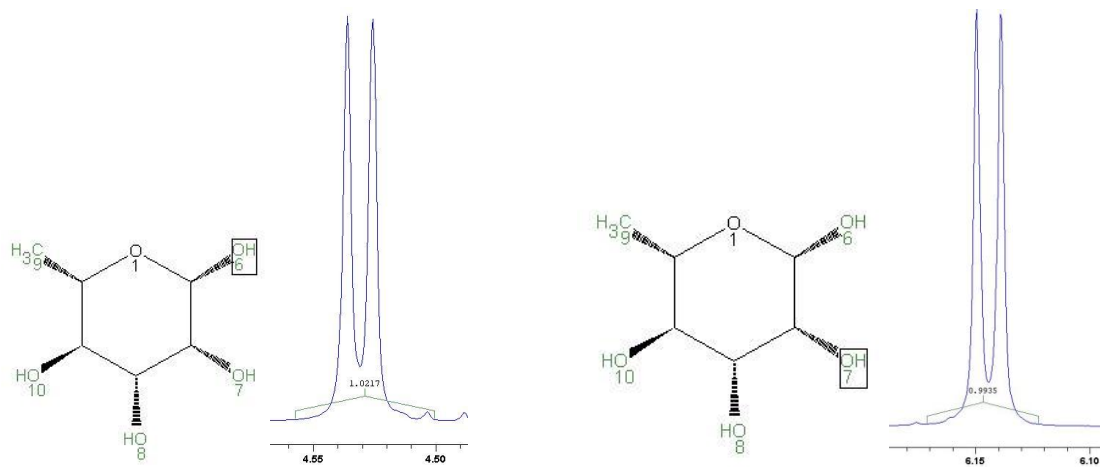
### 50. L-+-Rhamnose-Monohydrat



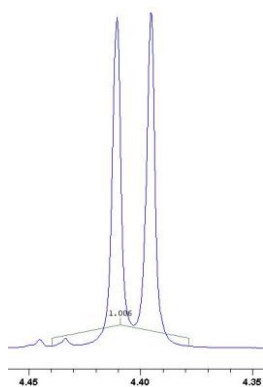
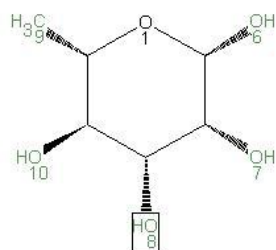
b



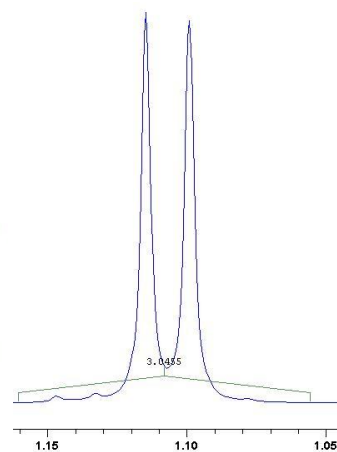
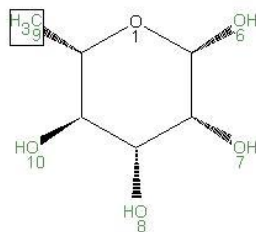
d



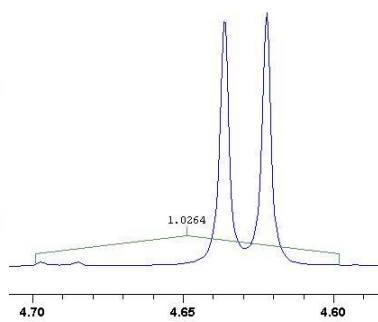
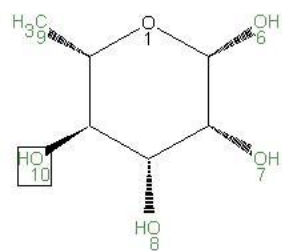
**f**



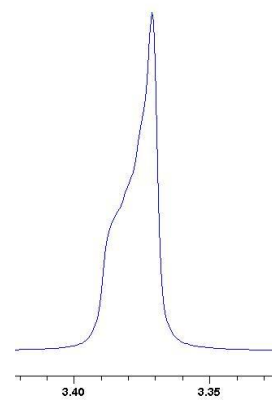
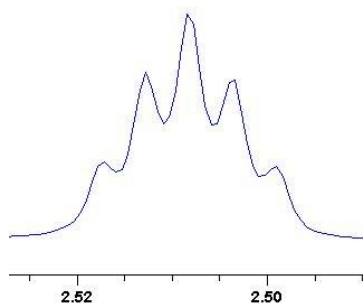
g



h

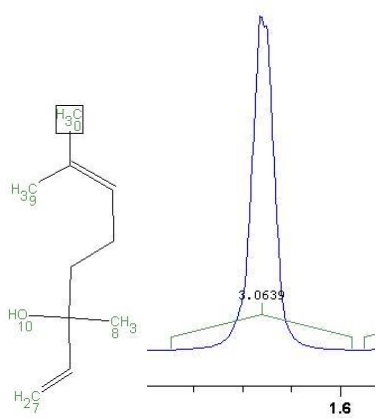
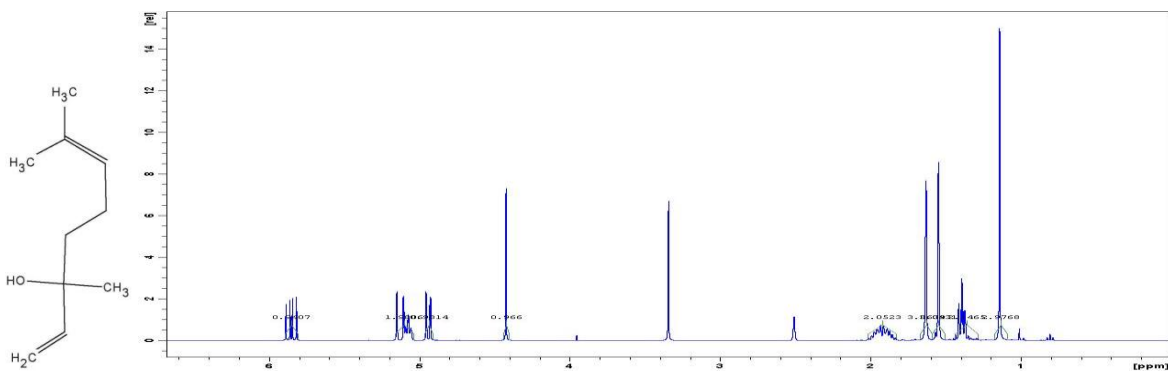


i

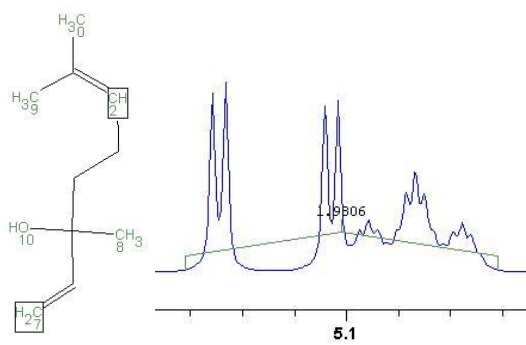
H<sub>2</sub>O

DMSO

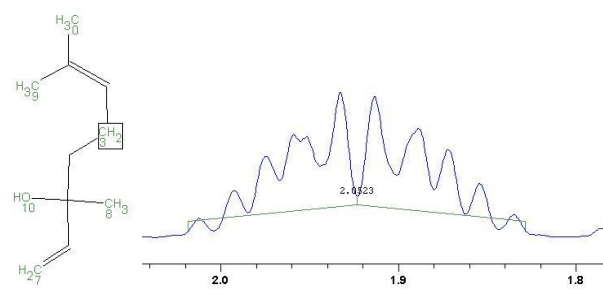
## 51. Linalool



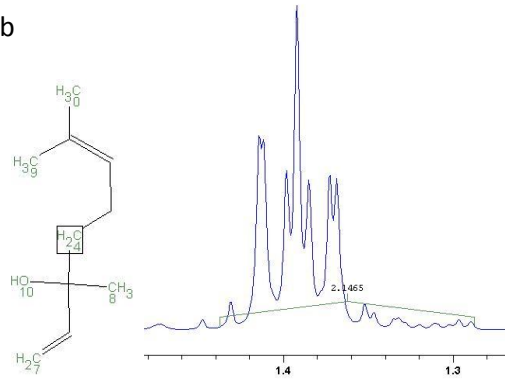
a



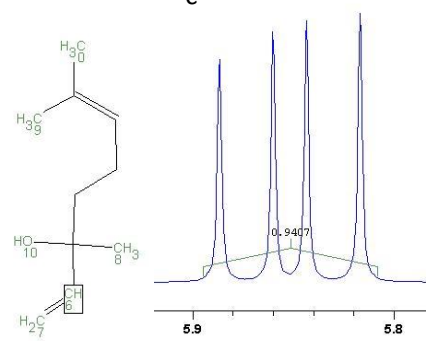
b



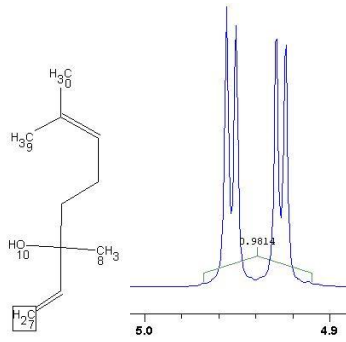
C



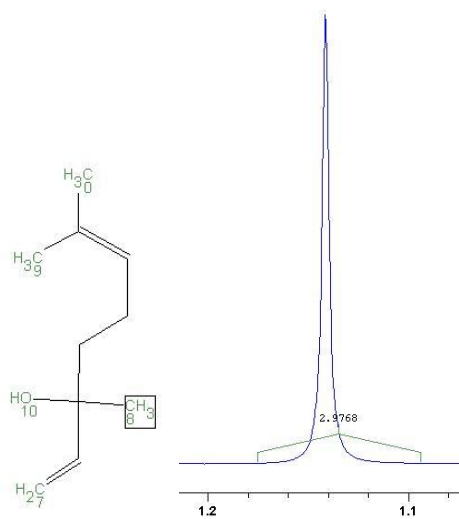
d



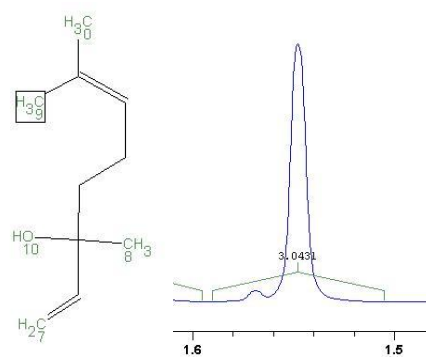
e



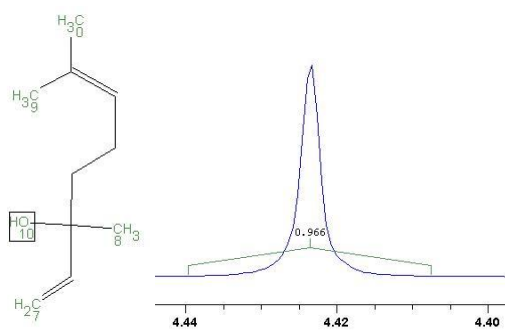
**f**



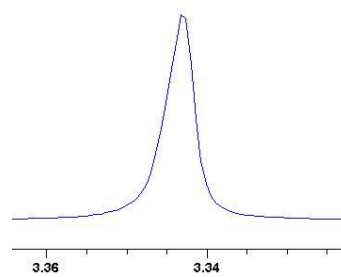
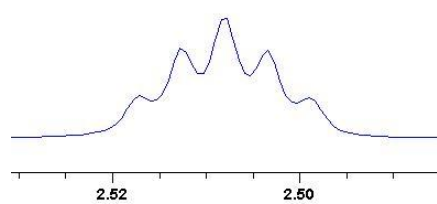
g



h

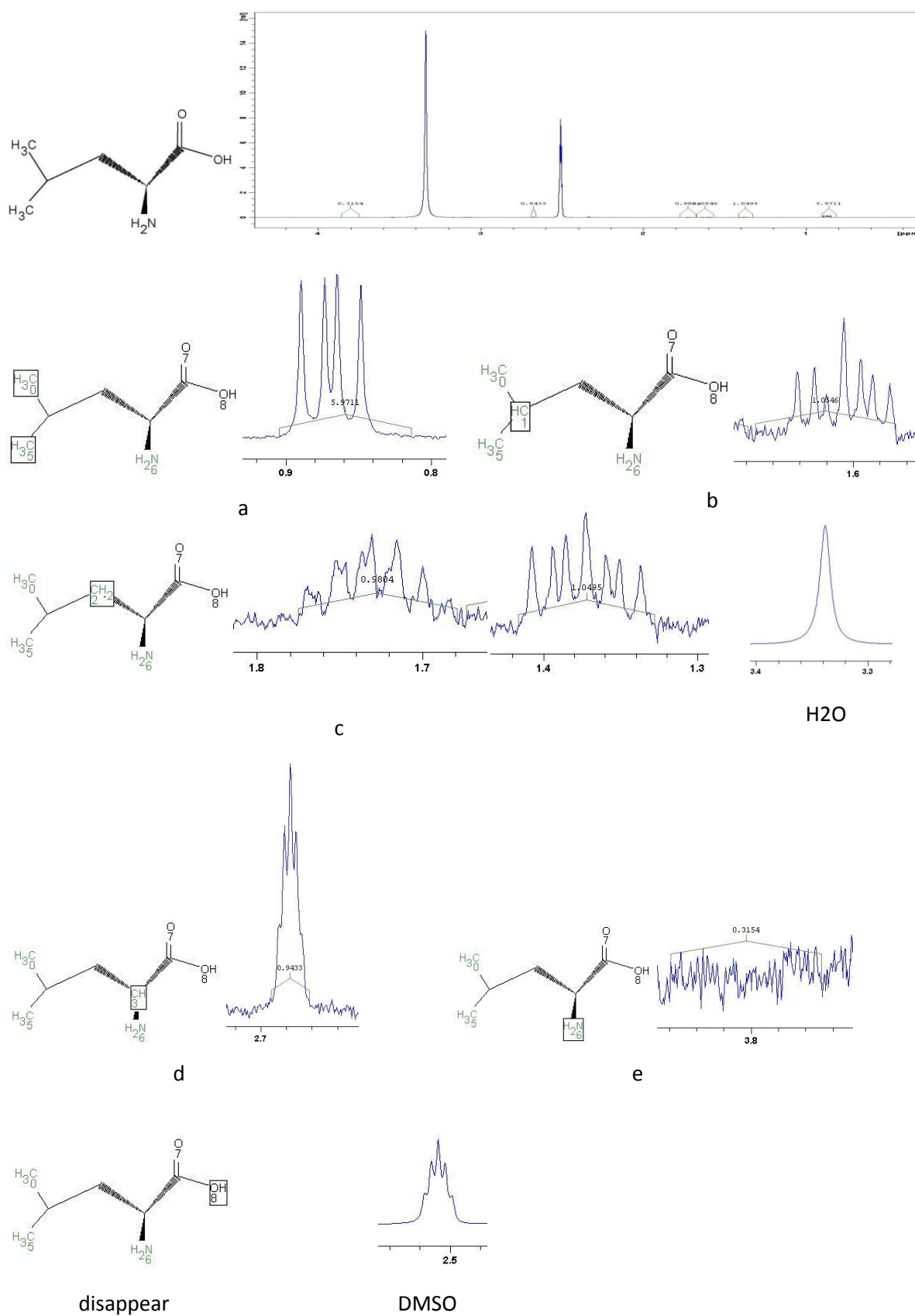


i

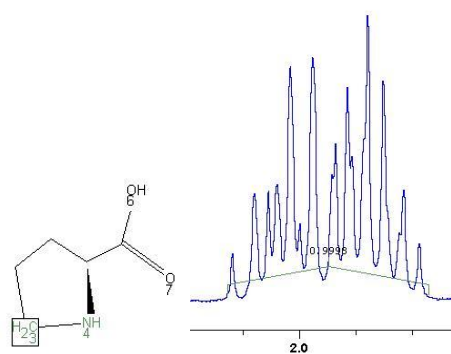
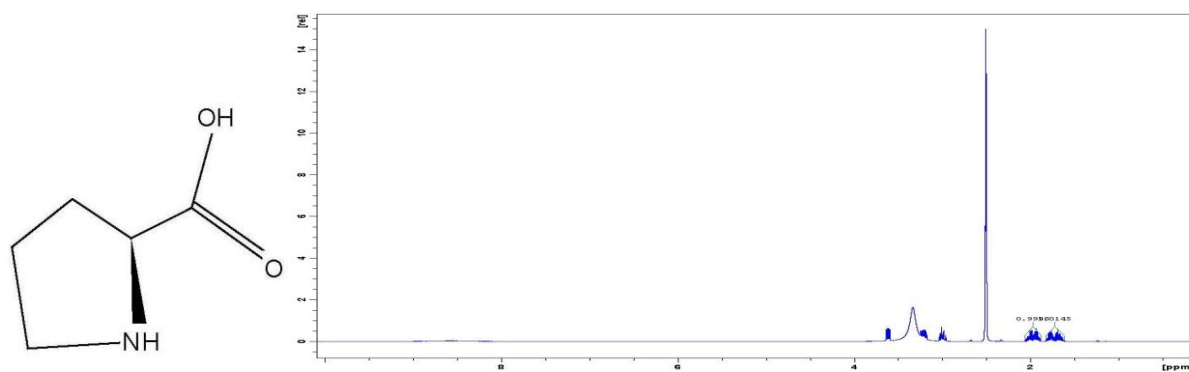
H<sub>2</sub>O

DMSO

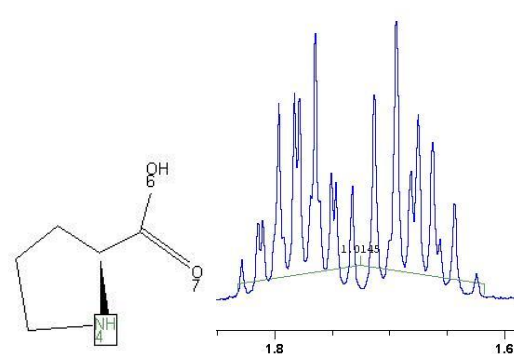
## 52. L-Leucin



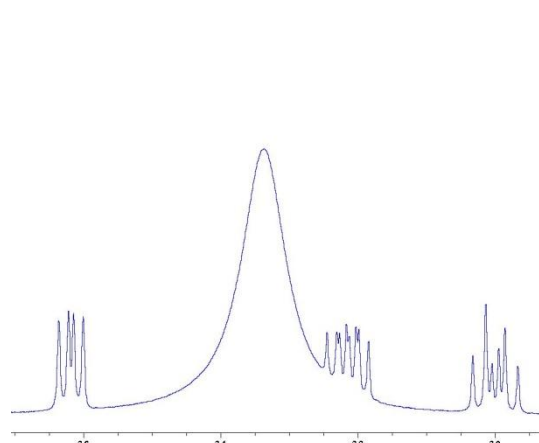
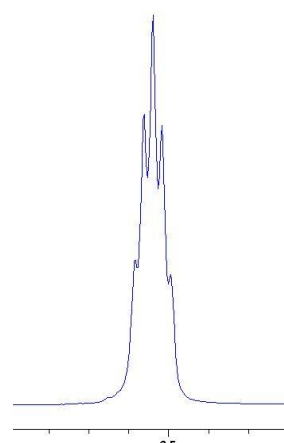
## 53. L-Prolin



a



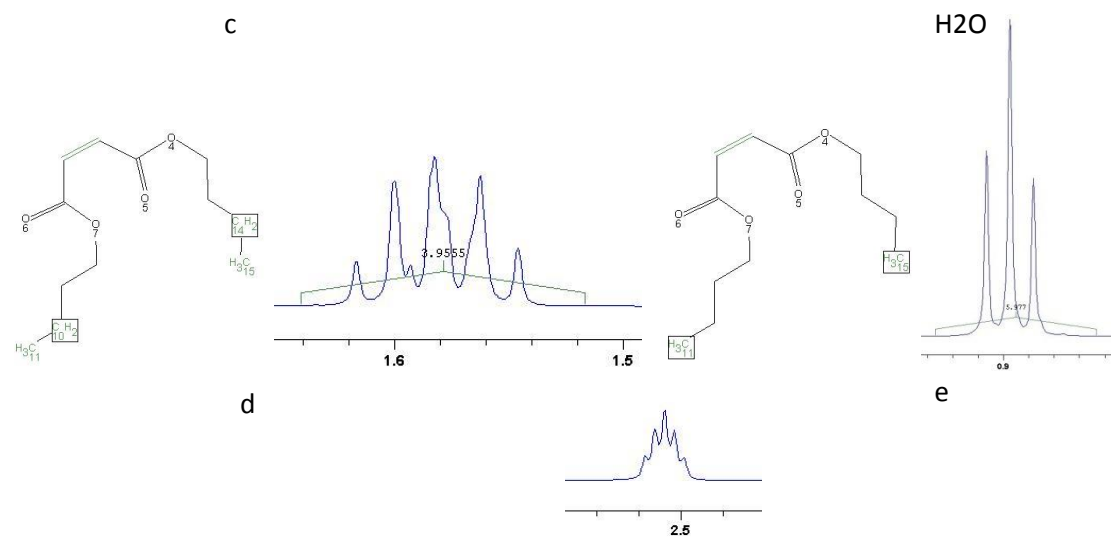
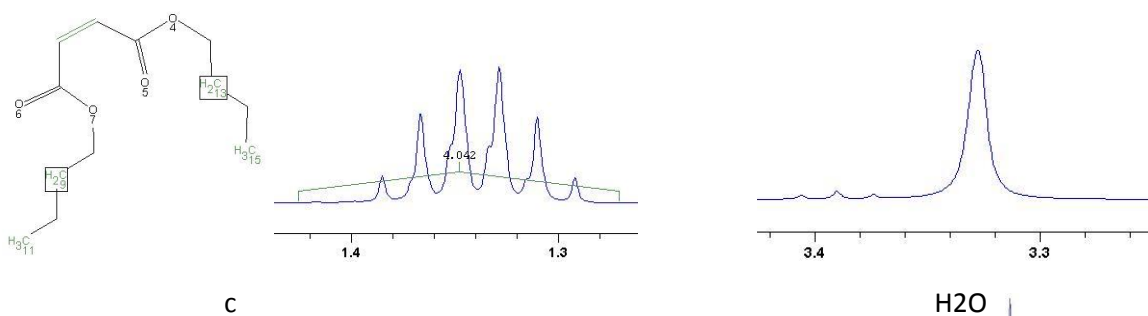
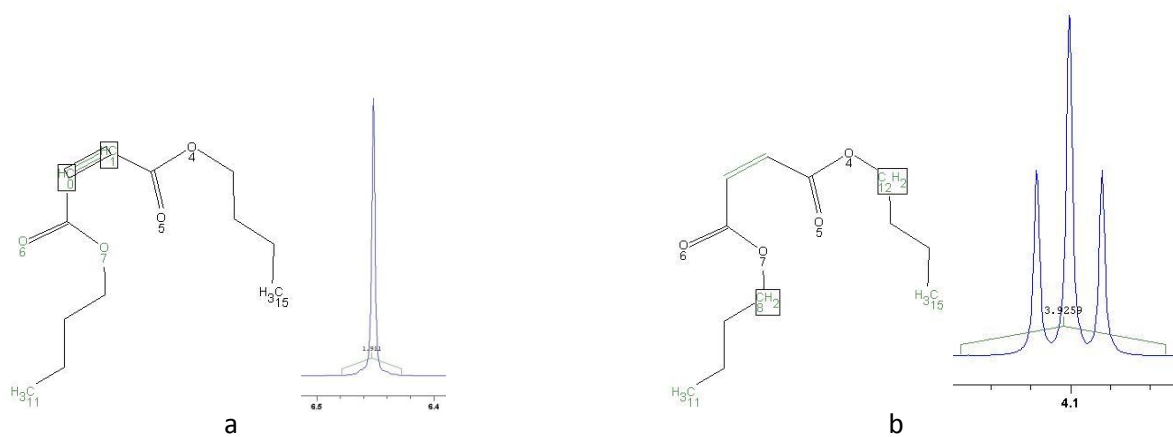
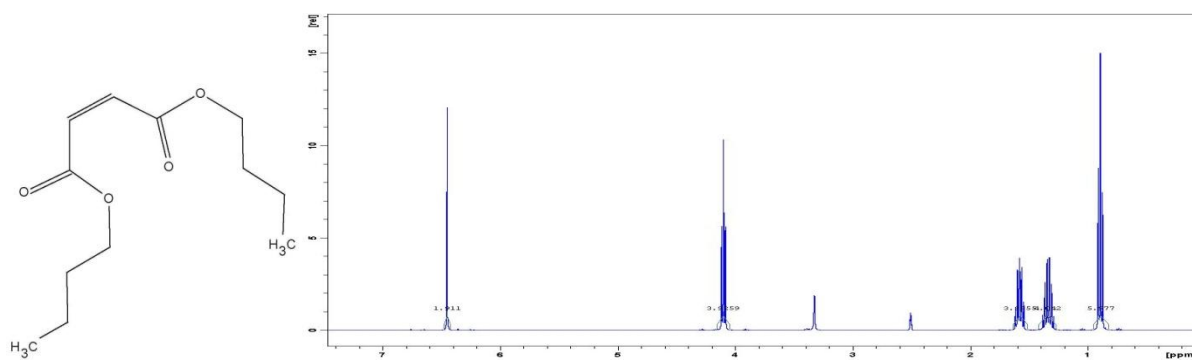
b

H<sub>2</sub>O

DMSO

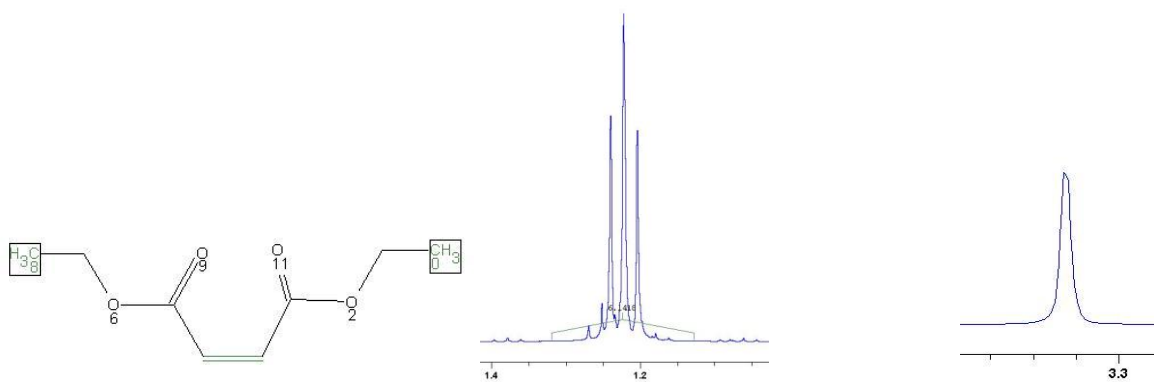
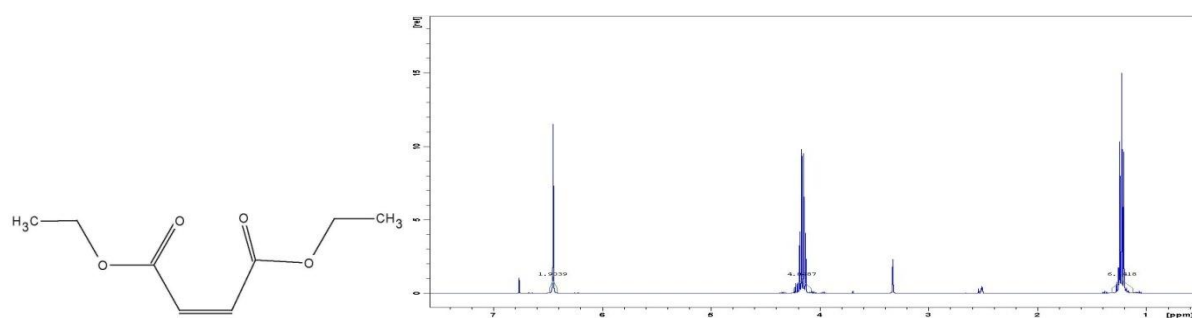


## 54. Maleinsaeure-dibutylester

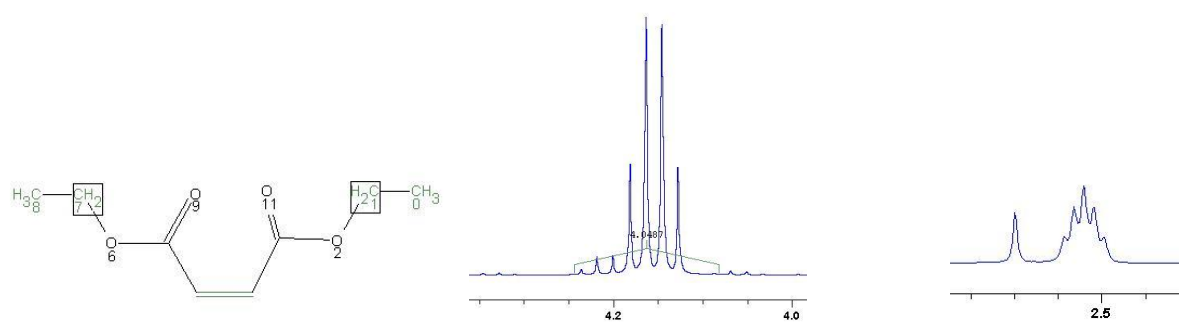


DMSO

## 55. Maleinsaeure-diethylester

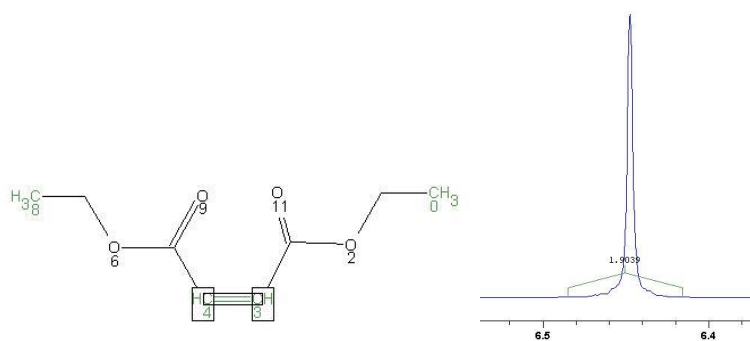


a

H<sub>2</sub>O

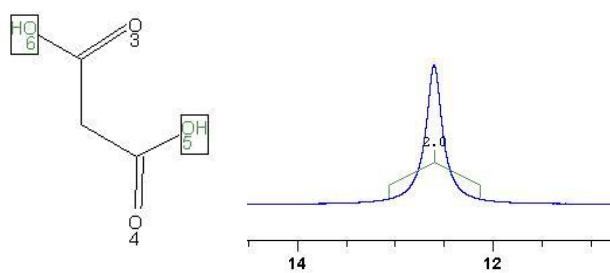
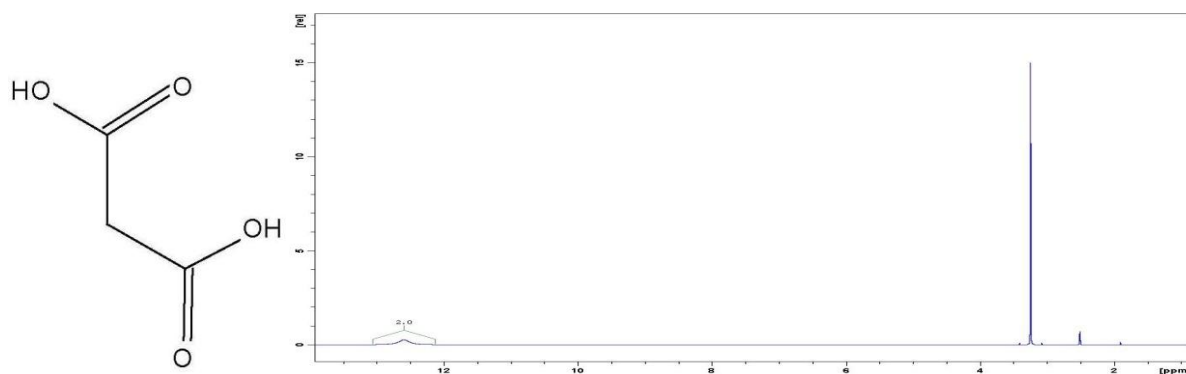
b

DMSO

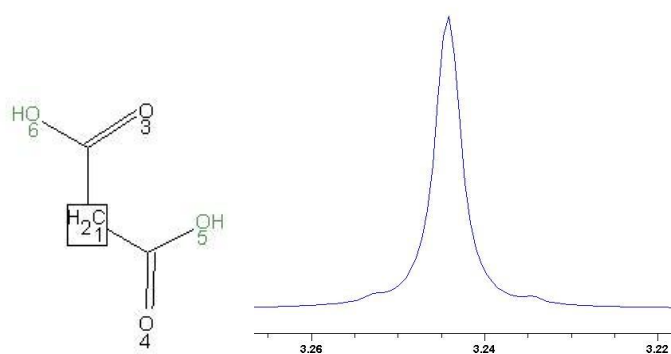


c

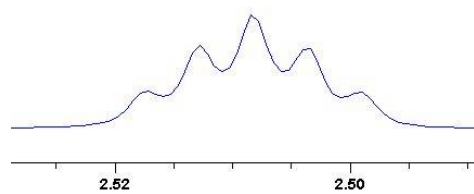
## 56. Malonsaeure



a

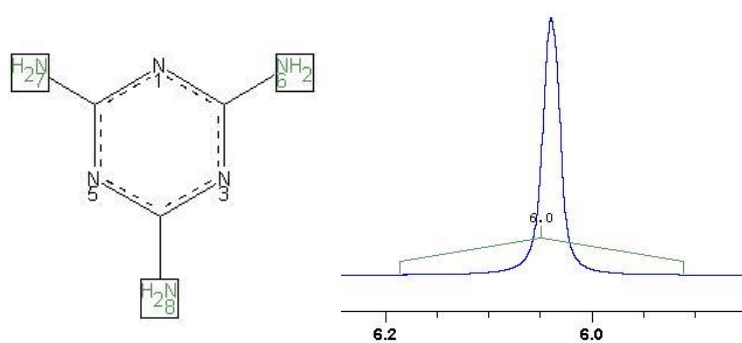
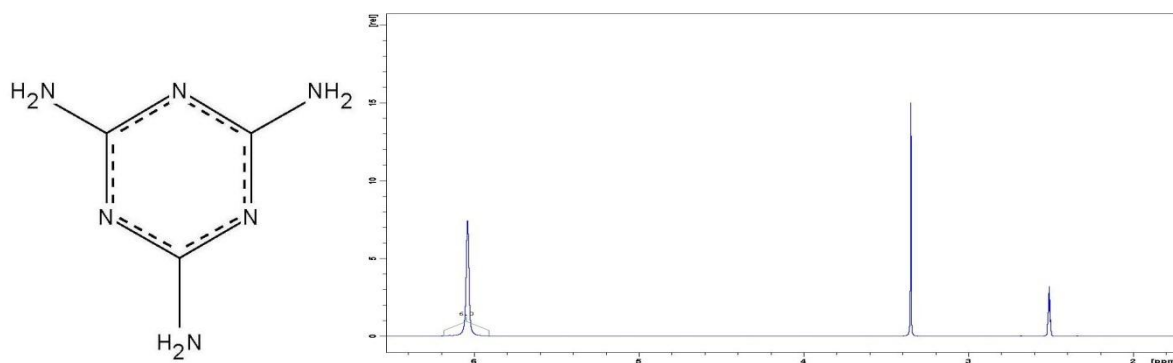


b and H2O

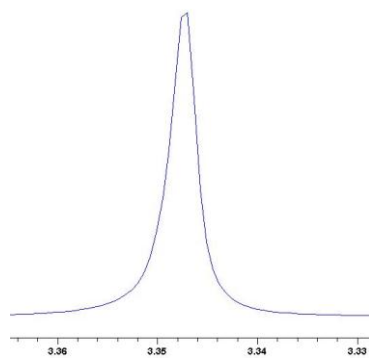
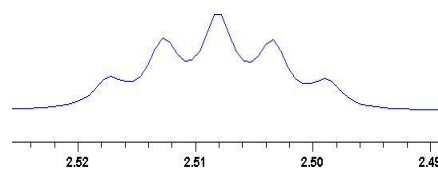


DMSO

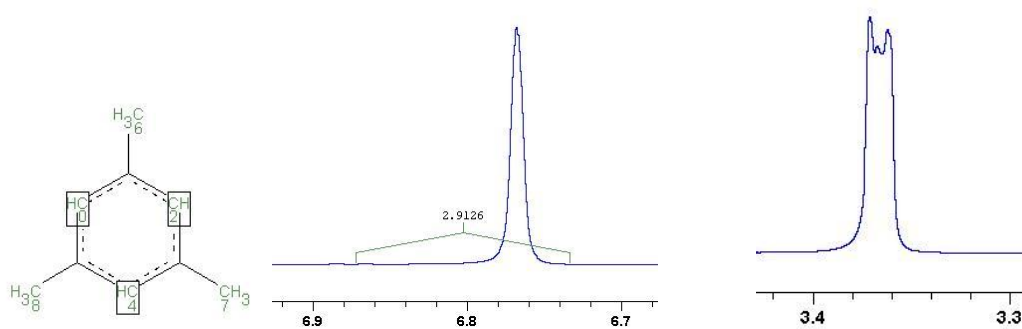
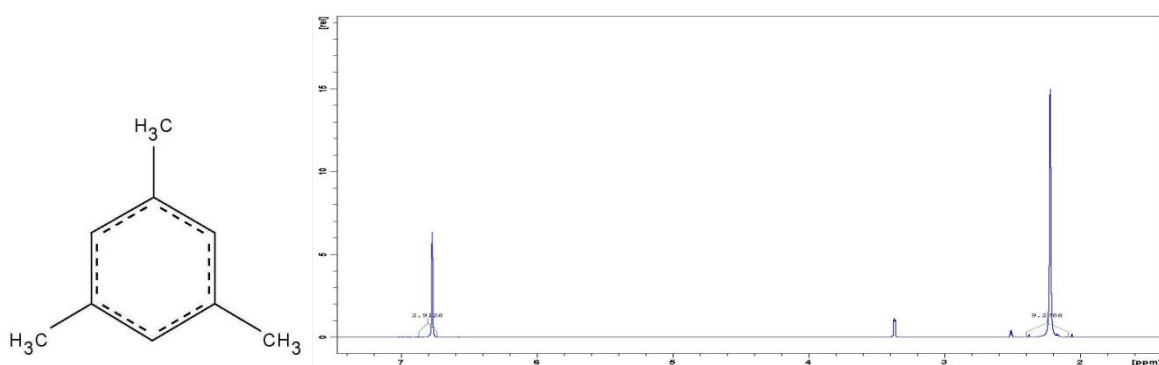
## 57. Melamin



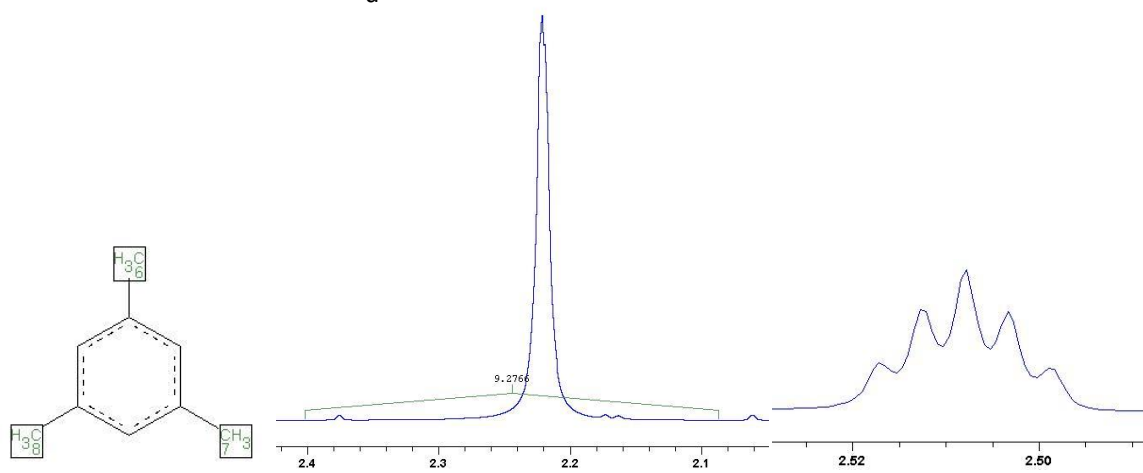
a

 $\text{H}_2\text{O}$  $\text{DMSO}$

## 58. Mesiyylen



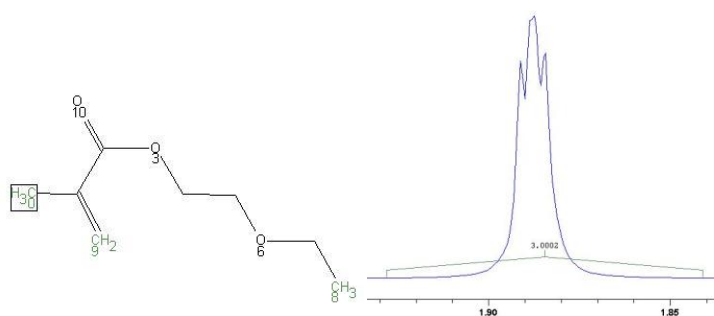
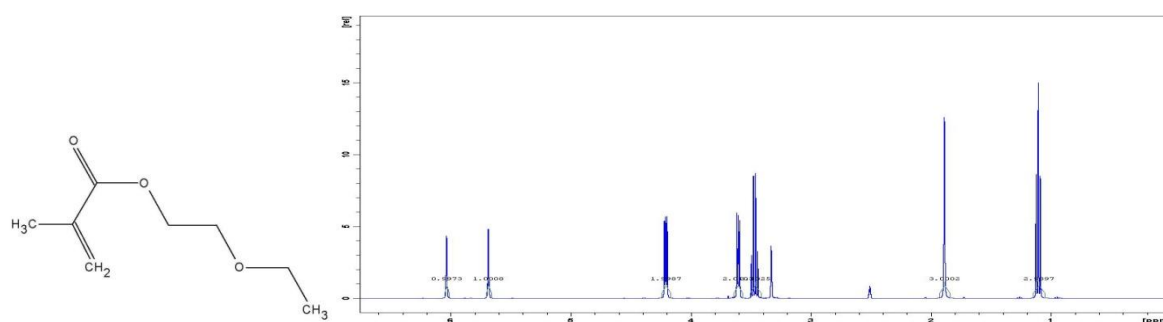
a

 $\text{H}_2\text{O}$ 

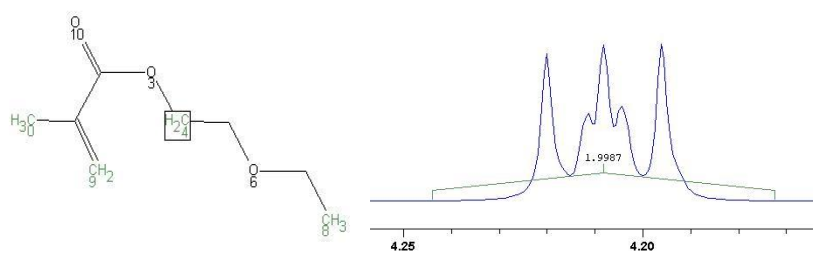
b

DMSO

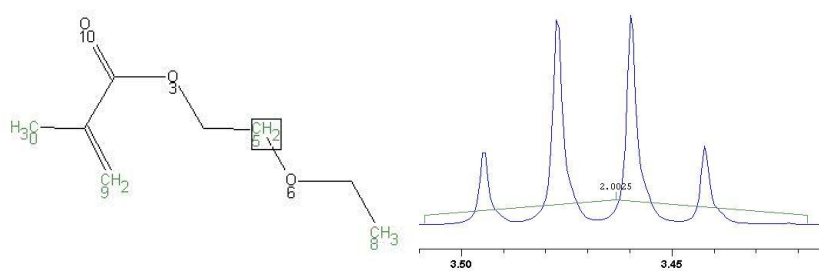
## 59. Methacrylsaeure-2-ethoxyethylester



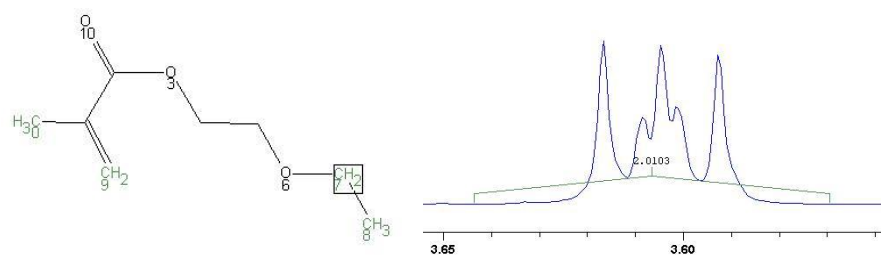
a



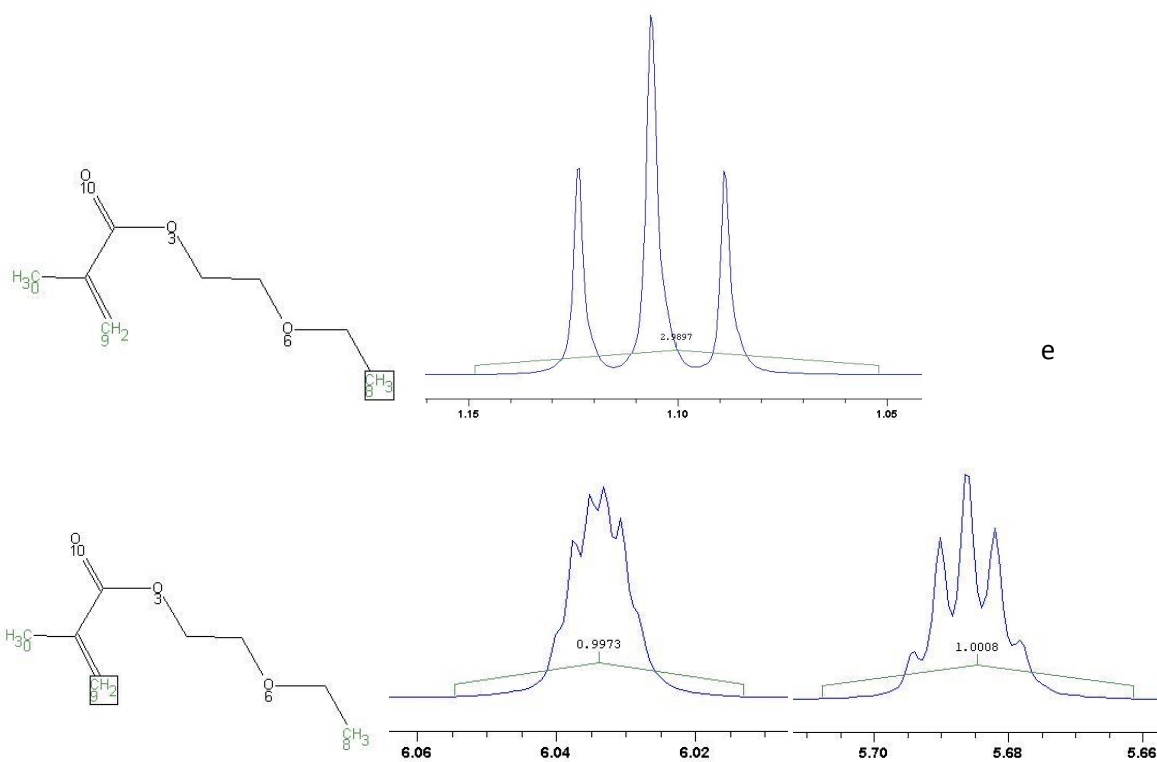
b



c

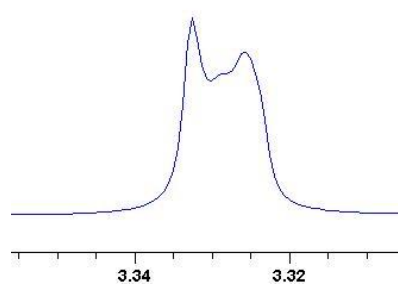


d

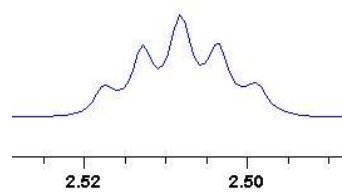


e

f

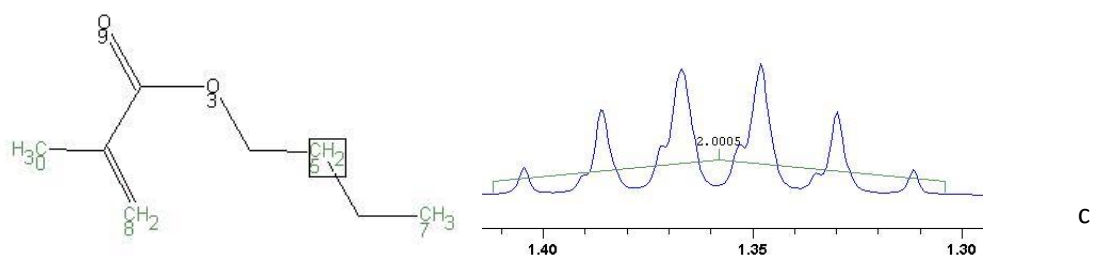
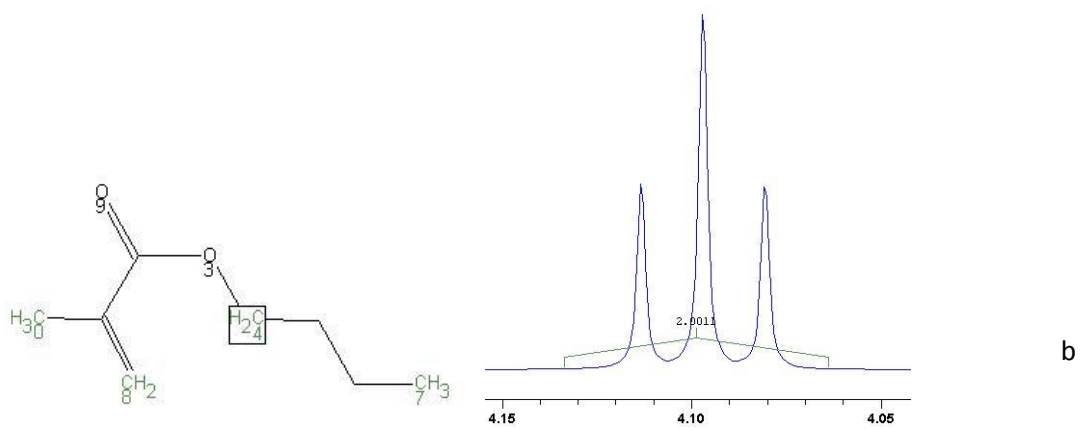
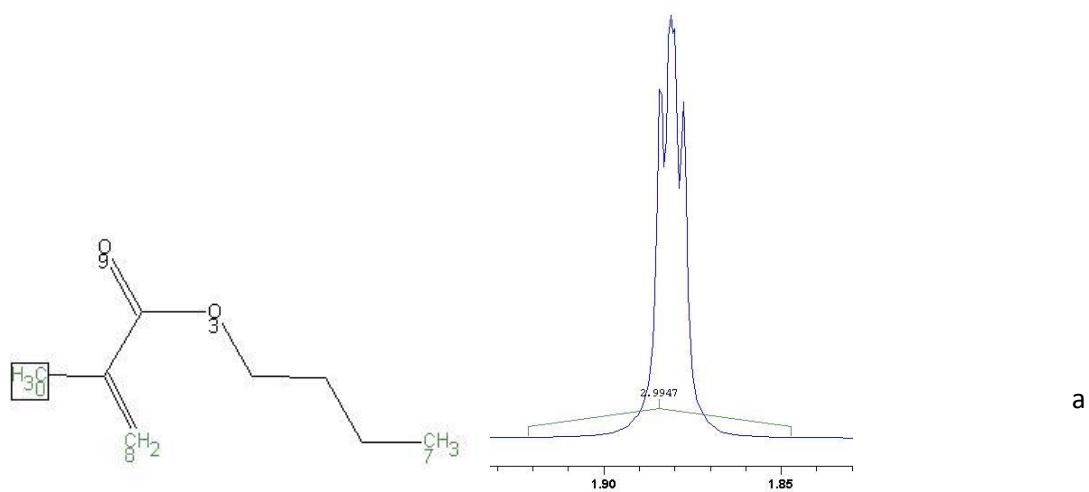
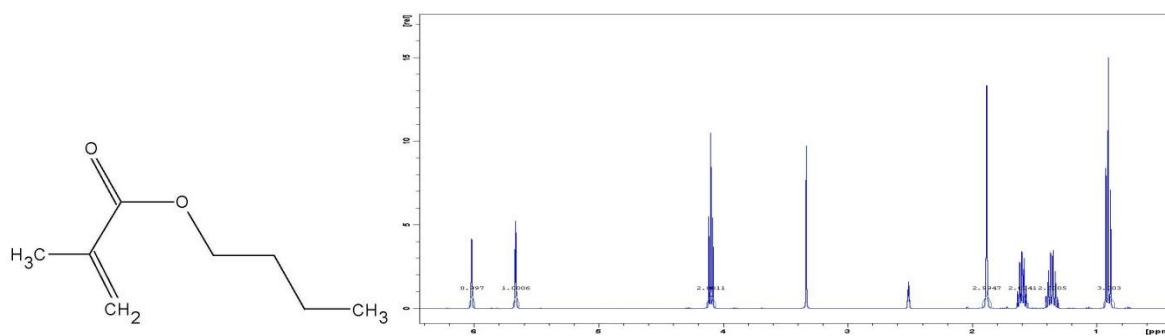


H2O

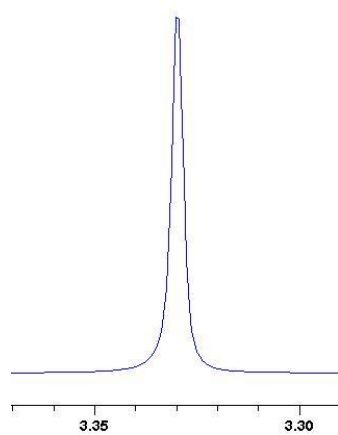
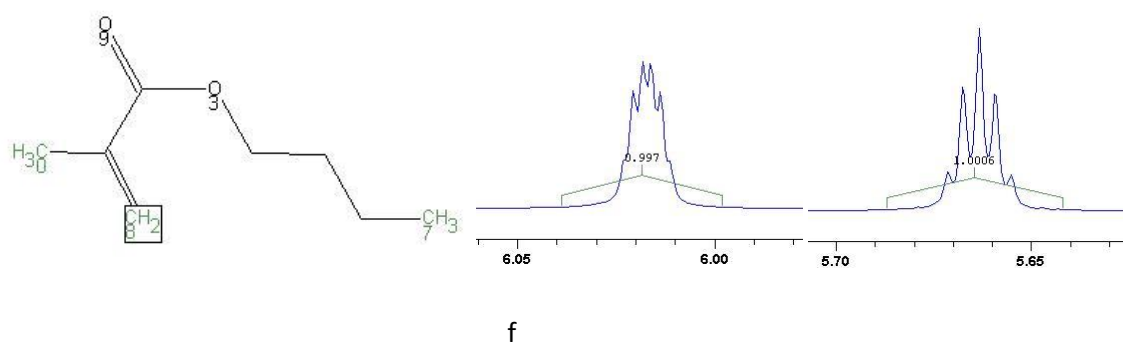
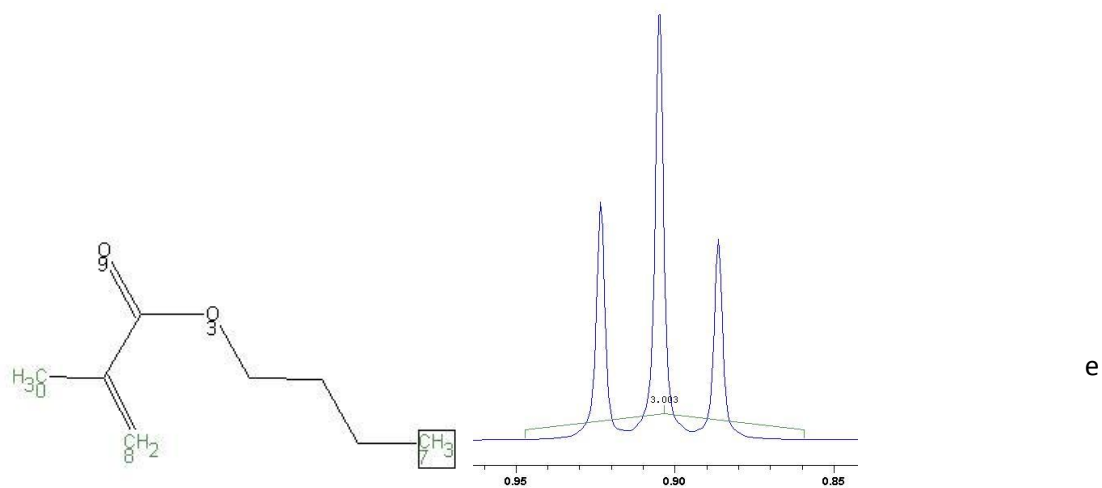
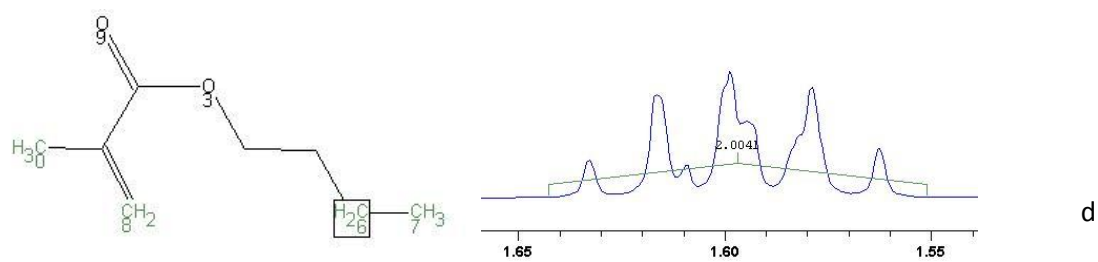


DMSO

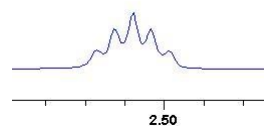
## 60. Methacrylsaeure-butylester





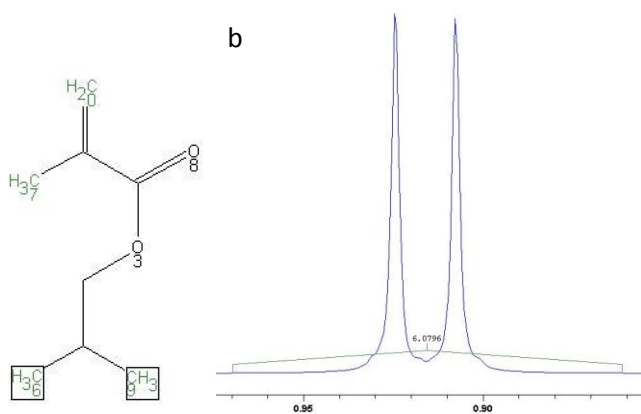
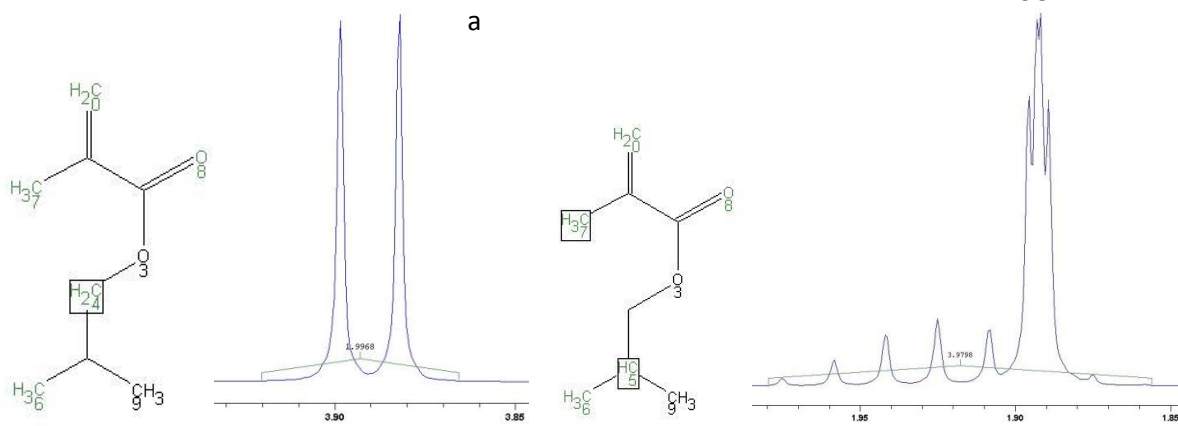
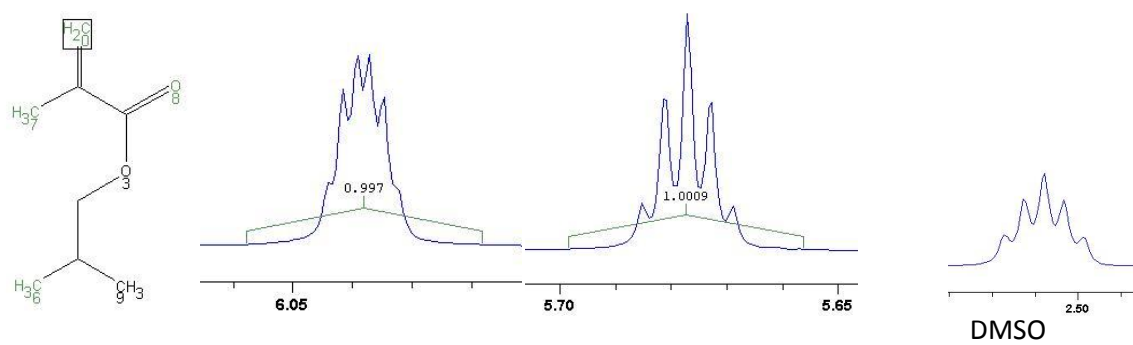
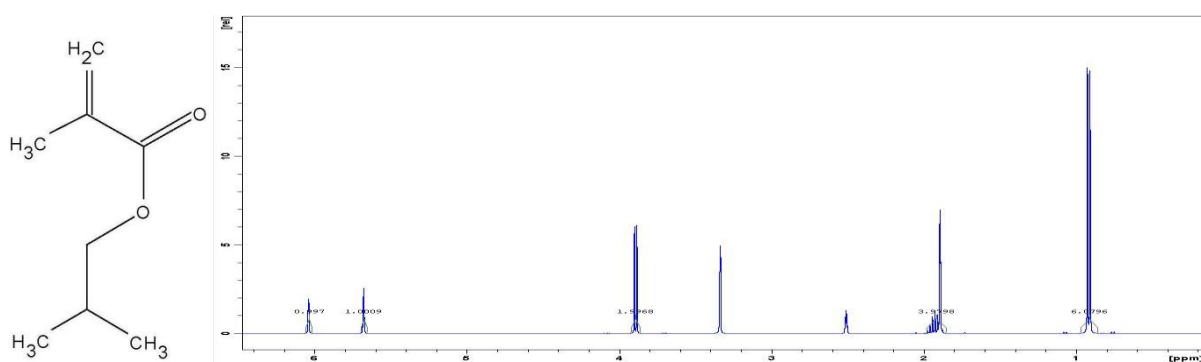


H<sub>2</sub>O



DMSO

## 61. Methacrylsaeure-isobutylester

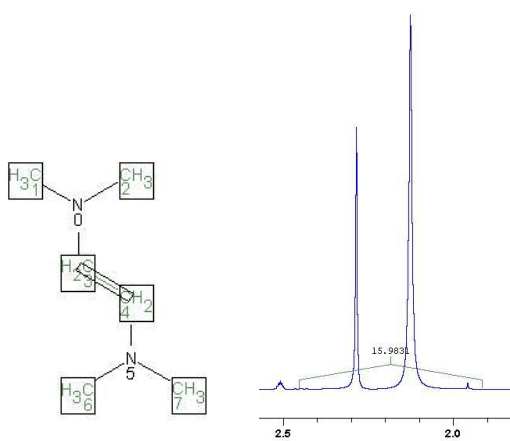
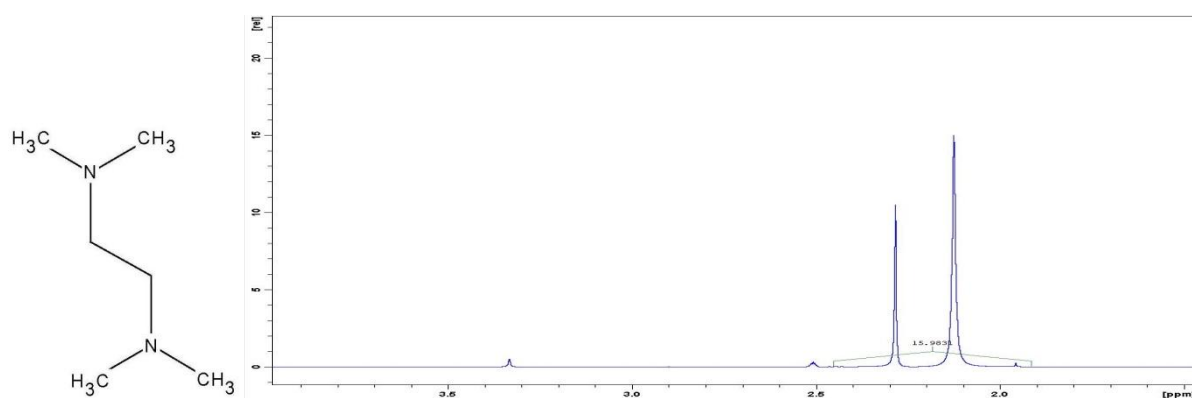


d

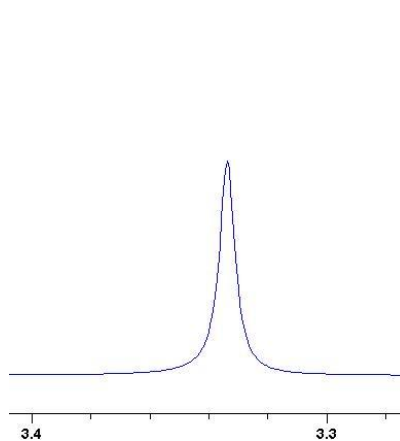
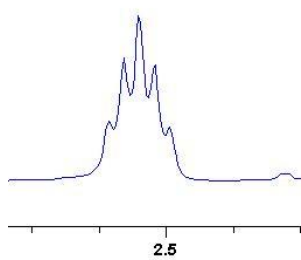
c

H<sub>2</sub>O

## 62. N,N,N,N-Tetramethyl-ethylendiamin

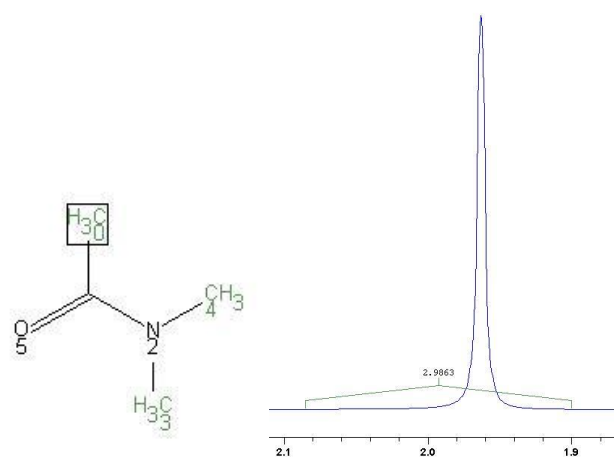
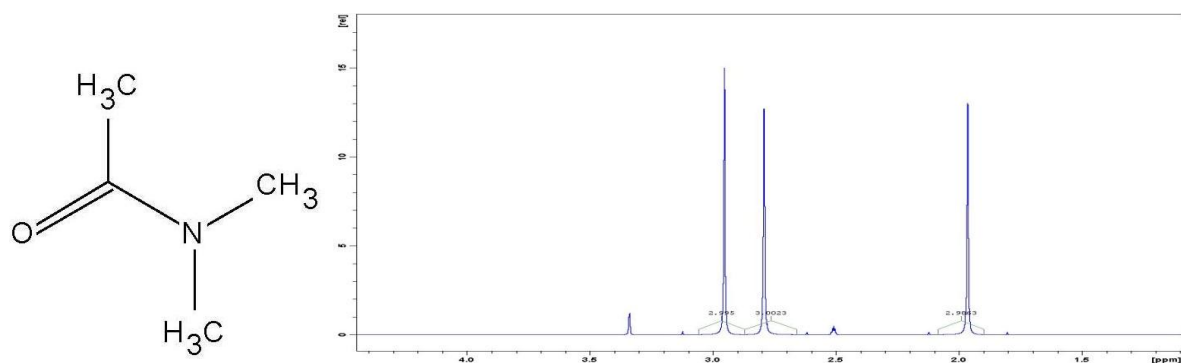


a

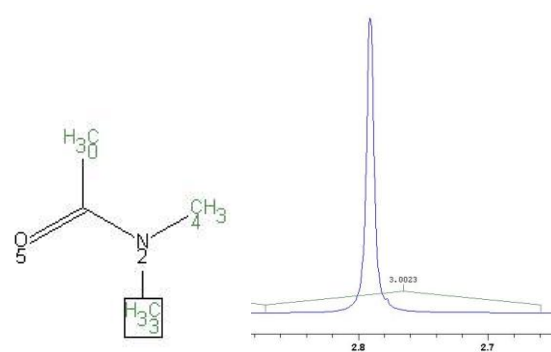
H<sub>2</sub>O

DMSO

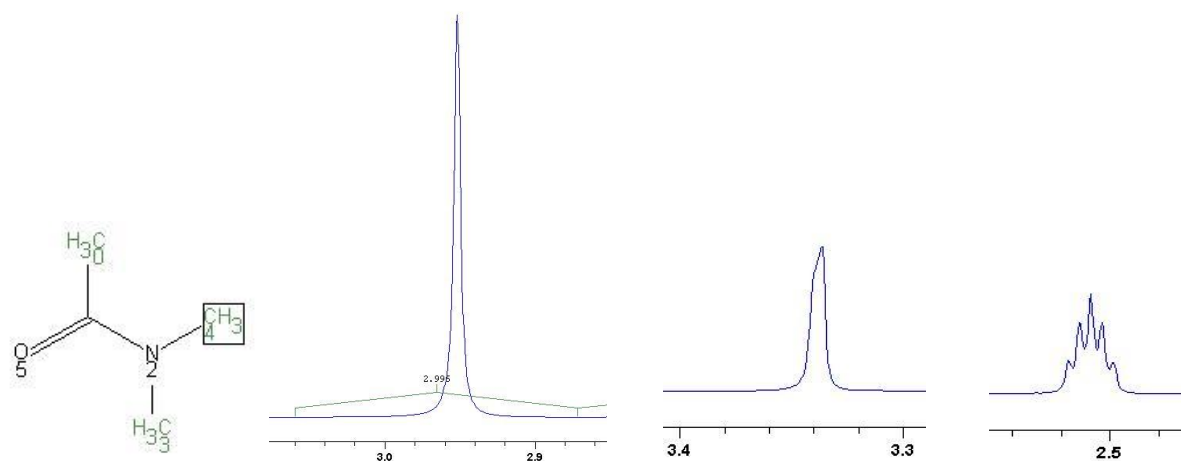
## 63. N,N-Dimethylacetamid



a



b

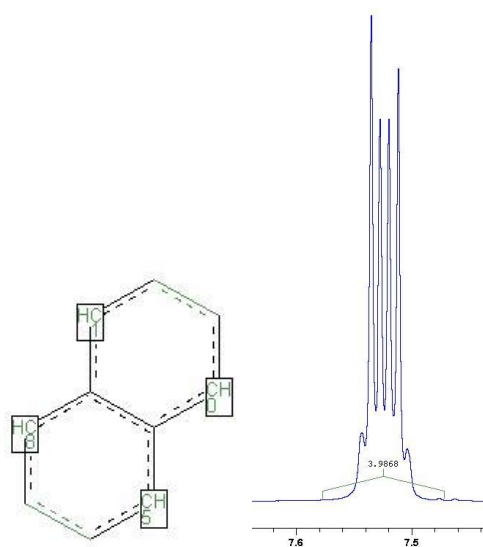
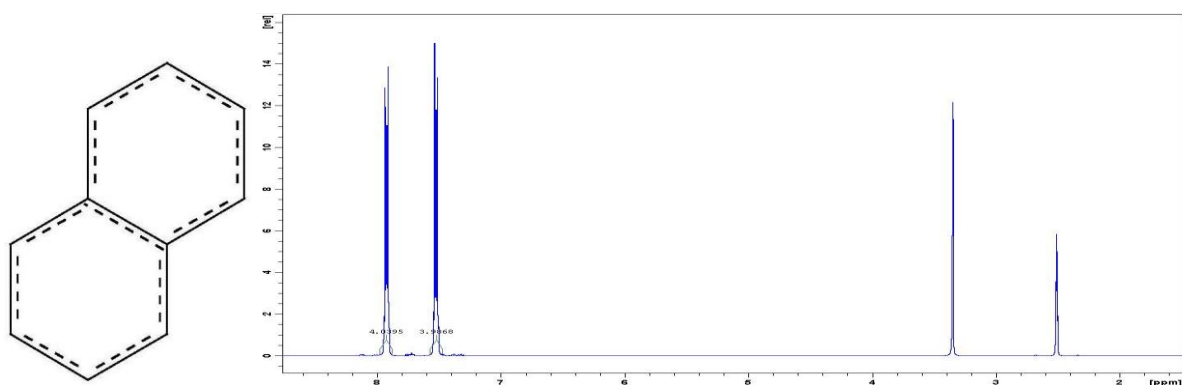


c

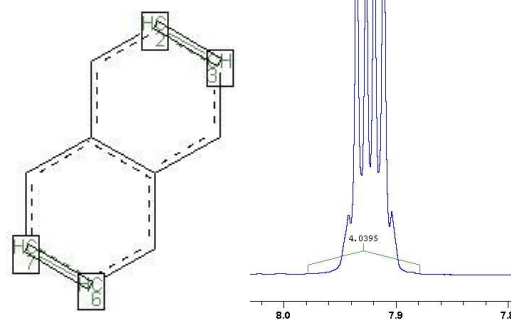
H<sub>2</sub>O

DMSO

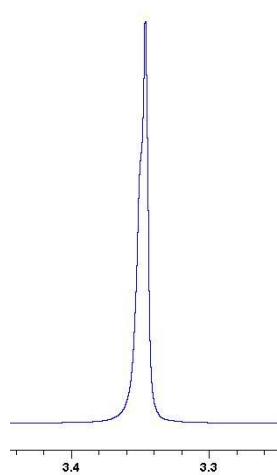
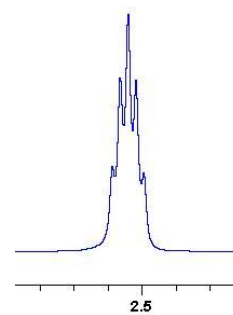
## 64. Naphthalin



a

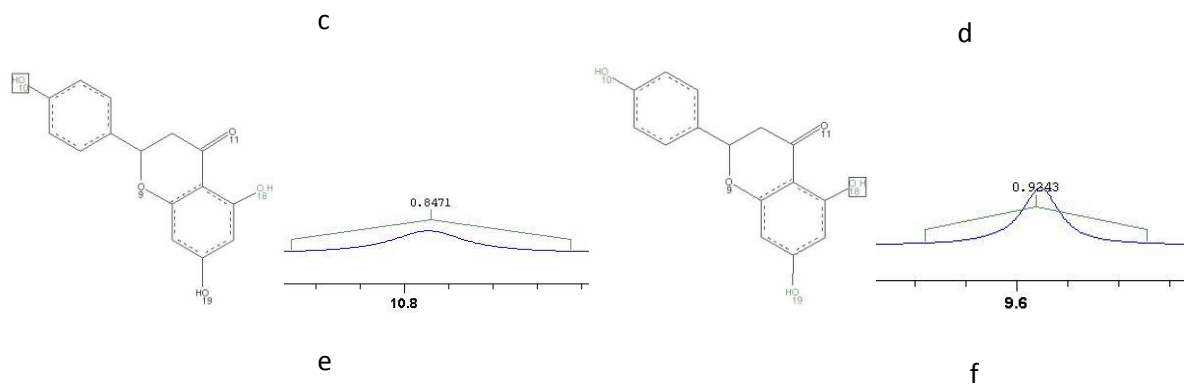
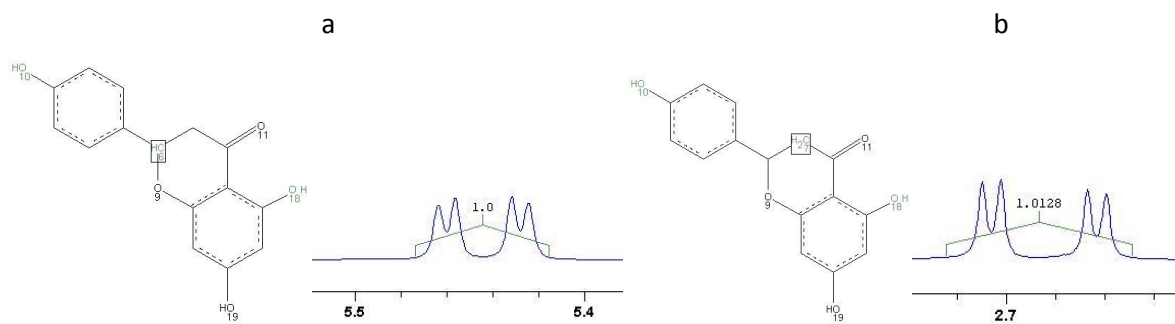
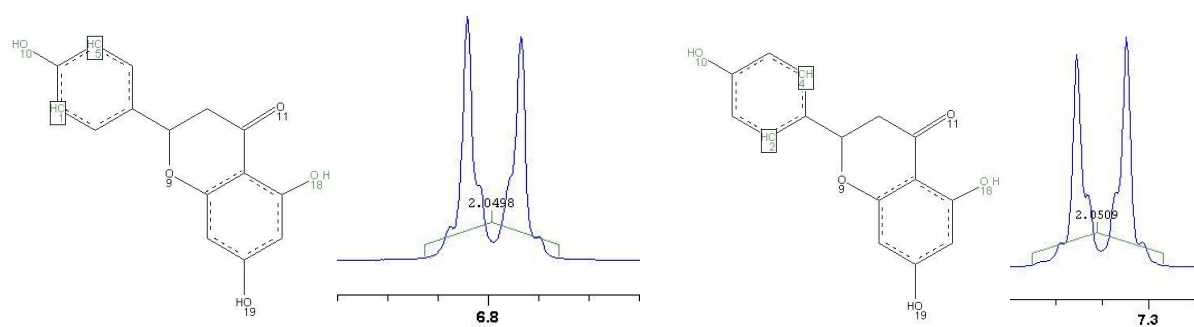
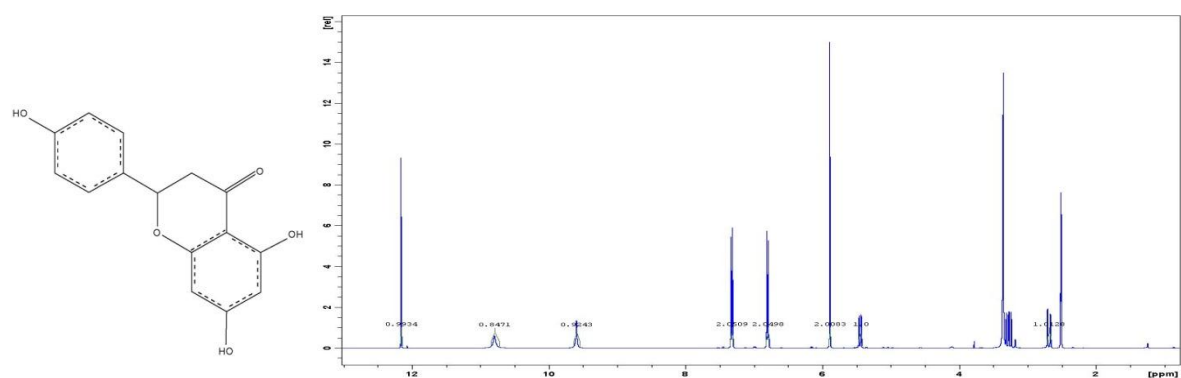


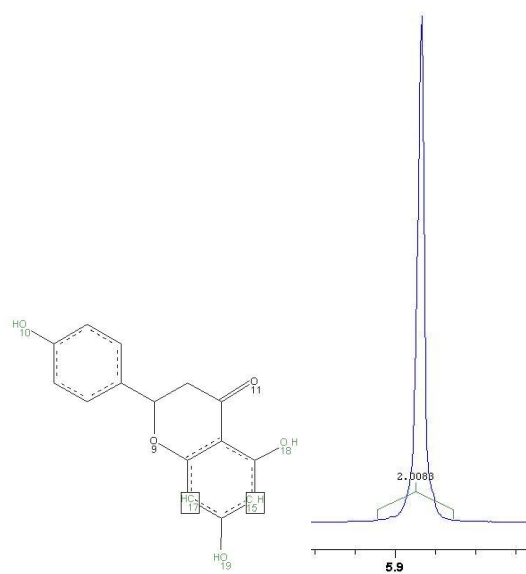
b

H<sub>2</sub>O

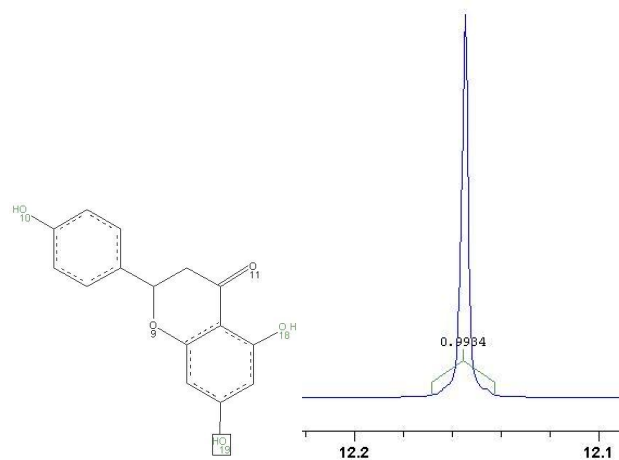
DMSO

## 65. Naringenin

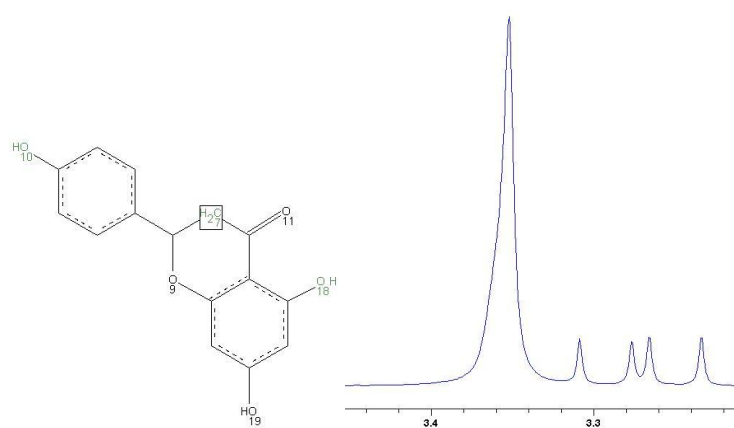




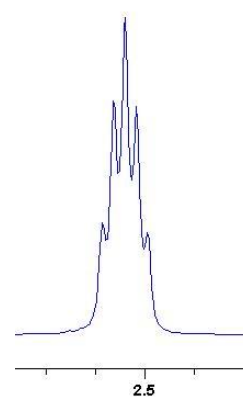
**g**



**h**

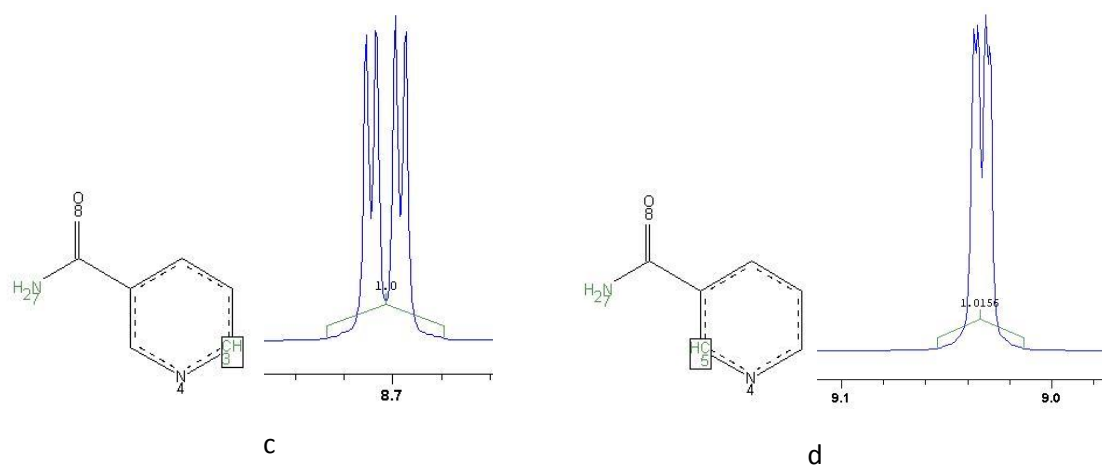
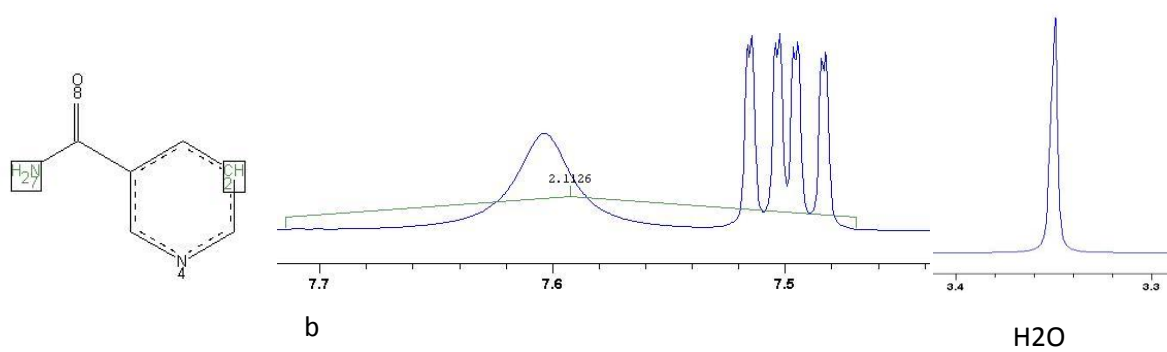
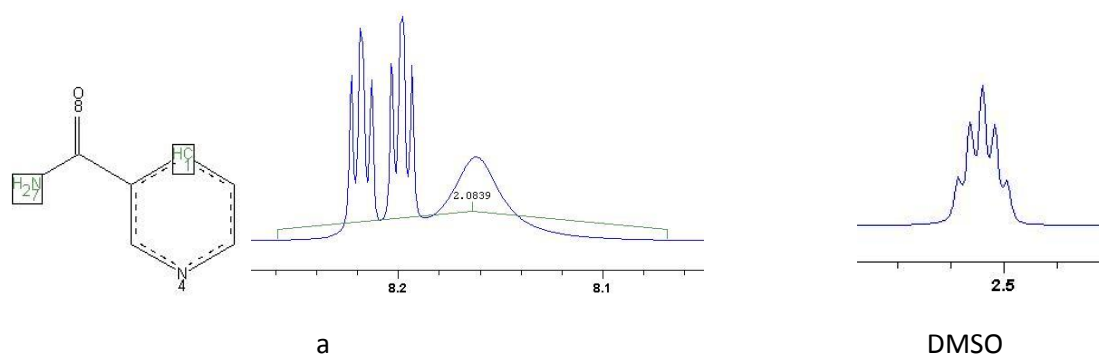
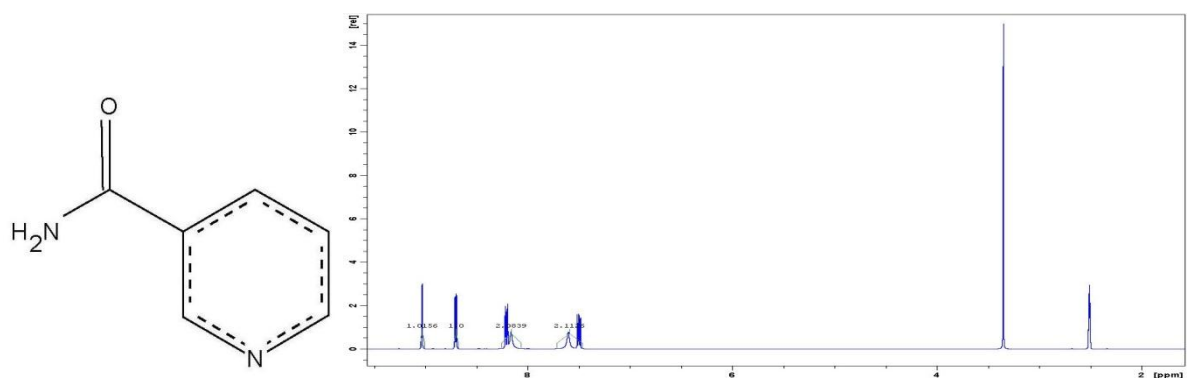


**i + H<sub>2</sub>O**



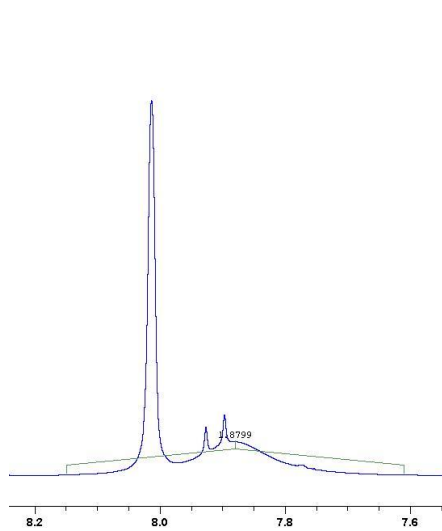
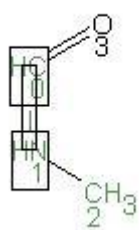
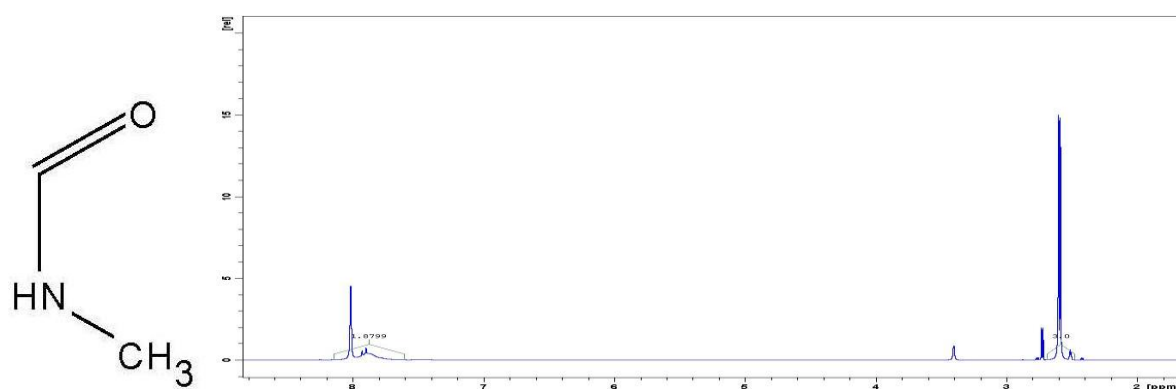
**DMSO**

## 66. Nicotinsaeureamid

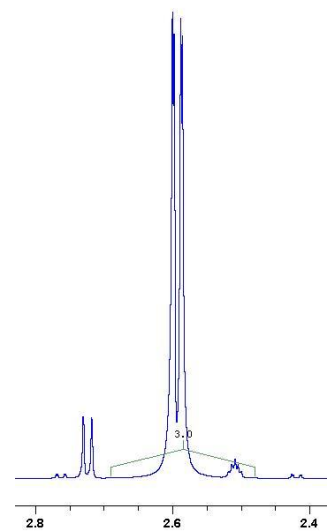
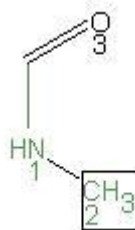




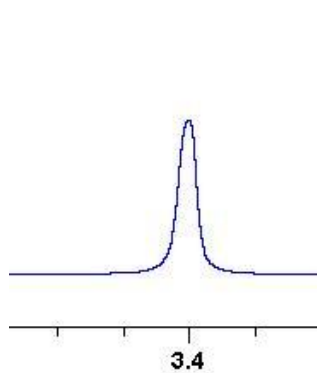
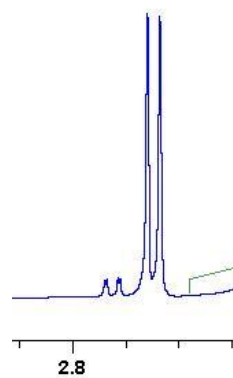
## 67. N-Methylformamid



a

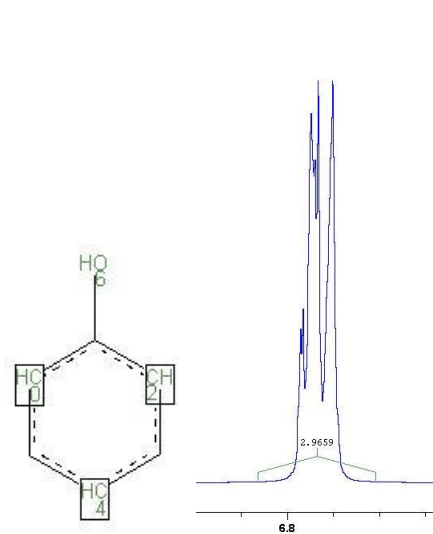
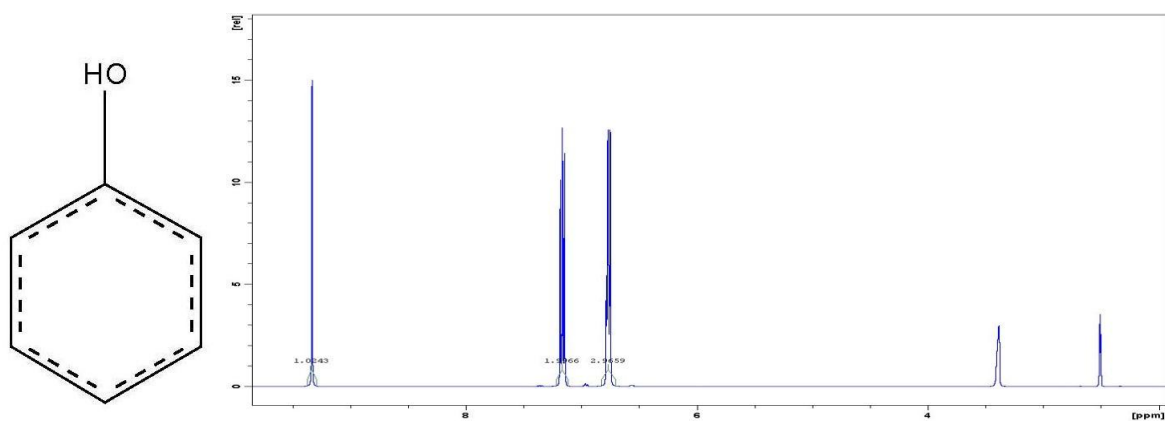


b

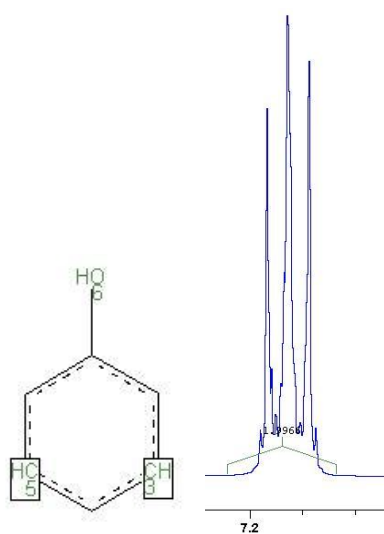
H<sub>2</sub>O

DMSO

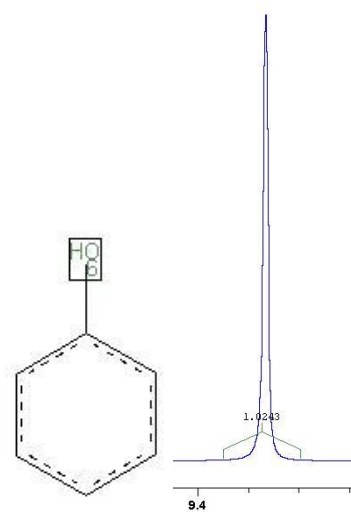
## 68. Phenol



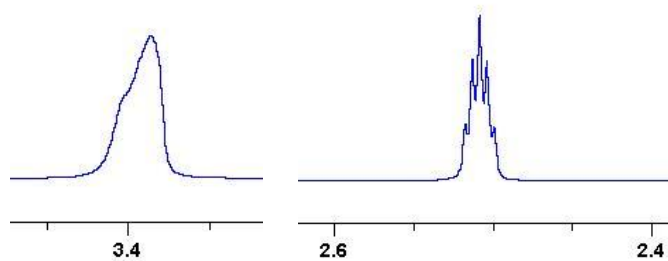
a



b

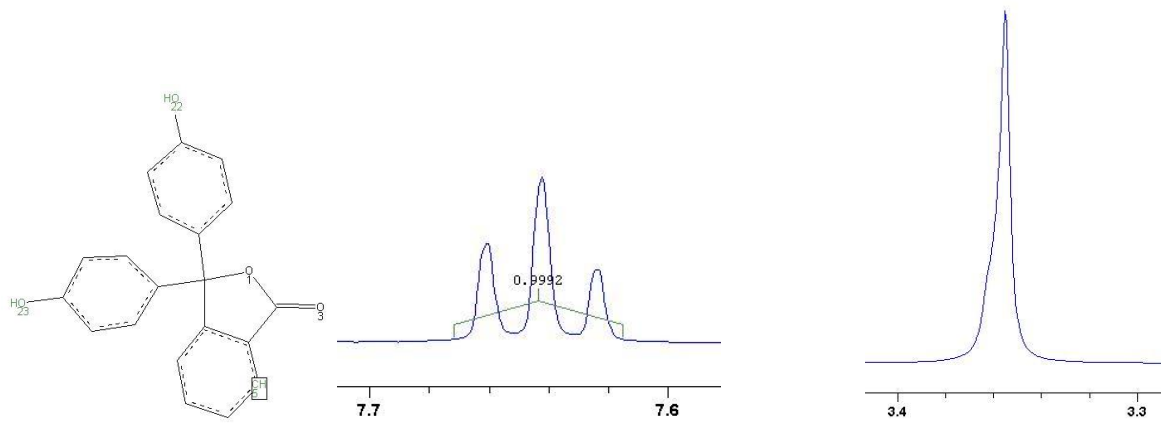
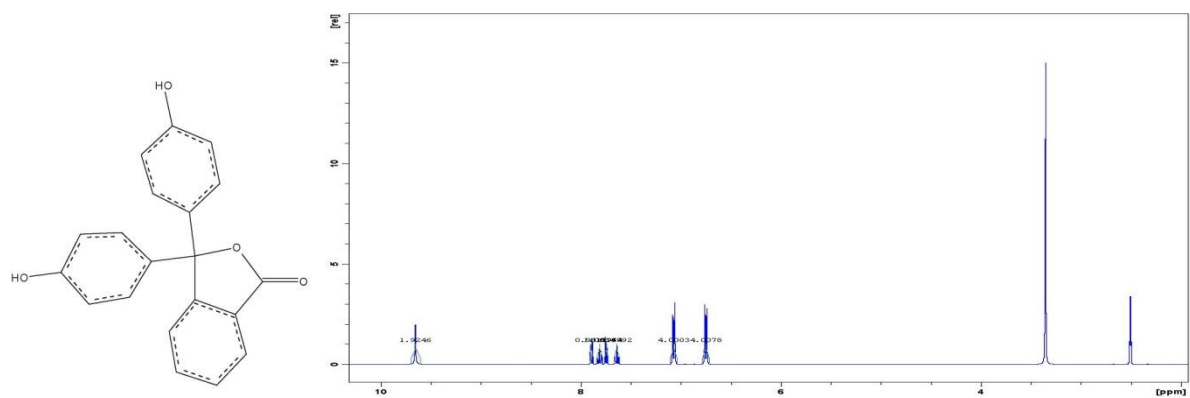


c

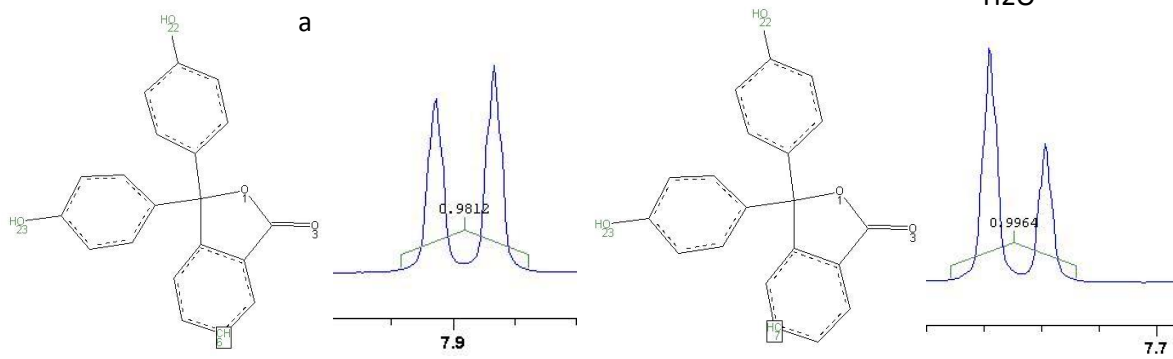
 $\text{H}_2\text{O}$ 

DMSO

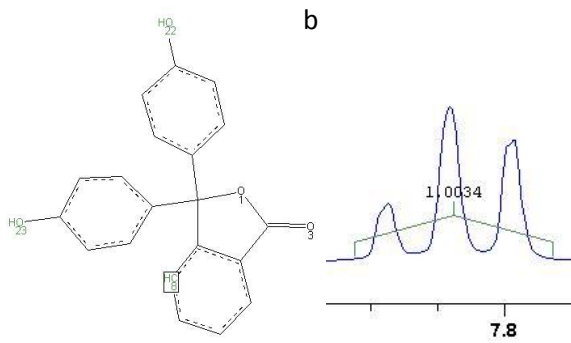
## 69. Phenolphthalein



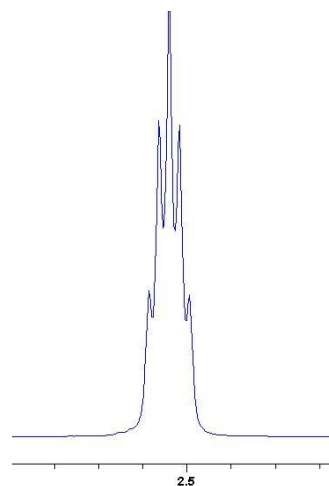
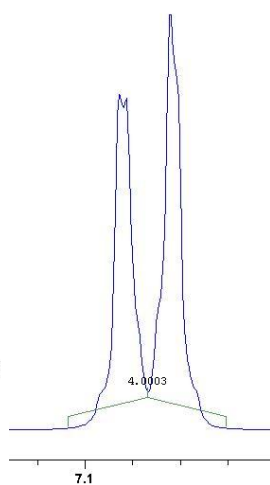
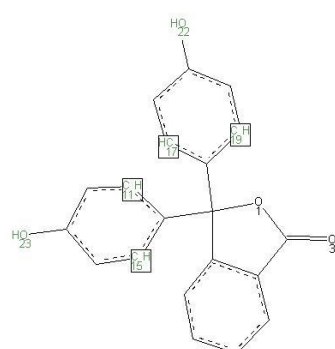
H<sub>2</sub>O



C

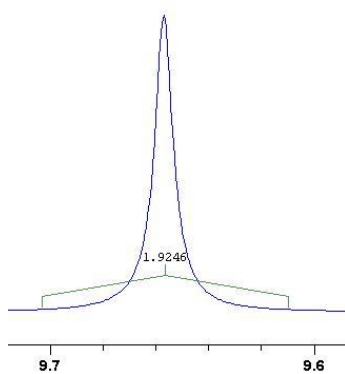
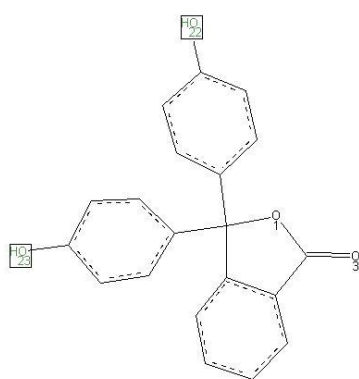


d

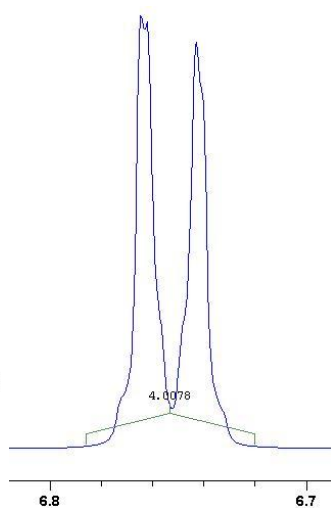
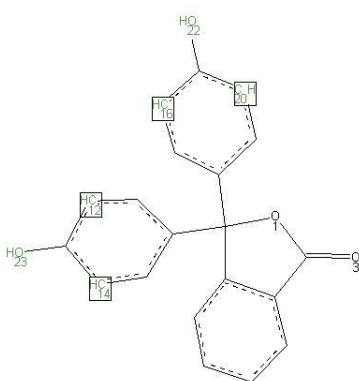


DMSO

e

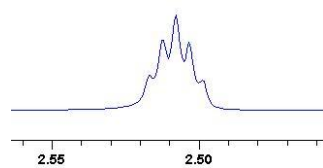
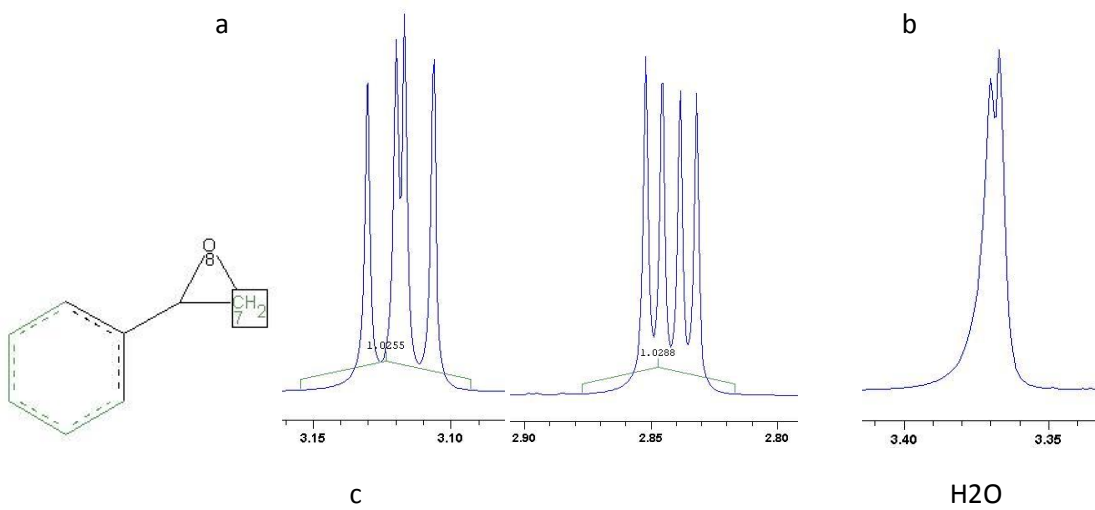
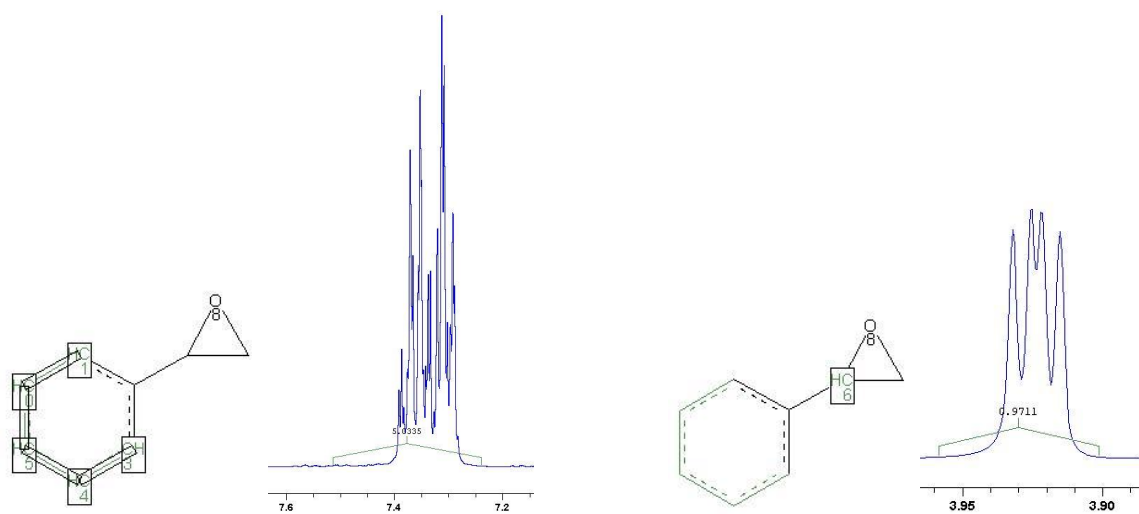
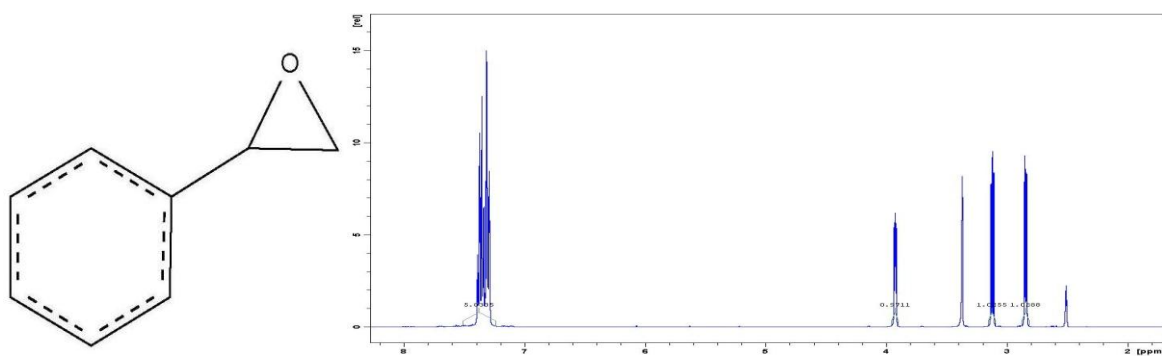


f



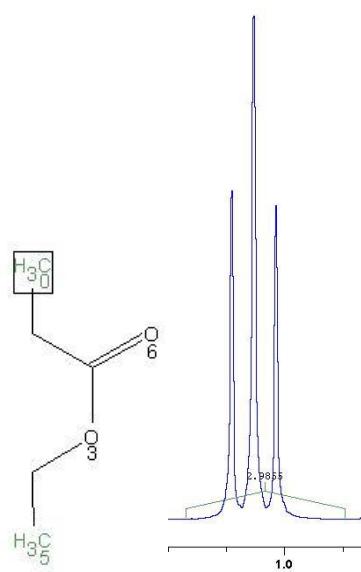
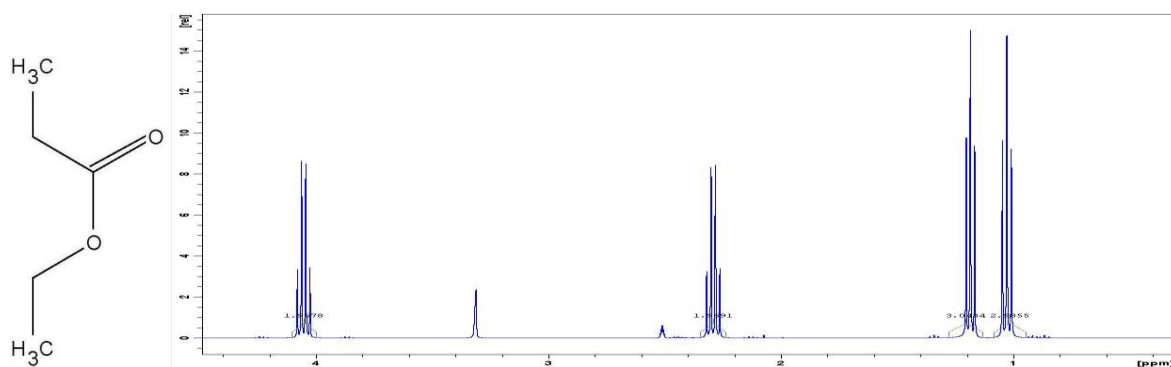
g

## 70. Phenylethylenoxid

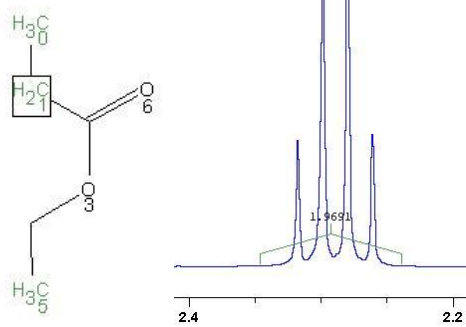


DMSO

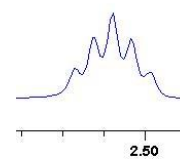
## 71. Propionsaeureethylester



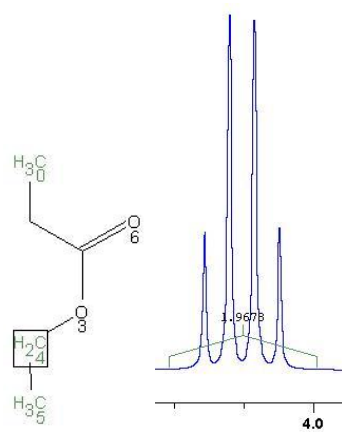
a



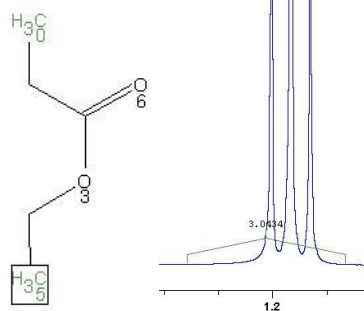
b



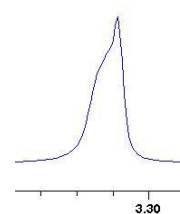
DMSO



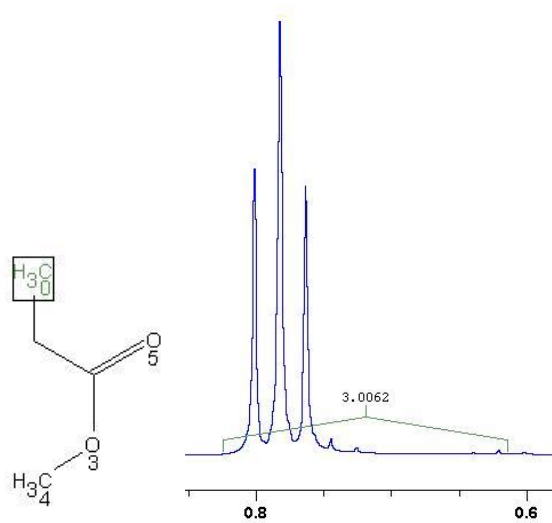
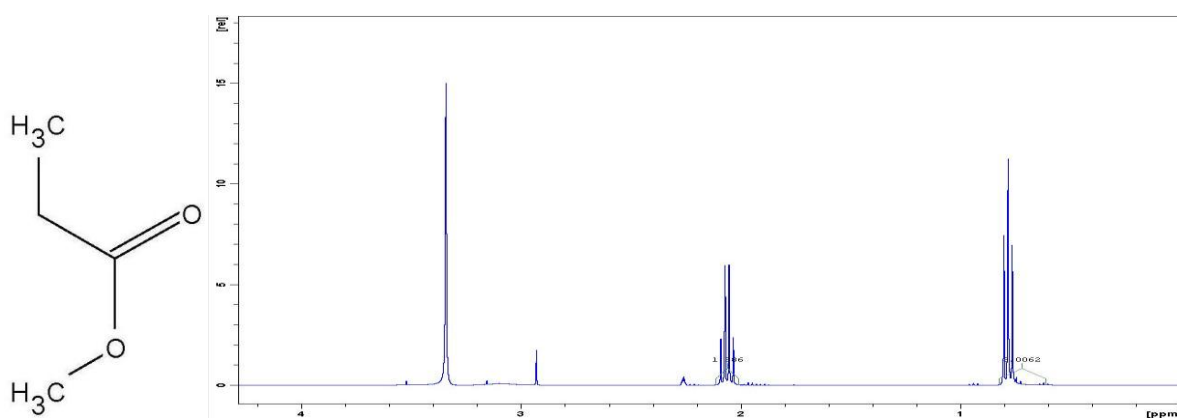
c



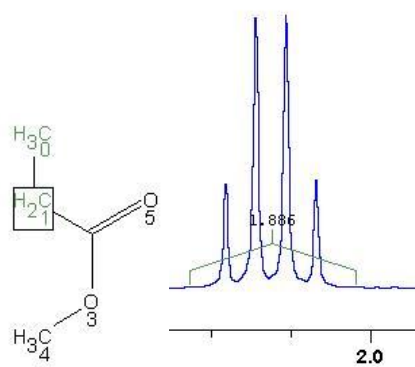
d

H<sub>2</sub>O

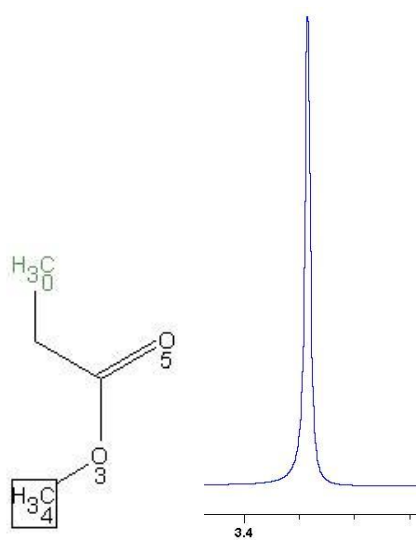
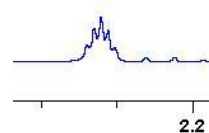
## 72. Propionsaeuremethylester



a

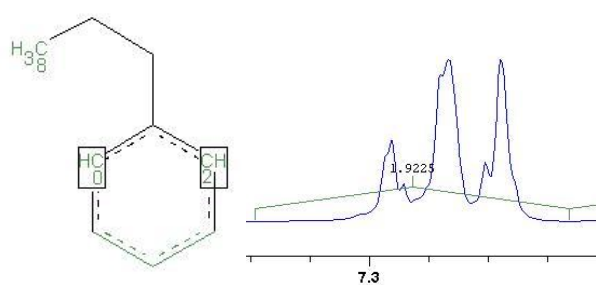
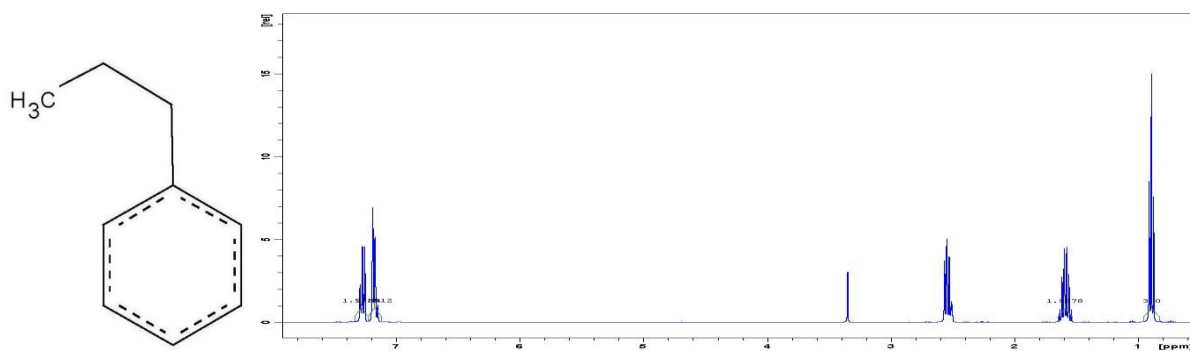


b

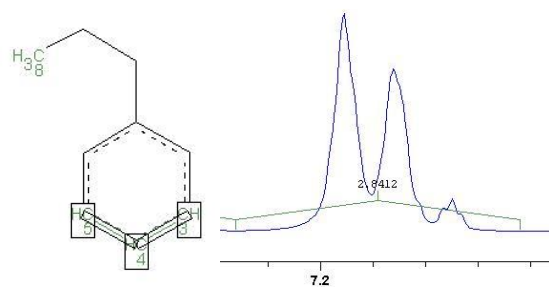
c + H<sub>2</sub>O

DMSO

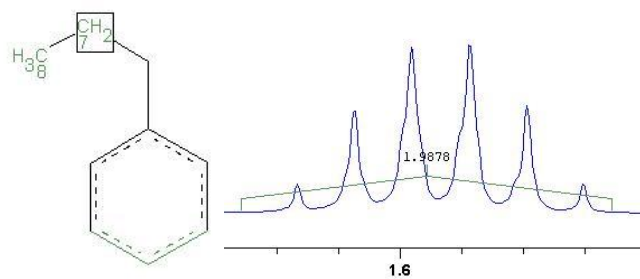
## 73. Propylbenzol



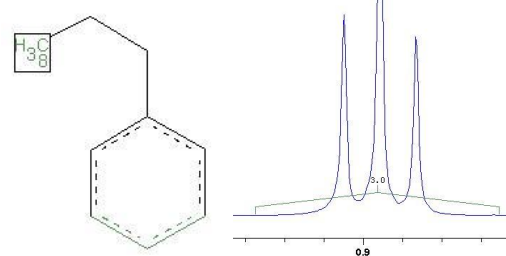
a



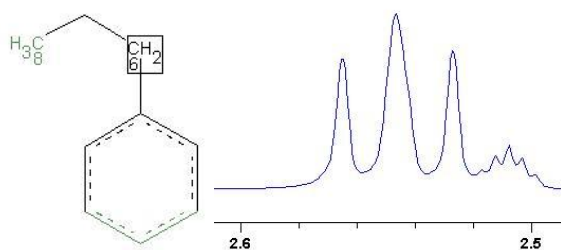
b



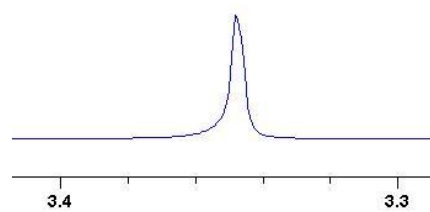
c



d



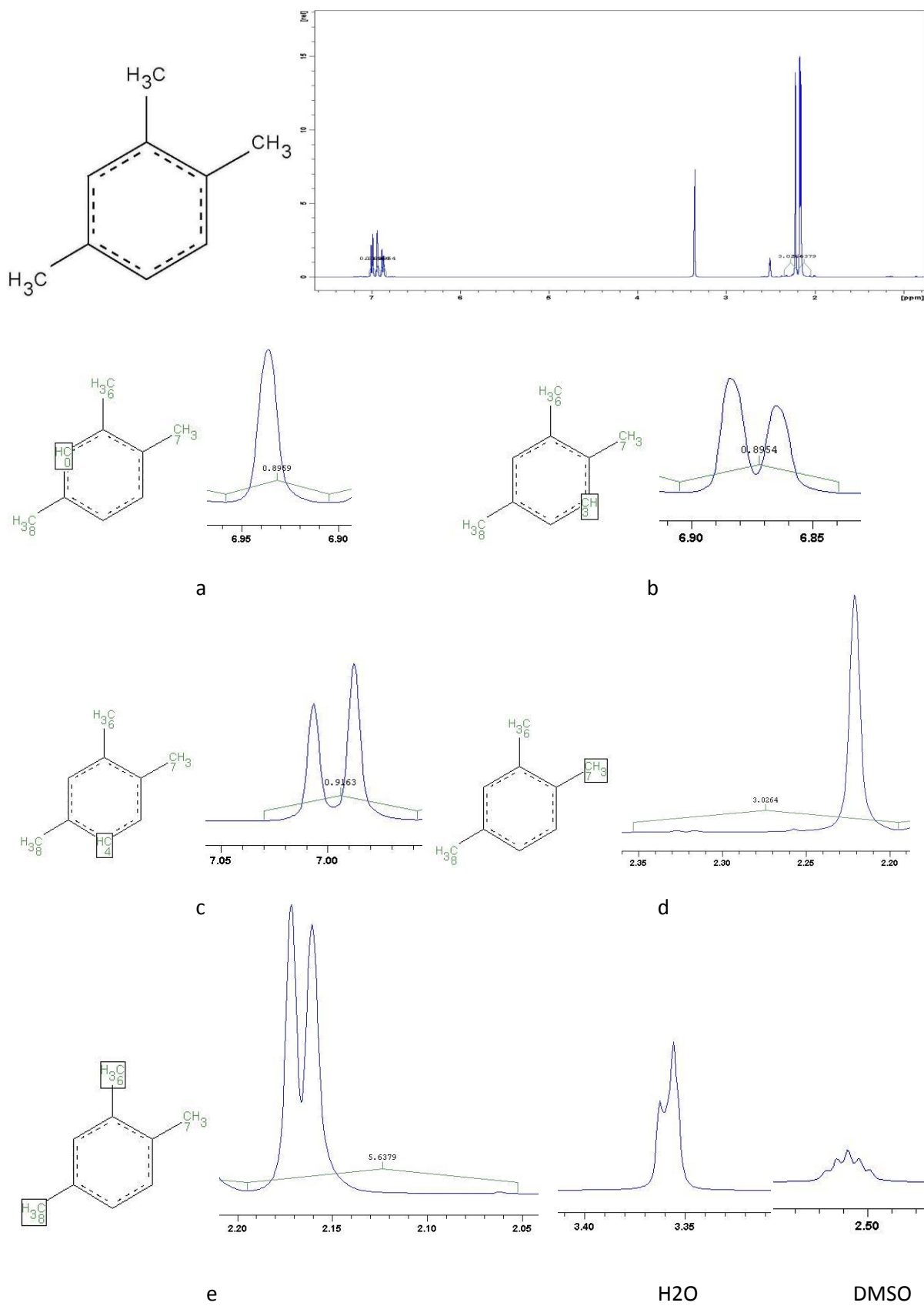
e + DMSO



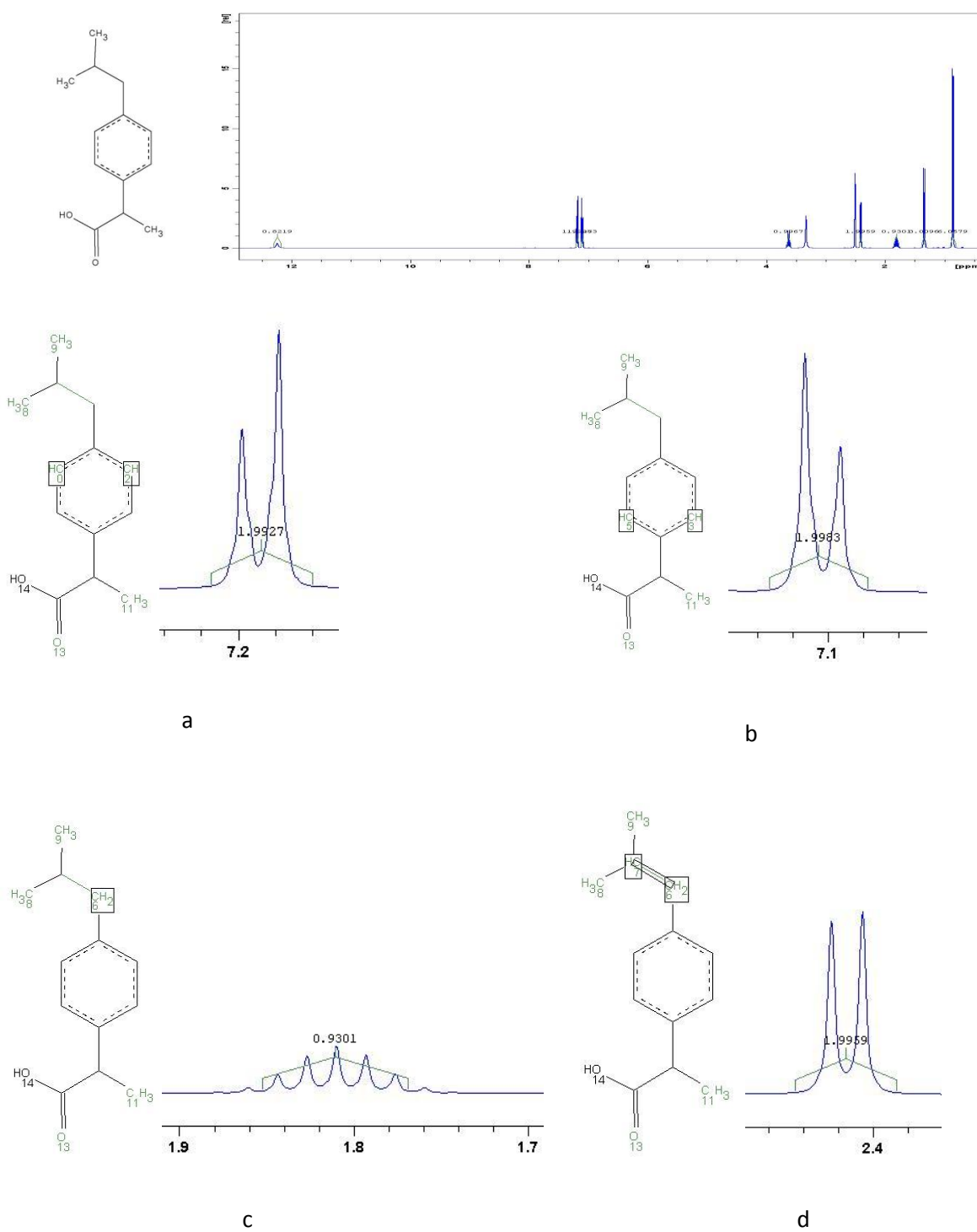
H2O

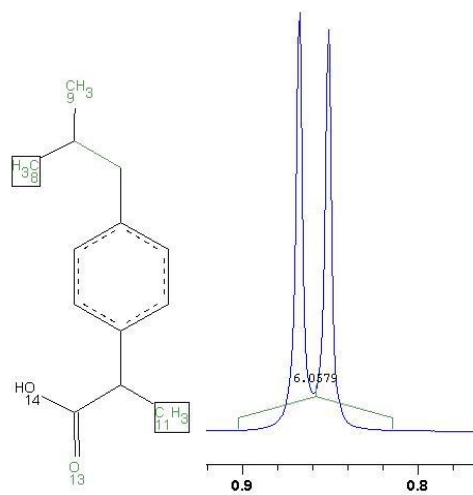


## 74. Pseudocumol

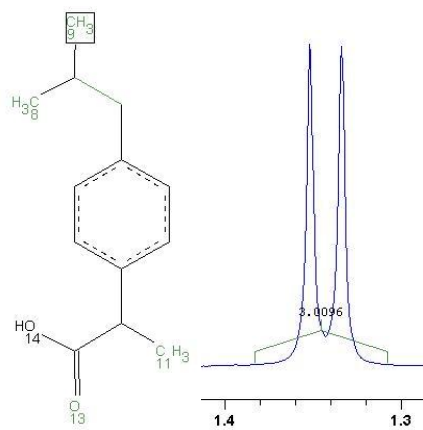


## 75. S+2-4-Isobutylphenylpropionsaeure

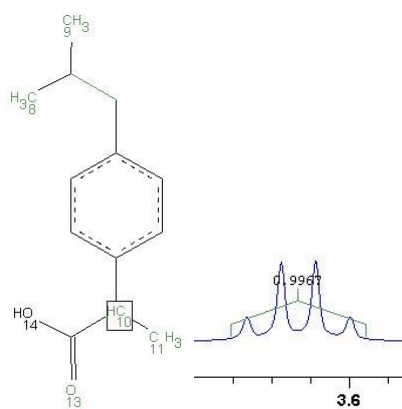




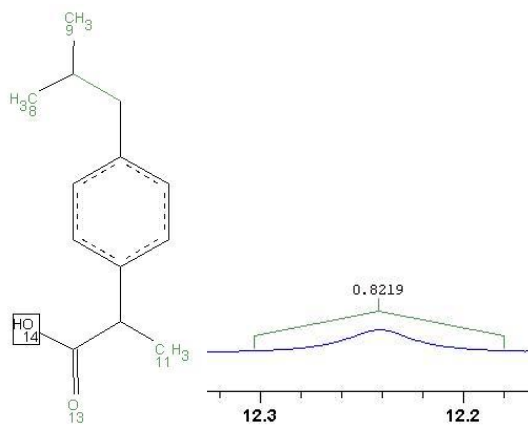
e



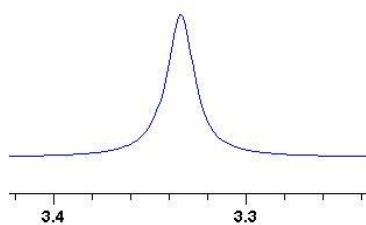
f



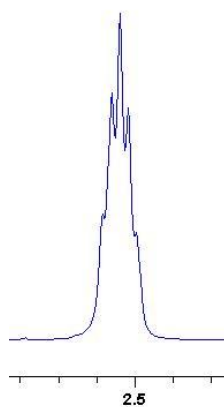
g



h

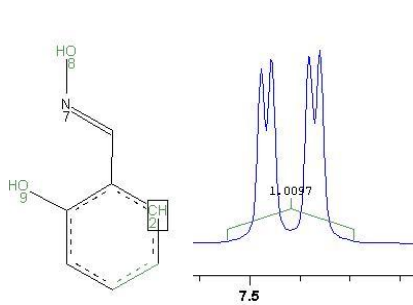
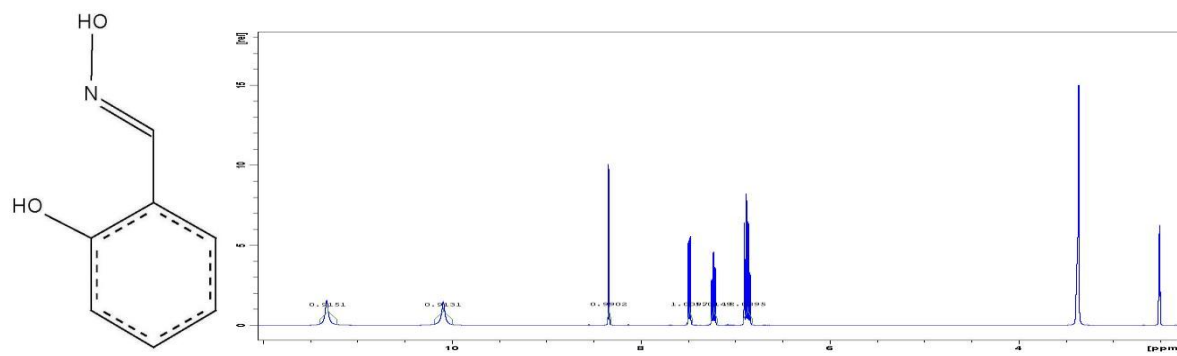


H2O

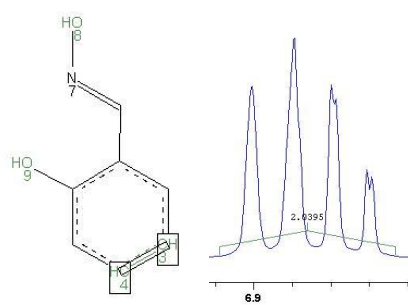


DMSO

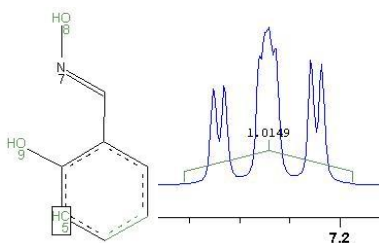
## 76. Salicylaldoxim



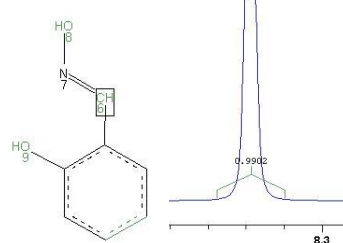
a



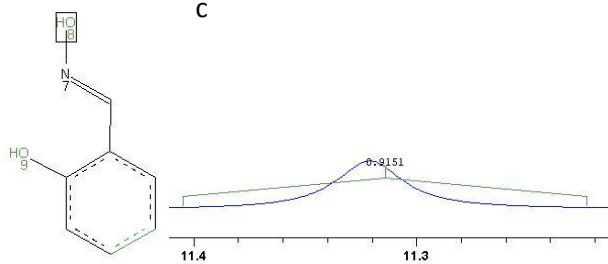
b



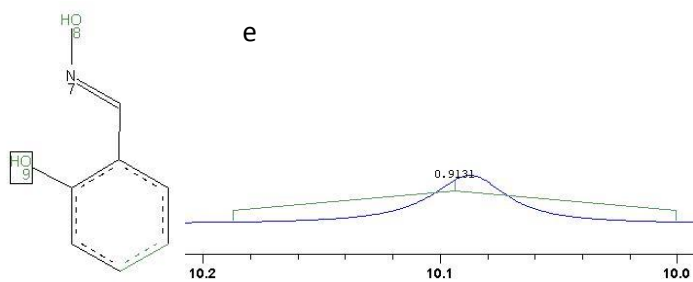
C



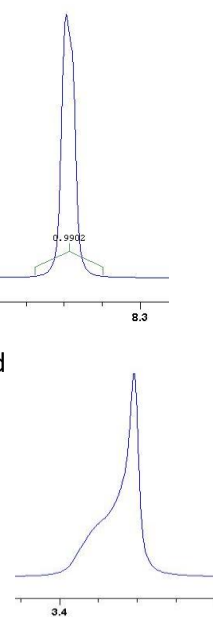
d



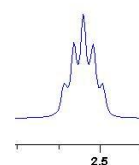
e



**f**

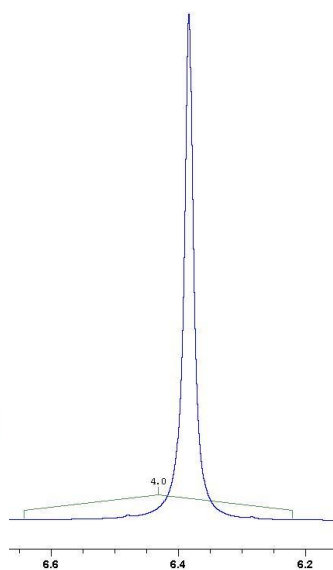
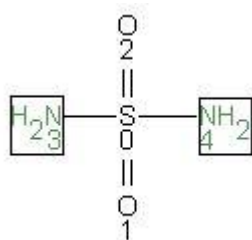
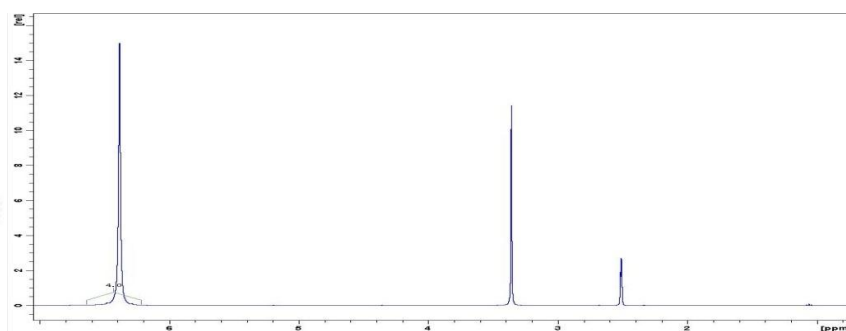
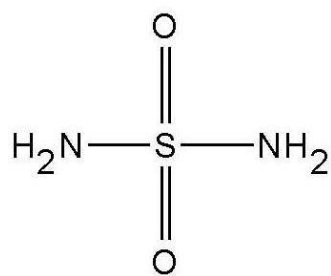


H<sub>2</sub>O

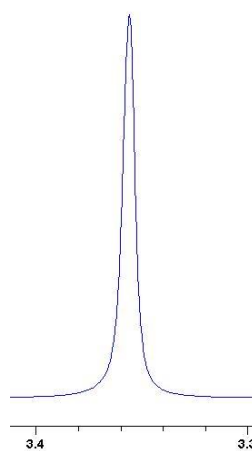
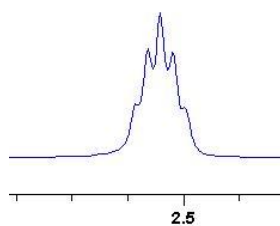


DMSO

## 77. Sulfamid

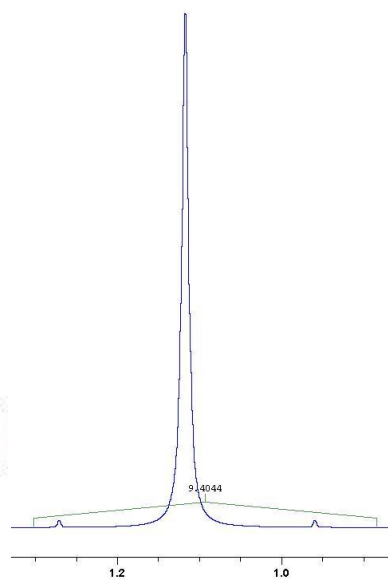
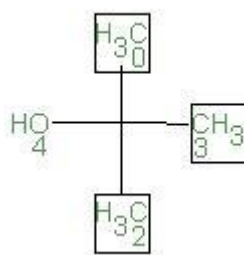
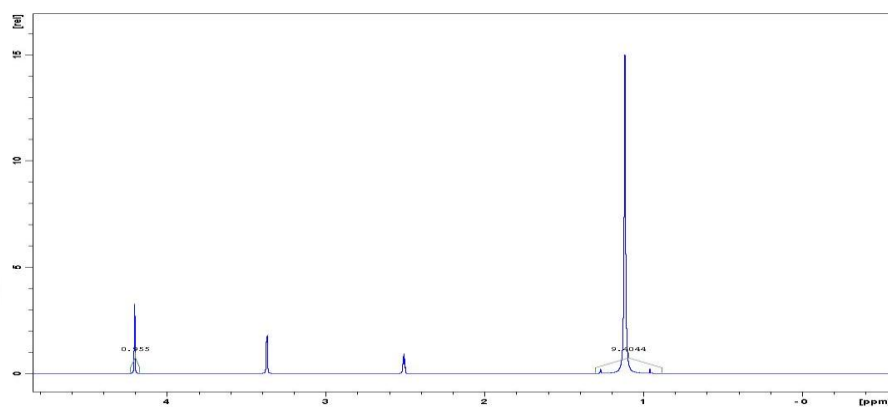
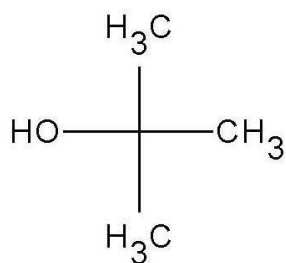


a

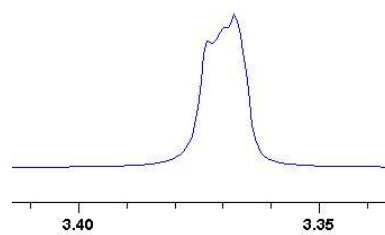
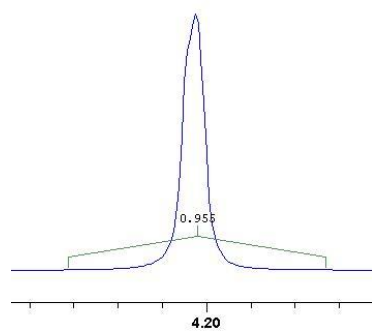
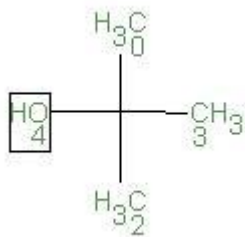
H<sub>2</sub>O

DMSO

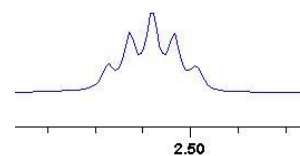
## 78. tert-Butylalkohol



a

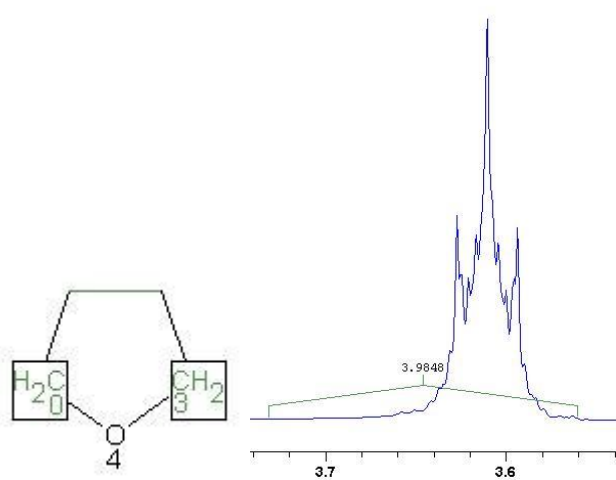
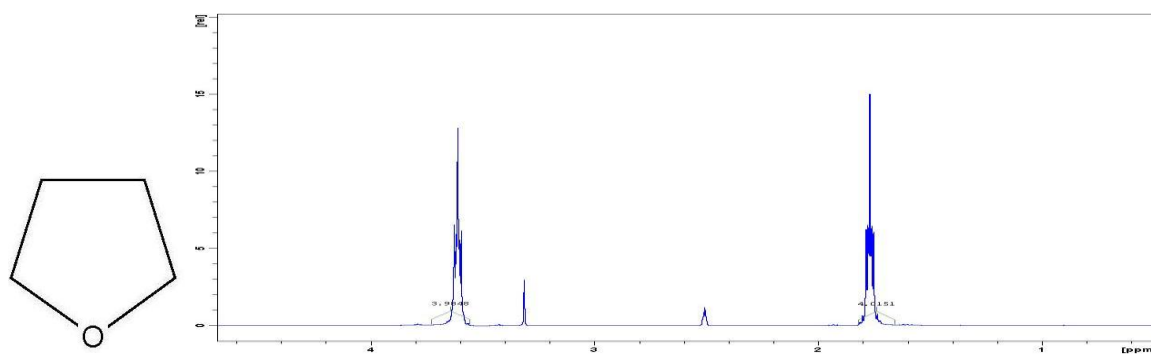
H<sub>2</sub>O

b

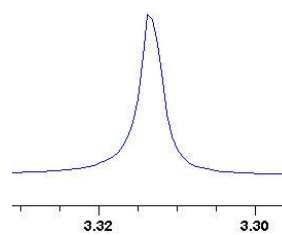
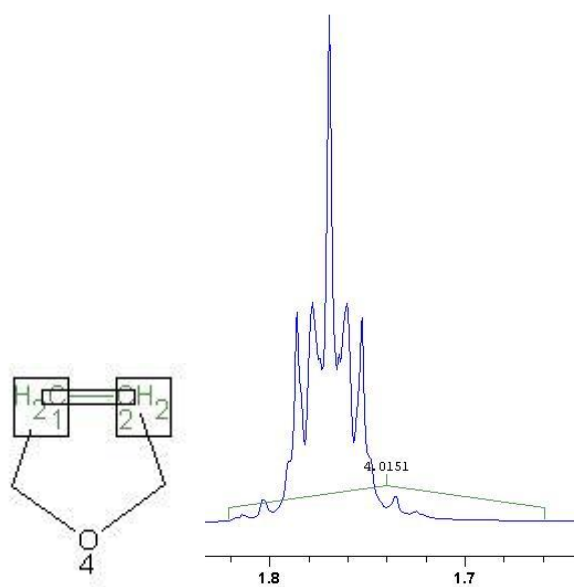


DMSO

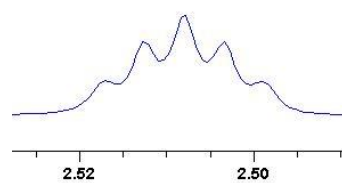
## 79. THF



a

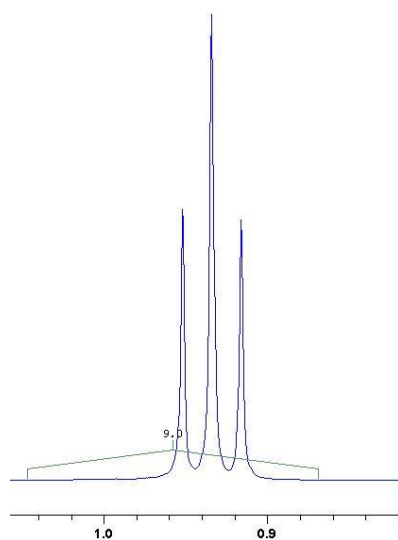
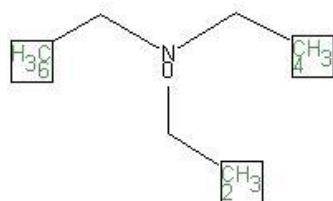
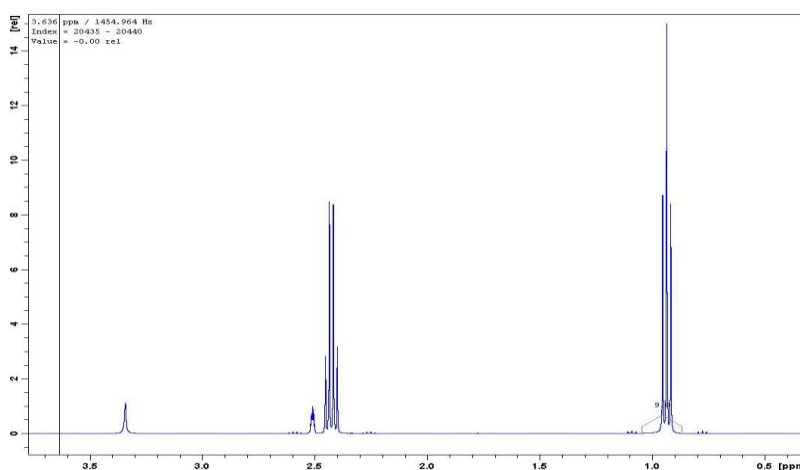
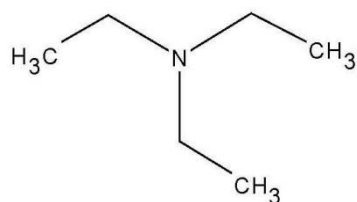
H<sub>2</sub>O

b

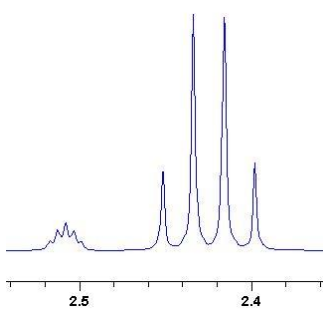
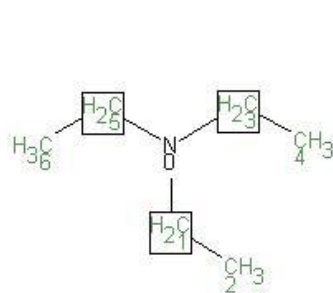


DMSO

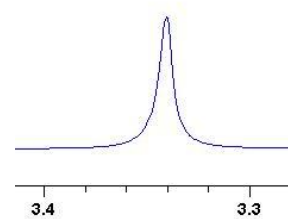
## 80. Triethylamin



a



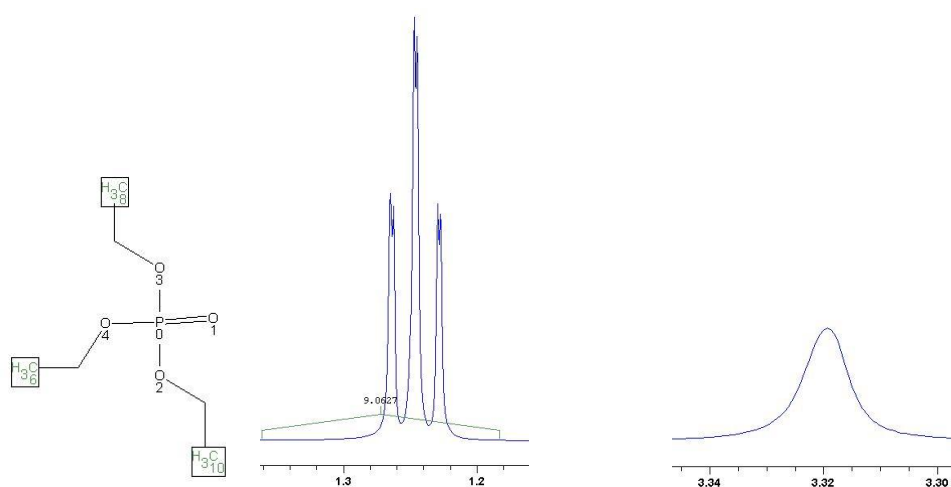
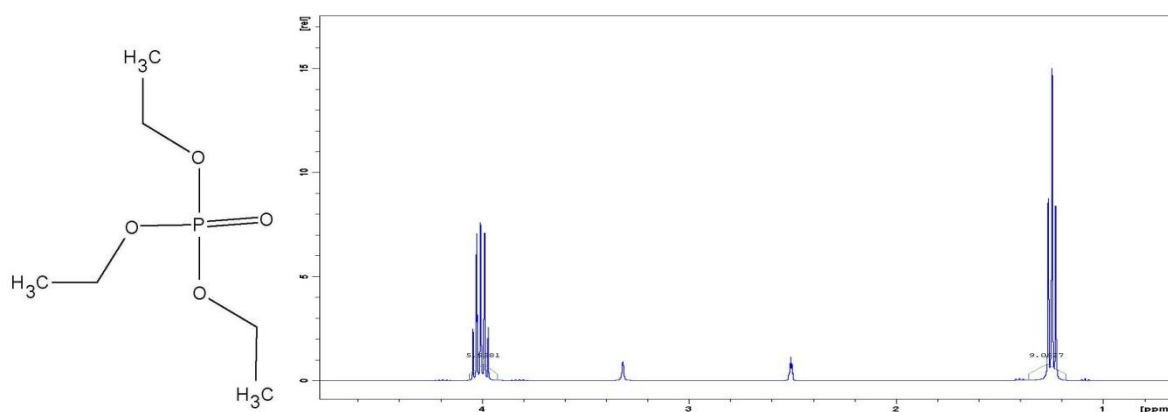
b + DMSO



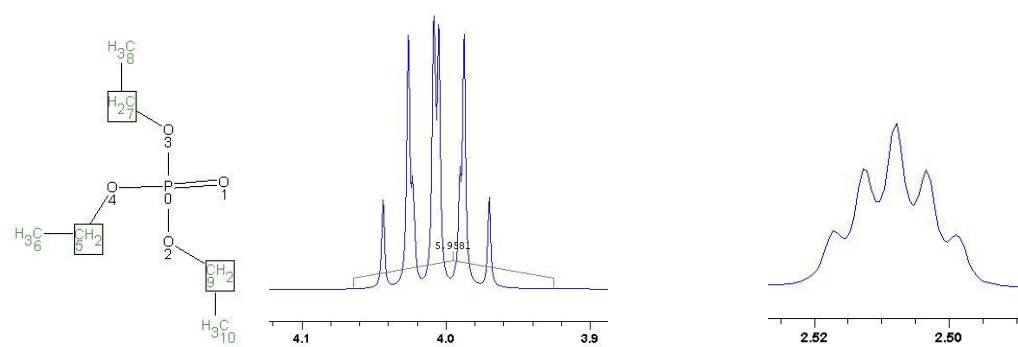
H2O



## 81. Triethylphosphat



a

H<sub>2</sub>O

b

DMSO

Note, in totally 85 consistent spectrum-structure pairs, the system only predict that 81 pairs are consistent. Therefore, here we only list the assignments on 81 pairs where the system predicts that they are consistent.

## D. Curriculum Vitae



### Personal information

First name / Surname **Jiwen Li**

Nationality Chinese

Date of birth 4<sup>th</sup> of March 1974

Gender Male

### Work experience

Dates 22 Jun 04 to 12 Dec 09

Position held Research Assistant

Name and address of employer Department of Informatics, University of Zurich  
Binzmühlestrasse 14, 8050 Zürich (Switzerland)

Dates 3rd Jan 2000 – 30th Sep 2002

Occupation or position held Technical & Q&A Support Engineer

Name and address of employer BD Medical - Pharmaceutical Systems Beijing Office  
7/A Donghuan SQUARE 9 Dong Zhong Street , 100027 Beijing (China)

Dates 5th Sep 1996 – 31st Dec 1999

Position held Technical Support Engineer

Name and address of employer China Iron & Steel Research Institute Group (CISRI)  
No.76 Xueyuan Nanlu,Haidian,Beijing 100081,China

### Education and training

Dates 22nd Jun 2004 – 11th Mar 2010

Title of qualification is going to award Ph.D in Computer Science  
Awarding University The University of Zurich, Zürich, Switzerland

Dates 20th Oct 2002 – 14th Feb 2004

Title of qualification awarded MS.c in Computer Science  
Awarding University The University of Bristol, Bristol, The United Kingdom

Dates 1st Sept 1992 – 31st Jul 1996

Title of qualification awarded B.Sc. in Electric Engineering  
Awarding University Zhejiang University , Hangzhou, China