Development of Python-Based Program to Elucidate Structure from NMR Data

A
dissertation submitted
in partial fulfilment of the requirement for the degree of

Master of Science

in

Chemistry

by

Vanshika Bansal
(Reg.No: 302102020)

Under the guidance of

Dr Manmohan Chhibber
Professor
School of Chemistry and Biochemistry

Dr Ruchika Lamba
Assistant Professor
Department of
Electrical and Instrumentation Engineering

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, PATIALA-147004

July 2023

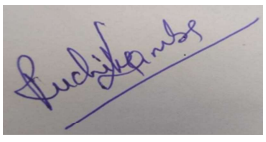
DECLARATION

I hereby declare that the dissertation entitled "**Development of Python-Based Program to Elucidate Structure from NMR Data**" being submitted in the partial fulfilment of the requirements for the award of the degree of **Master of Science in Chemistry** to the School of Chemistry and Biochemistry, Thapar Institute of Engineering and Technology (TIET), Patiala is a record of my own work carried out under the supervision of **Dr Manmohan Chhibber**, Professor, School of Chemistry and Biochemistry, TIET and **Dr. Ruchika Lamba**, Assistant Professor, Department of Electrical and Instrumentation Engineering from January -July, 2023. Further, any work of this dissertation has not been submitted to any University for the ward of any other degree or diploma.

Vanshika Bansal
(Reg. No: 302102020)
Signature of Candidate

Place: Patiala

This is to certify that the above statement made by the candidate is correct and true to the best of our knowledge



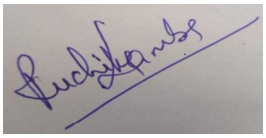
Dr Manmohan Chhibber
Professor
School of Chemistry and Biochemistry

Dr Ruchika Lamba
Assistant Professor
Department of
Electrical and Instrumentation Engineering

CERTIFICATE

This is to certify that the dissertation entitled "**Development of Python-Based Program to Elucidate Structure from NMR Data**" being submitted by **Ms. Vanshika** to the School of Chemistry and Biochemistry, Thapar Institute of Engineering and Technology, Patiala in partial fulfilment of the requirements for the award of the degree of **Master of Science in Chemistry**, is an authentic record of the work carried out by the candidate under our guidance and supervision. She has fulfilled the requirements for submitting this dissertation, which to my knowledge has reached the requisite standard.

The results embodied in the dissertation have not been submitted in part or whole to any other University or Institute for the award of their degree or diploma.



Dr Manmohan Chhibber
Professor
School of Chemistry and Biochemistry

Dr Ruchika Lamba
Assistant Professor
Department of
Electrical and Instrumentation Engineering

ACKNOWLEDGEMENT

First and foremost, I extend my heartfelt gratitude to Almighty God for His abundant blessings, and the opportunities that have enabled me to successfully accomplish my thesis.

I am deeply indebted to **Dr. Satnam Singh**, Professor and Head of the School of Chemistry & Biochemistry, for providing me with the invaluable opportunity to explore the realm of research through this dissertation.

I wish to express my sincere appreciation to my supervisors**, Dr Manmohan Chhibber** and **Dr Ruchika Lamba**, for their constant encouragement, constructive feedback, and active involvement throughout my master's thesis.

A special acknowledgment goes to my python instructor, **Jose Portilla**, and the platform **Udemy**, for upskilling me. I extend my heartfelt thanks to my uncles, **Mr. Arpit Gupta** (Data Architect at AWS) and **Mr. Sunil Garg** (Development Manager- Digital Transformation at FIS), for generously sharing their time and valuable input to encourage me by suggestions and constructive criticism in shaping the outcome of my work.

I cannot overlook the support of my parents, whose love and encouragement have been a driving force behind my achievements. They are my ultimate role models. Lastly, I want to extend a special appreciation to my sister, Pearl, for her unconditional love and inspiration. Her belief in me has been a constant source of motivation throughout this journey.

I am also grateful to the Thapar Institute of Engineering and Technology (TIET), Patiala, for providing me with an exceptional learning environment and granting me the opportunity to successfully complete my studies, earning a master's degree in science.

Vanshika Bansal
(Reg. No: 302102020)
Signature of Candidate

Place: Patiala

ABSTRACT

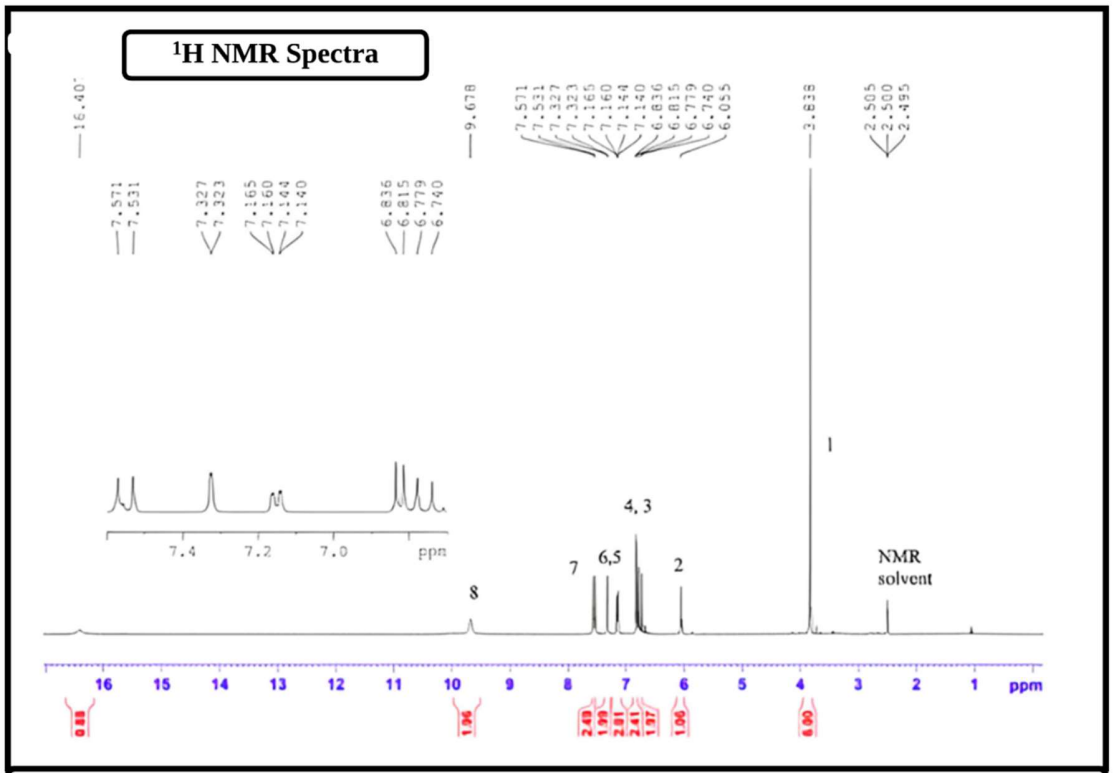
This work aims to broaden the accessibility of structure determination using NMR data beyond chemistry experts. The objectives of this work include gathering comprehensive NMR data, developing an algorithm to analyze and compare the entered data with a reference database library, and programming the desired output in the form of a chemical structure. A Python-based program has been developed to determine the structures of organic compounds by inputting the textual format of NMR data for a given set of molecules. Through a library, the program analyse the data, providing output. The program offers a user-friendly interface with a single window display of precise molecule structures. Although the current library includes only 14 molecules, the program can be expanded to handle larger datasets. Future versions could incorporate additional components such as coupling constants and other types of NMR spectra and compatibility with a wider range of solvents. One gap in research is the lack of a program that directly reads and processes NMR data to generate the corresponding structure. Developing such a program would enable students, researchers, and publishers to verify the structure of organic compounds from published data, ensuring accurate interpretation of NMR data.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Determining the structure of molecules plays a pivotal role in organic chemistry, medicinal chemistry, material science and many others. Nuclear Magnetic Resonance (NMR) stands out as a powerful tool for structure elucidation and other studies. While ^{1}H NMR alone is often sufficient to predict and determine molecular structures, its accuracy and reliability can be enhanced by utilizing complementary techniques such as infrared spectroscopy (IR), mass spectrometry (MS) and ^{13}C NMR.



^{1}H NMR Data in text form

(400 MHz, DMSO-*d*6, TMS): δ 16.47 (br, 1H, Enol-H), 9.66 (br, 2H, -OH), 7.55 (d,*J*= 15.8 Hz, 2H, Vinyl-H), 7.32 (d,*J*= 1.89 Hz, 2H, Aryl-H), 7.15 (dd,*J*= 8.2; 1.93 Hz, 2H, Aryl-H), 6.82 (d,*J*= 8.13 Hz, 2H, Aryl-H), 6.75 (d,*J*= 15.81 Hz, 2H, Vinyl-H), 6.06 (s, 1H, Methine-H), 3.84 (s, 6H, -OCH3)

Figure- 1.1 (Shows ^{1}H NMR spectra and ^{1}H NMR data of a complex compound)

Nowadays, several software applications elucidate the structure of almost all organic compounds from the NMR spectra and vice versa. The NMR data is usually written in a particular format for records or publishing in manuscripts, as shown in **Figure-1.1**. However, to conclude the structure of an organic compound by interpreting the text of NMR data demands both time and expert knowledge of the subject. Consequently, extracting the molecular structures from the NMR data text becomes arduous, making it relatively inaccessible to other domain scientists like biotechnologists, engineers, physicians, physicists and forensics experts. Thus, a pressing need exists to create a software solution that, upon inputting NMR data, automatically generates the corresponding molecular structure.

The present work broadens the accessibility of structure determination using NMR data, beyond the realm of chemistry experts. A program has been developed using Python computer language that focuses on the structure determination of organic compounds by inputting the textual format of the NMR data for a given set of molecules.

CHAPTER 2

LITERATURE REVIEW

The fields of chemistry and computer programming are quite related since early nineteen century The valence bond theory (1927)[1], density function theory (1964)[2,3], autodock (1989)[4]. are some of the reference in time and domain that describe the above relation and its evolution. With the surge in the field of computer science, a number of programs have been reported in literature recently, following paragraphs describe the development in the field in context of present work.

2.1 SMILES - Simplified Molecular Input Line Entry System (1987)[4]

SMILES was perhaps the first algorithm developed by D Weininger in late 1980s. The introductory methodology and encoding followed an algorithm to generate a unique SMILES notation that depicted graphical structure structures as shown in **Figure-2.1**.One of the advantage of SMILES is that it uses compact notation well suited for high speed machine processing for many chemical computer applications.

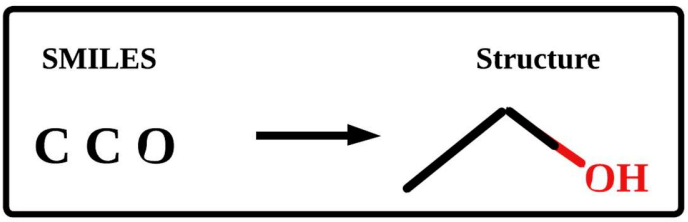


Figure-2.1 (SMILES Code for Ethanol and corresponding structure)

2.2 NMR View - A computer program for the visualization and analysis of NMR data (1994)[5]

The purpose of NMR View is to provide researchers and scientists with an integrated tool for analyzing NMR (Nuclear Magnetic Resonance) data. The software includes a robust peak picker that identifies peak position, shape, width, and intensity. It enables interactive deletion of spurious peaks. NMR View supports the addition of new capabilities through the TCL command language. It has specific commands for assigning NOESY peaks (which can use 2D, 3D, or 4D

criteria for matching the chemical shift assignment data) and facilitates iterative structure refinement. NMR View allows for the visualization, extraction, and analysis of multiple data sets, It also facilitates refinement of molecular structures, and automation of spectral analysis processes. Overall, NMR View aims to enhance the efficiency and accuracy of NMR data analysis.

2.3 StrucEluc expert software system for the structure elucidation of Natural Products from 1D and 2D NMR Data[8,9]

First-generation expert systems use only 1D NMR data were limited to molecules containing about 25 skeletal atoms. Second-gen systems are available using 2D NMR data such as SESAMI[6], CHEMICS[7], and StrucEluc[8,9]. The purpose of StrucEluc is to help in the process of structure elucidation, which involves determining chemical structure of a compound based on spectroscopic data. It analyzes various types of spectroscopic data, including 1D and 2D NMR, IR, MS, and ^{13}C spectral data. The program require one possible molecular formula as input. It has a (GUI). The program analyzes^{13}C and ^{1}H NMR chemical shifts and generates correlation tables from fragments and structures in the database. The generated structures can be verified using spectral filtering and spectrum-structure correlation libraries. The program checks for structure identity and ranks structures based on the average deviation (dF) between estimated spectra and experimental 1D spectrum, with the correct structure typically appearing at the top of the ranked list. In addition to spectroscopic data, StrucEluc can incorporate supplementary information such as ^{1}H NMR chemical shifts, full ^{1}H NMR spectra, assignable fragments from mass spectra, and physiochemical properties.

2.4 CASE(Computer-assisted structure elucidation) program (2019)[10]

The purpose of CASE (Computer-Assisted Structure Elucidation) programs is to assist in the process of determining the chemical structure of a compound based on available NMR data. These programs analyze NMR spectra, such as COSY and HMBC spectra, and use coupling constants and chemical shift correlations to generate and rank potential structures (based on the probabilities calculated from available NMR data). The goal is to identify the most likely structure or a list of

probable structures, taking into account empirical formulas and experimental NMR data (as input). CASE programs help researchers overcome challenges such as violations of coupling constant assumptions and low proton-to-carbon ratios, which can make structure determination difficult.

Some CASE programs are listed below.

1. ACD Labs Structural Elucidator[8,10,11,12]

ACD Labs Structure Elucidator is a software tool developed by Advanced Chemistry Development (ACD) Labs for determining the structure of complex organic molecules using NMR data. The software requires several types of input data to perform its analysis. These include the empirical formula of the molecule, ^{13}C chemical shifts, as well as HSQC and HMBC data. Additionally, COSY data can be optionally provided. The program can also handle other 2D NMR data and import 1D and 2D FID (Free Induction Decay) data from various spectrometers. The program initially processes the input data and generates a list of possible structures that are consistent with the provided information. To make a more definitive choice among the generated structures, ACD Labs scientists recommend utilizing DFT calculations of ^{13}C and ^{1}H chemical shifts. **Figure-2.2** MCD of plebeianiol.

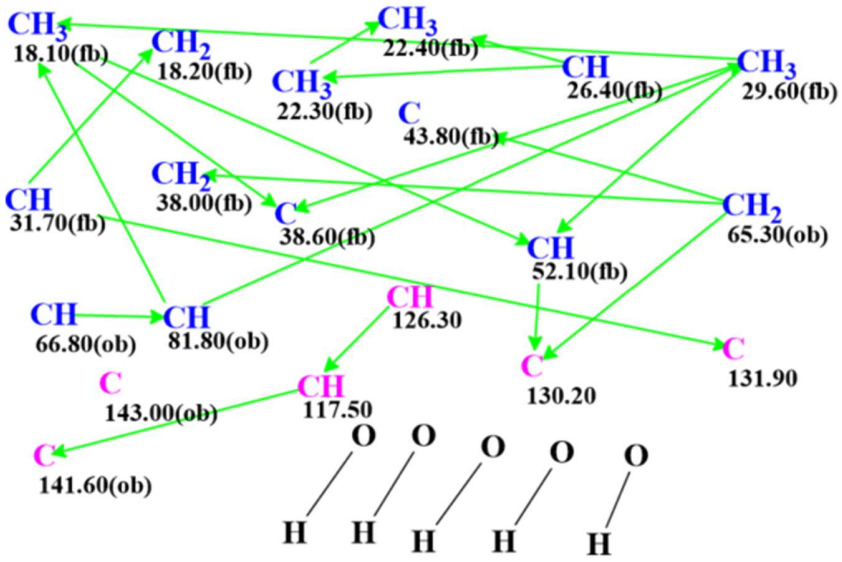


Figure- 2.2 (Molecular connectivity diagram (MCD) of plebeianiol A. Hybridizations of carbon atoms are marked by the corresponding colours : sp2 - violet, sp3 - blue. Labels "ob" and "fb" are set by the program to carbon atoms for which neighbouring with heteroatom is either obligatory (ob) or forbidden (fb). HMBC connectivities are marked by green arrows.)[13]

2. Bruker CMC-se program[14,15,16]

Bruker CMC-se is a software program developed by Bruker Corporation for the structure elucidation of unknown compounds. It was first introduced in 2012. The program requires a minimum input of the molecular formula and ^{1}H, HSQC, and HMBC spectra. Additional 2D spectra and correlation data can also be included to enhance the analysis. It generates a list of potential structures that are consistent with the provided information. It has ability to handle suspected long-range correlations. It provides options for user-designated numbers of allowed long-range COSY and/or HMBC correlations. To determine the most probable structure, Bruker CMC-se compares calculated chemical shifts for different structural possibilities with the experimental values obtained from the spectra. By assessing the level of agreement between calculated and experimental chemical shifts, the program assists in identifying the most likely structure among the generated candidates. All the process done by CMC-se is shown in **Figure-2.3.**

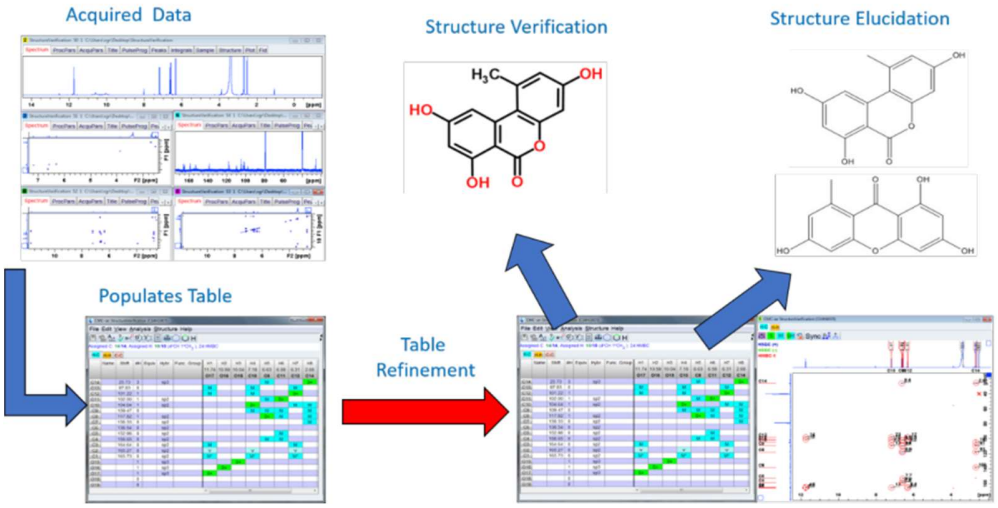


Figure- 2.3 (CMC-se supports both structure elucidation as well as structure verification workflows.)[24]

3. MestrelabMNova structure elucidation program[17,18]

MestrelabMNova Structure Elucidation is a software program developed by Mestrelab Research for the purpose of confirming structural assignments of natural products and assisting in the determination of new natural product structures. The recommended input for MNova CASE studies within the program includes FIDs corresponding to 1H, 13C, HSQC, HMBC, and COSY spectra. In the newer version of the program, previously ignored peaks can now be used for

structure generation. MNova also offers an option to automatically search for and eliminate a specified number of HMBC correlations. This feature assists in refining the list of potential structures by considering the long-range correlations between atoms, which can be particularly important in complex natural product structure

4. Nuzillard's LSD program[16,19,20]

   Nuzillard's LSD (Logic for Structure Determination) program is a software tool developed by Professor Jean-Luc Nuzillard for the purpose of structure elucidation in organic chemistry, specifically focusing on the analysis of natural products. Recommended input, includes 1H NMR, 13C NMR, COSY, HSQC, and HMBC data. Additionally, the program can handle other 2D data that may be available.The LSD program incorporates substructures specifically designed to aid in the elucidation of natural products. LSD program is freely available.

5. CSEARCH[21]

   CSEARCH is a non-commercial program used for dereplication, which is the process of determining whether a compound is known or new to avoid duplicating efforts on known compounds. The program works by comparing the 13C spectrum of an unknown compound with the spectra of known compounds in the database. The primary purpose of CSEARCH is to assist in the identification and characterization of organic compounds

2.5 DP4-AI automated NMR data analysis: straight from the spectrometer to structure(2020)[22]

DP4-AI (Dereplication and Prediction by Artificial Intelligence) is an automated NMR data analysis tool that utilizes artificial intelligence techniques to predict relative stereochemistry and assist in structure determination. DP4 analysis, stands for Discrete Probability (DP4) analysis, With DP4-AI, the process of DP4 calculation has been streamlined using the PyDP4 Python script. The program requires the molecule structure and assigned experimental 1D NMR spectra from the user to perform the analysis. DP4-AI incorporates an assignment algorithm (AA) that assigns atoms to observed peaks using predicted chemical shifts. This algorithm takes into account the variability of NMR spectra and allows for user-defined adjustments, such as adjustments to the DFT solvent model and integration of multiplets in ^{1}H spectra. The

performance of DP4-AI, has been evaluated using a test set of 47 diverse molecules with multiple stereocenters. All the processes executed by this program are shown in **Figure-2.4.**

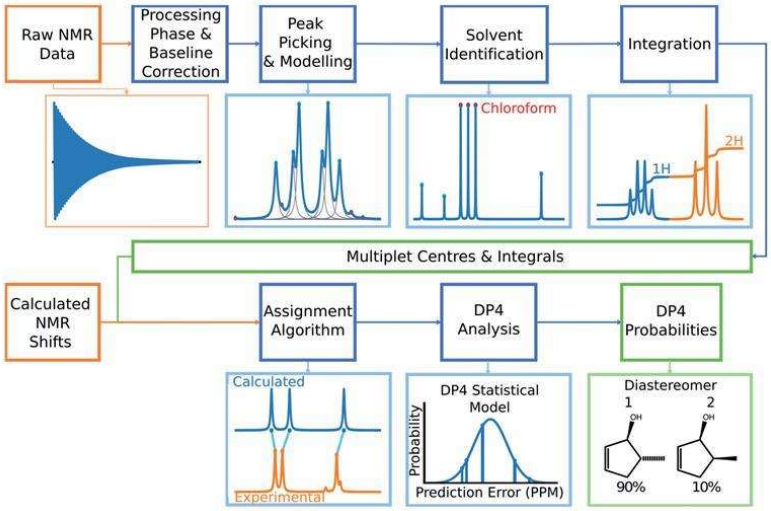


Figure-2.4 (DP4-AI processes raw NMR data in a series of stages to yield experimental multiplet shift values and their integrals. The program then takes shifts calculated using DFT for each atom in the molecule and assigns them to the experimental peaks. This assignment is then used to calculate a DP4 probability for each diastereomer.)[22]

2.6 SHERLOCK(2023)[23]

SHERLOCK is a free and open-source CASE software designed for the structure elucidation of small organic molecules using 1D and 2D NMR ^{1}H, ^{13}C, DEPT, HSQC, HMBC, and COSY spectra along with the given molecular formula determined by mass spectrometry (MS), to form the foundation for structure generation. The software provides a versatile GUI that allows the user to control various operations in the structure elucidation process. The software takes into account the substructures and restrictions derived from the NMR data and can generate all possible structures that satisfy the given constraints.

CHAPTER 3

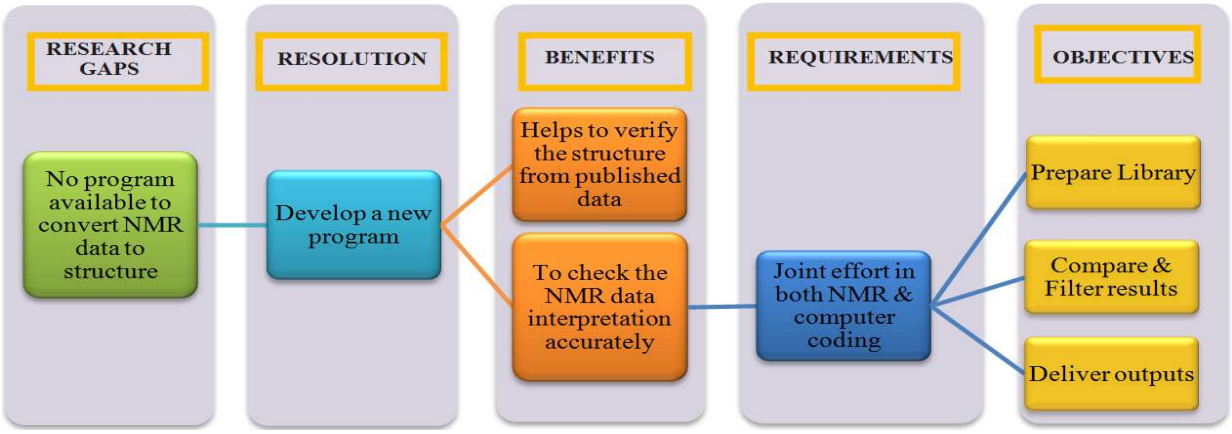
GAPS IN RESEARCH & OBJECTIVES

3.1 Gaps in Research

The accurate interpretation of NMR spectra is vital in structure elucidation. However, as discussed in previous sections, several software can predict the structure of organic compounds based on the NMR graph or vice versa. Some programs employ data from 1D, 2D NMR, and other spectroscopic techniques to deduce the structure. But no program is currently available that can read or process NMR data directly into its corresponding structure. The development of such a program can enable students, researchers, and publishers to verify the structure of organic compounds from the published data. This will enable the user to check the NMR data interpretation accurately. It is crucial that while developing such a program care should be taken that it has a user-friendly interface intuitive and easy to navigate. The development of such a program requires collaborative knowledge and expertise in both NMR spectroscopy and computer coding language. The present work has developed a Python-based program to elucidate structure from NMR Data.

3.2 Objectives

- To gather, compile and organize comprehensive NMR data for selected molecules.
- To develop/write an algorithm to analyze and compare the entered data with the reference database library and filter the results.
- To program the desired output in the form of a chemical structure.



CHAPTER 4

METHODOLOGIES

4.1 General: The algorithm was written in Python, and NMR data was compiled in Excel. Pandas Library was used to read, filter, and extract information. The output was expressed in SMILES language and tagged to " http://hulab.rxnfinder.org/smi2img/ " for display in the output window.

4.2 Data Collection: A set of ten alcohols and hydrocarbons, mostly homologous, was shortlisted and their ^{1}H NMR Graphs collected [25]. from https://docbrown.info/page06/spectra/0spectra-nmr1h.htm. The NMR data with the number of signals, number of equivalent protons, chemical shift, their splitting, and output in terms of SMILES language was compiled and organized in different columns of an Excel sheet as shown in **Table-4.1.**

Table 4. 1 Format of ^{1}H NMR of data for ten homologous alcohols and hydrocarbons with SMILES output

Molecule	number_signal	number_equivalent_protons	chemical_shift	splitting	output	Finaloutput
Methane	1	4	0.9	s	C	C
Methanol	2	3	3.43	s	C	CO
Methanol	2	1	3.66	s	O	CO
Isopropanol	3	6	1.2	d	C	CC(C)O
Isopropanol	3	1	4.01	m	C(C)	CC(C)O
Isopropanol	3	1	2.16	s	O	CC(C)O
Butan-2-ol	5	3	1.17	d	C	CC(O)CC
Butan-2-ol	5	1	3.71	m	C	CC(O)CC
Butan-2-ol	5	1	2.37	s	(O)	CC(O)CC
Butan-2-ol	5	2	1.46	m	C	CC(O)CC
Butan-2-ol	5	3	0.93	t	C	CC(O)CC
2-methylpropan-2-ol	2	9	1.26	s	CC(C)(C)	CC(C)(C)O
2-methylpropan-2-ol	2	1	2	s	O	CC(C)(C)O
2,2-dimethylpropane	1	12	0.9	s	CC(C)(C)C	CC(C)(C)C
2-methylpropane	2	1	2	m	C	C(C)(C)(C)
2-methylpropane	2	9	0.8	d	(C)(C)(C)	C(C)(C)(C)
2-Methylbutan-1-ol	5	3	0.93	t	C	CCC(C)O
2-Methylbutan-1-ol	5	2	1.46	m	C	CCC(C)O
2-Methylbutan-1-ol	5	1	3.71	m	C	CCC(C)O
2-Methylbutan-1-ol	5	3	1.17	d	(C)	CCC(C)O
2-Methylbutan-1-ol	5	1	2.37	s	O	CCC(C)O
Cyclopropane	1	6	0.2	s	C1CC1	C1CC1
1,2-dioxane	2	4	3.55	t	O1CC	O1CCCCO1
1,2-dioxane	2	4	1.45	t	CCO1	O1CCCCO1

Ethoxyethane	2	6	1.21	t	C(C)	C(C)OCC
Ethoxyethane	2	4	3.47	q	OC(C)	C(C)OCC
Butanol	5	3	0.94	t	C	CCCCO
Butanol	5	2	1.39	m	C	CCCCO
Butanol	5	2	1.53	m	C	CCCCO
Butanol	5	2	3.63	t	C	CCCCO
Butanol	5	1	2.24	s	O	CCCCO
2-methoxypropane	3	6	1.2	d	CC	CC(C)OC
2-methoxypropane	3	1	3.7	m	(C)O	CC(C)OC
2-methoxypropane	3	3	3.2	s	C	CC(C)OC
1-methoxypropane	4	3	3.337	s	CO	COCCC
1-methoxypropane	4	2	3.336	t	C	COCCC
1-methoxypropane	4	2	1.59	m	C	COCCC
1-methoxypropane	4	3	0.93	t	C	COCCC

4.3 Graphical User Interface:

Using Python as a programming language a graphical user interface was created to facilitate the user to enter data as shown in **Fig 4.1**. The algorithm was featured with a self-validation algorithm to check the authenticity of the entered data and suggest a correction in case of error as shown in **Fig 4.2**

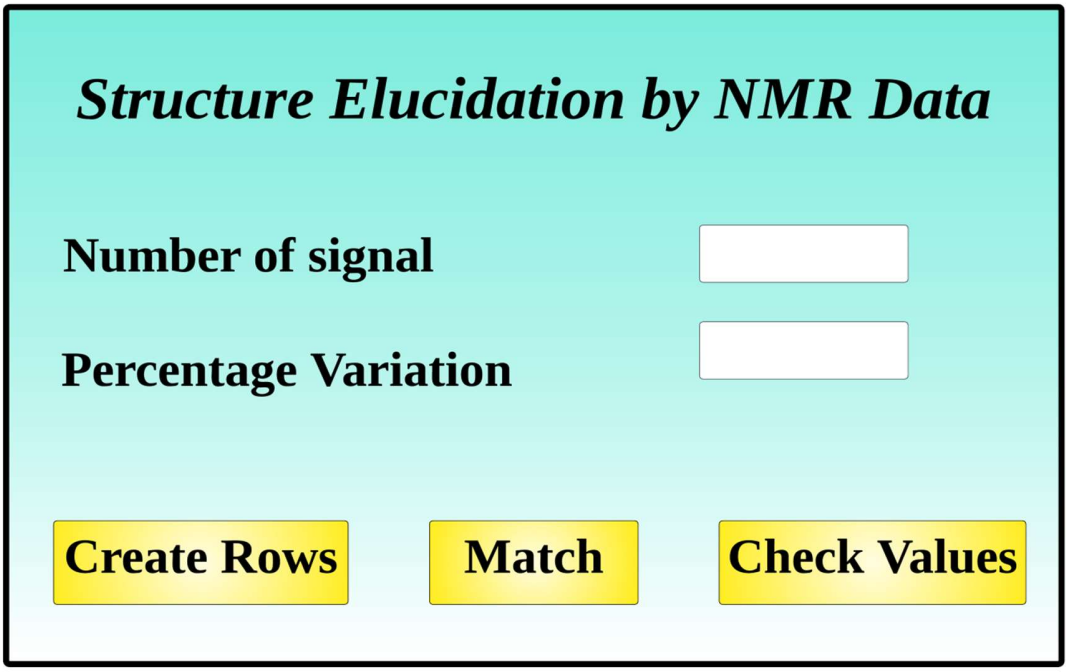


Figure 4. 1 (GUI (Graphical User Interface) where the user can enter the NMR data, In this GUI user will enter the Number of Signal and percentage variations, then click on given buttons i.e. 'Create rows', 'Match', 'Check Values' for desired purpose)

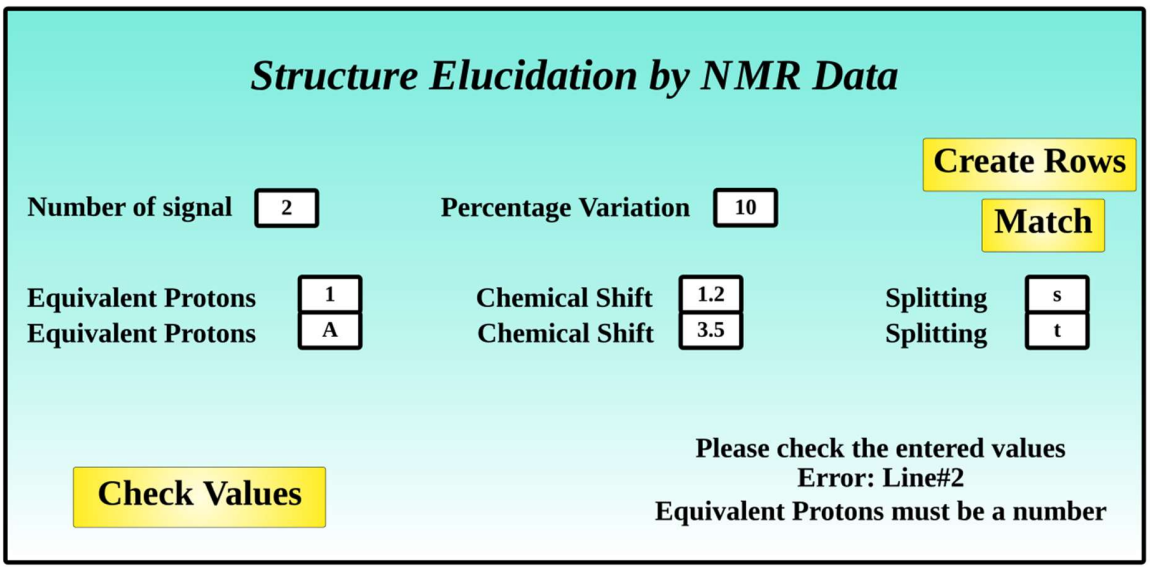


Figure 4. 2 (An Error message is shown when an invalid value is entered in the entry box of 'Equivalent Protons'. A similar message will be shown if 'Chemical Shift' be entered other than float and 'Splitting' be entered other than 's, d, t, q, or m))

4.4 Linking Excel Sheet, External Website, and Filtering & Sorting Data:

The Excel sheet was imported into the program and a for loop was created. This facilitated the output in SMILES language on entering the desired information as per the columns of the Excel sheet (Number of Signals, Number of Equivalent protons, Chemical Shift, and splitting). The sorted data from the program was facilitated to automatically enter as an input to a URL http://hulab.rxnfinder.org/smi2img/{final_output}. The final output from the URL was received output in real-time to give a structure of the molecule for the data entered.

CHAPTER 5

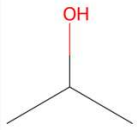
RESULTS AND DISCUSSIONS

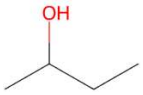
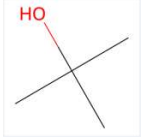
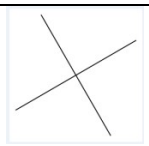
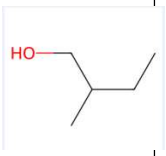
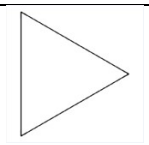
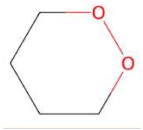
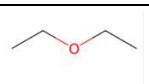
As discussed in Chapter-3 it was envisioned to create a program efficiently processing NMR data to provide output in the form of SMILES Code, molecule names, and molecular structure. Therefore an algorithm was developed using Python as a computer language.

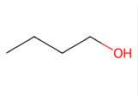
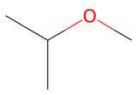
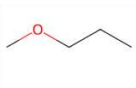
5.1 Creating a Library

A set of 14 molecules were listed to include diverse structures like primary, secondary, tertiary, branched, cyclic, and isomeric hydrocarbons and alcohols. The published data of all such molecules including the number of signals, number of equivalent protons, chemical shift, and splitting was compiled as shown in **Table 4.1**. It can be seen that data was segregated as input by the user and output by the proposed program(**Table 5.1**).

Table 5.1 (Input by the user includes the number of signals, number of equivalent protons, chemical shift, and splitting and output by the proposed program includes the image of molecular structure and SMILES code)

Sr. No.	Molecules	Input				Output	
		No. of Signal	No. of Equivalent Protons	Chemical Shift	Splitting	Structure	SMILES
1.	Methane	1	4	0.9	s		C
2.	Methanol	2	3 1	3.43 3.66	s s		CO
3.	Propan-2-ol	3	6 1 1	1.2 4.01 2.16	d m s		CC(C)O

4.	Butan-2-ol	5	3 1 1 2 3	1.17 3.71 2.37 1.46 0.93	d m s m t		CC(O)CC
5.	2-methylpropan-2-ol	3	9 1	1.26 2	s s		CC(C)(C)O
6.	2,2-dimethylpropane	1	12	0.9	s		CC(C)(C)C
7.	2-methylpropane	2	1 9	2 0.8	m d		C(C)(C)(C)
8.	2-methylbutan-1-ol	5	3 2 1 3 1	0.93 1.46 3.71 1.17 2.37	t m m d s		CCC(C)O
9.	Cyclopropane	1	6	0.2	s		C1CC1
10.	1,2-dioxane	2	4 4	3.55 1.45	t t		O1CCCCO1
11.	Ethoxyethane	2	6 4	1.21 3.47	t q		C(C)OCC

12.	Butan-1-ol	5	3 2 2 2 1	0.94 1.39 1.53 3.63 2.24	t m m t s		CCCCO
13.	2-methoxypropane	3	6 1 3	1.2 3.7 3.2	d m s		CC(C)OC
14.	1-methoxypropane	4	3 2 2 3	3.337 3.336 1.59 0.93	s t m t		COCCC

5.2 Writing Code for the development of program

The program was developed using the Python programming language. Graphical User Interface (GUI) was developed using Python tkinter library to facilitate users to enter and process the data. The process of Dynamic entry boxes is implemented to sequence the data. In order to avoid any errors, it was ensured that clear instructions were provided by introducing text in the GUI to guide the users for data input. The variability observed in NMR peaks due to different instruments, solvents, and synthetic methods, the program was algorithmed to accomodate ± 10% difference in the chemical shift values. To verify the correctness of the input data a for loop code was introduced to display error messages. The message suggested modifying the input if its incorrect or reviewing the input if it is correct.

5.3 Linking a program with an Excel sheet

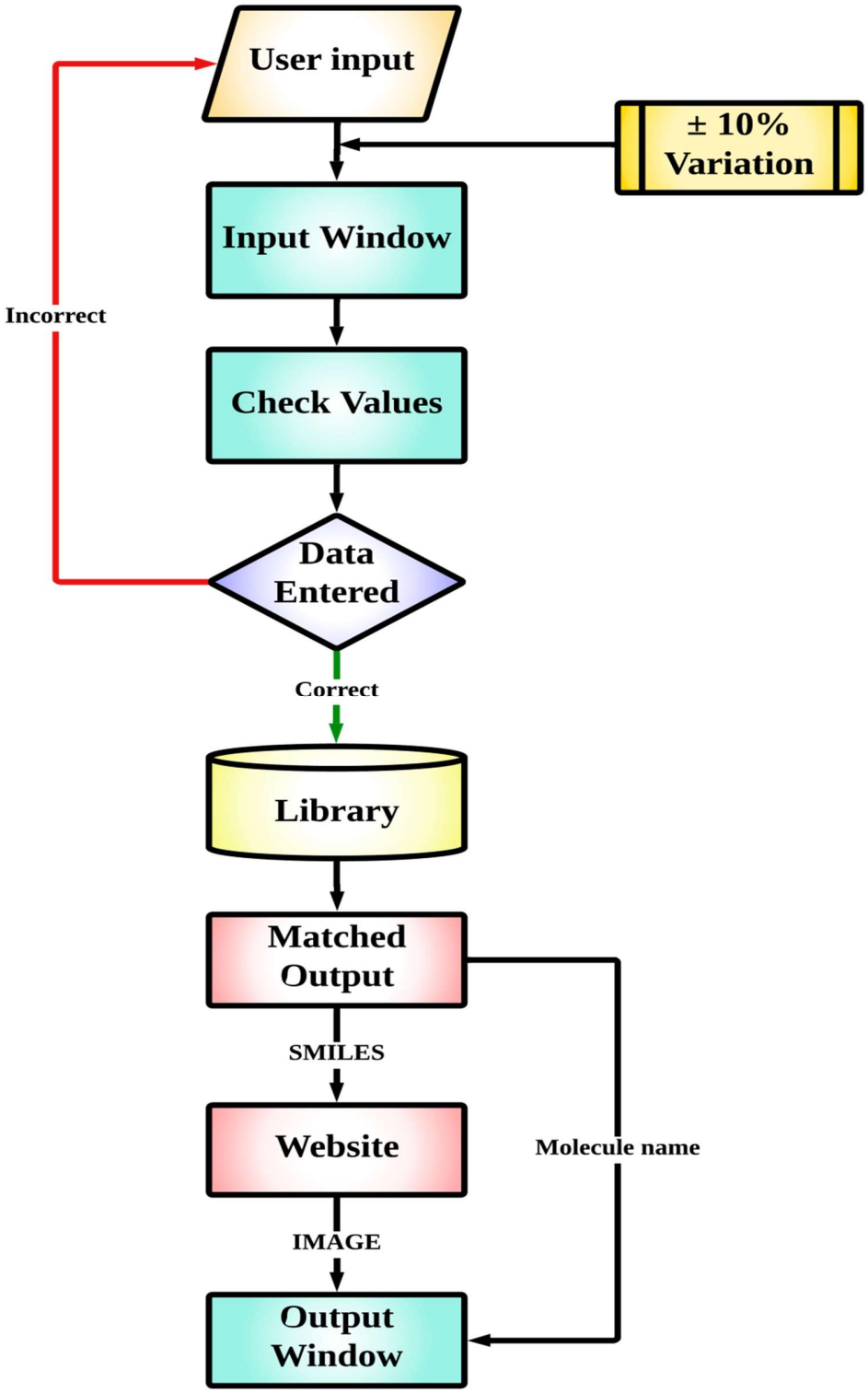
Using the Pandas library available in Python, the program was integrated with the Excel sheet. The Excel sheet served as the program's core database. The algorithms to get input data, filter using a

database, and finally, display the output were introduced in the program. Thus, upon entering data in the input window, the program was algorithmed to retrieve the user-provided information and proceeded to search and match the relevant data within the Excel sheet. To facilitate this process, the program was empowered to employ various filtering and sorting techniques to ensure accurate and efficient retrieval of the desired output at the press of the 'Match' tab. Thus, the program offered a streamlined and efficient solution for processing and presenting the desired results like the name of the molecule, SMILES code, and constituents of the molecule. This however displayed only the SMILES code, the name of the molecule, and the atoms present. The desired chemical structure, as anticipated in the objectives could not be realized. Therefore it was envisioned to link the program with an open license *platform* that converts SMILES code into the corresponding molecular structure.

5.4 Linking a program with a website

To facilitate this integration, the desired Python libraries were imported and utilized to facilitate real-time retrieval of the chemical structures. The web-browser library enabled the program to interact with the website. The PIL (Python Imaging Library) imported the retrieved molecular structure images within the program. Additionally, the requests library used a smooth connection with the website to enable data exchange.

Finally, the web integration functionality enabled the web browser, PIL, and requests libraries to obtain real-time molecular structure images based on SMILES input. Scheme-1 shows the summary of inputs and process to display the final output window



Scheme- 1 Represents the whole process of the user inputting data into the input window and allows +- 10% variation, then checking the validity of data, if the data is valid then proceed to Library if not valid then move to user input, after matching from the library, matched output in form of SMILES code is sent to a website to display molecule's image , name, and SMILES code on the display window.

5.5 Challenges

The project facilitated the prediction of the molecular structure of organic compounds with ^{1}H NMR. However, at various stages, several challenges were encountered. Some of the challenges faced and their resolution are worth mentioning and have been described in the following lines.

(1) Accommodating Variability of Input Data

The real data has variable **δ (**Chemical shift)values depending on the solvent, instrument, or the conditions used for the NMR analysis. For example, **Figure- 5.1** shows variable values (δ 5.0 to 1.0 ppm) of OH for different concentrations of Ethanol in CCl4.

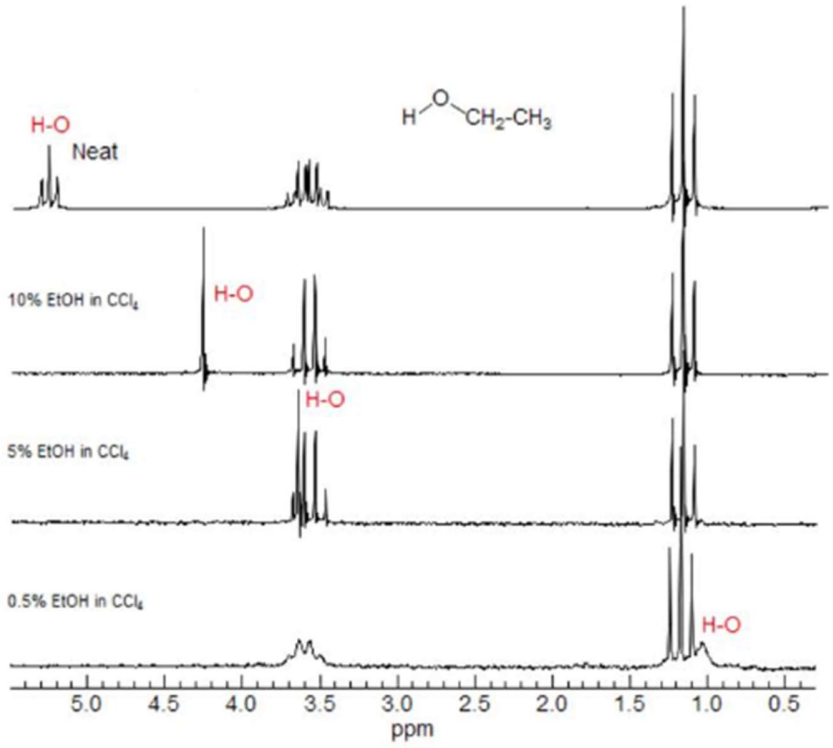


Figure - 5.1 (60 MHz spectra shows variable values (δ 5.0 to 1.0 ppm) of OH for different concentrations of Ethanol in CCl4.)

The challenge to get precise structure from the library, even if there is a large variation in data was resolved by introducing the variability option in the program. This gave freedom to users to accommodate the variability in the input data from the standard value. It is worth mentioning that due to large deviation in the acidic group protons (-OH, -NH_2,-COOH) δ value variability ranged between 10% to 25% while other protons gave precise results within 10% variability.

(2) Sorting of NMR data:

The input ^{1}H NMR data is generally written either as increasing or decreasing δ values. The program was initially designed to take input in a particular sequence so ass to give output as per desired structure. The wrong order of the user gave a wrong or redundant output. Therefore a sorting functionality was introduced into the Excel sheet to give the same correct output irrespective of the order of the input values. **Figure 5.2** illustrated the above-mentioned point.

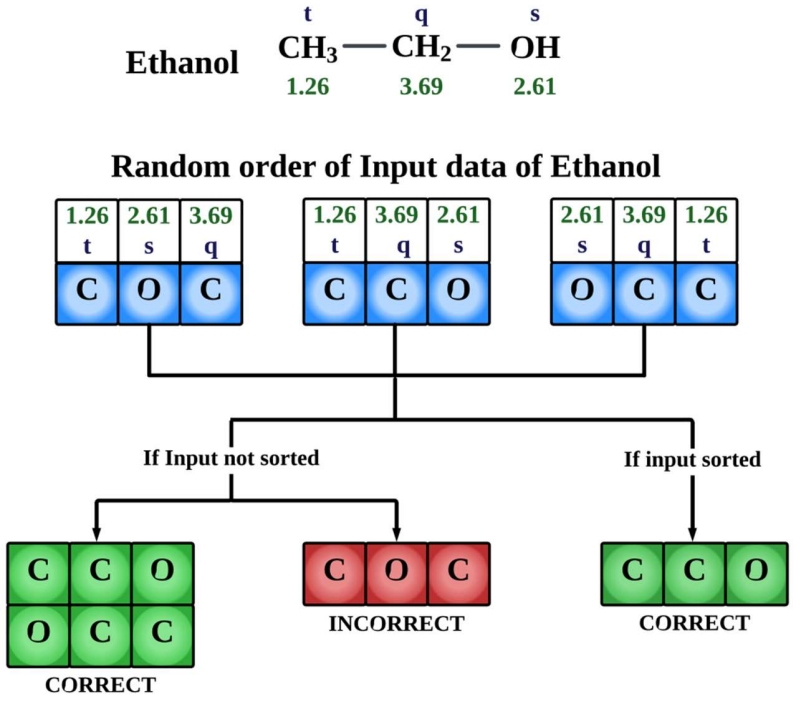


Figure- 5.2 (Ethanol is indicated with δ values and a splitting pattern. then the user can enter ^{1}H NMR data in any random order if the input is not sorted, the chance of getting correct and incorrect output is 50% but if the input is sorted, then the chance of getting correct output is 100%)

(3) Real-Time Image import from Website:

To overcome the challenge of drawing the large and complex structure in excel sheet, it was envisaged to give output of entered ^{1}H NMR data as SMILES code. A number of open license websites can then be used to generate the final structure from SMILES code. This, however, made structure determination using the above application a two-step process, Integrating the open license website into the program window gave an opportunity to complete the structure determination process as a single step. Therefore using a web browser, PIL, and request library was used to integrate the functionality into the program.

CHAPTER 6

CONCLUSION AND FUTURE DISCUSSION

6.1 Conclusion

In conclusion, we have developed a Python based program from scratch that enables users to predict the structures by submitting the ^{1}H NMR data including the number of signals, equivalent protons, chemical shifts, and splitting. It matches, sorts and gives output in the form of SMILES language using a library compiled in excel sheet. The highlight of the program is single window display of precise molecule structure and giving same output despite the sequence of data entered, Although the library has only 14 molecules but same program can be extended to 100 to 500 molecules depending upon the speed of the processor and coding used. The present version only incorporates ^{1}H NMR. However, there are several requirements that need to be fulfilled, which can be addressed in future versions of the program shown in **Figure-6.1**. For instance, additional components such as coupling constants, other types of NMR such as 2D, HMBC, HSQC, etc., and the ability to work with a greater number of solvents should be included.

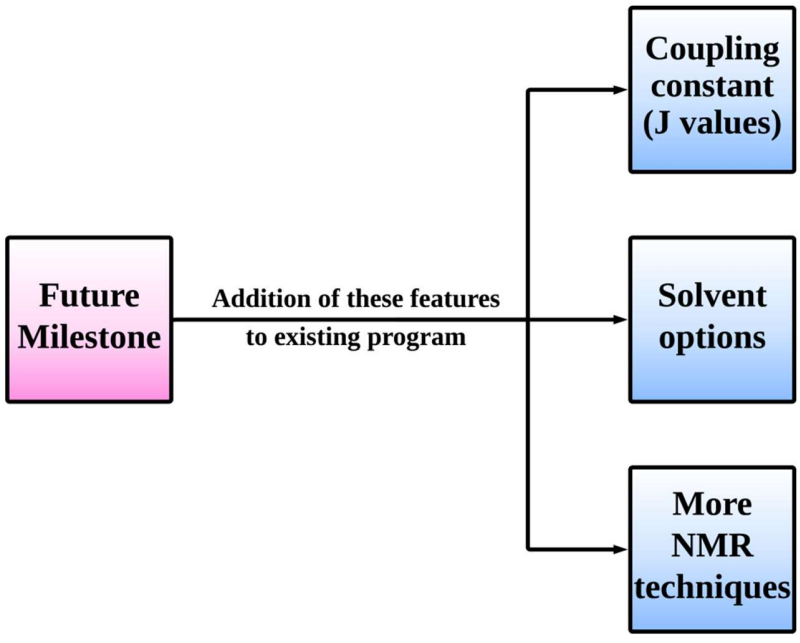


Figure- 6.1 (The program developed is the first version designed to elucidate the structure of molecules. To accurately elucidate the structure of larger molecules, the system must take into account coupling values(J), solvent option and more types of NMR techniques such as 2D NMR, C13, and others.)

REFERENCES

(1) Heitler, W.; London, F. WechselwirkungNeutralerAtome Und HomöopolareBindungNachDerQuantenmechanik. *Zeitschrift für Physik volume***1927**, *44* (6-7), 455–472. https://doi.org/10.1007/bf01397394.

(2) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Physical Review***1964**, *136* (3B), B864–B871. https://doi.org/10.1103/physrev.136.b864.

(3) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review***1965**, *140* (4A), A1133–A1138. https://doi.org/10.1103/physrev.140.a1133.

(4) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Modeling***1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(5) Johnson, B. A.; Blevins, R. A. NMR View: A Computer Program for the Visualization and Analysis of NMR Data. *Journal of Biomolecular NMR***1994**, *4* (5), 603–614. https://doi.org/10.1007/bf00404272.

(6) Munk, M. E.; Christie, B. D. The Characterization of Structure by Computer. *Analytica Chimica Acta***1989**, *216*, 57–68. https://doi.org/10.1016/s0003-2670(00)82004-8.

(7) KimitoFunatsu; Nobuyoshi Miyabayashi; Sasaki, S. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.***1988**, *28* (1), 18–28. https://doi.org/10.1021/ci00057a003.

(8)Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Martirosian, E. R.; Molodtsov, S. G. Application of a New Expert System for the Structure Elucidation of Natural Products from Their 1D and 2D NMR Data. *Journal of Natural Products, 2002, Vol. 65, No. 5***2002**, *65* (5), 693–703.

(9)Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Molodtsov, S. G.; Martin, G. E.; Martirosian, E. R. Structure Elucidator: A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments. *Journal of Chemical Information and Computer Sciences***2004**, *44* (3), 771–792. https://doi.org/10.1021/ci0341060.

(10)Burns, D. C.; Mazzola, E. P.; Reynolds, W. F. The Role of Computer-Assisted Structure Elucidation (CASE) Programs in the Structure Elucidation of Complex Natural Products. *Natural Product Reports***2019**, *36* (6), 919–933. https://doi.org/10.1039/c9np00007k.

(11)Elyashberg, M. E.; Williams, A. J. *Computer–Based Structure Elucidation from Spectral Data*; Springer, 2015.

(12)Elyashberg, M. E.; Williams, A. J. *Magnetic Resonance in Chemistry*; 2018.

(13)Elyashberg, M. E.; Sriram Tyagarajan; Mandal, M.; Buevich, A. V. Enhancing Efficiency of Natural Product Structure Revision: Leveraging CASE and DFT over Total Synthesis. *Molecules* **2023**, *28* (9), 3796–3796. https://doi.org/10.3390/molecules28093796.

(14)Kessler, P.; Godejohann, M. Identification of Tentative Marker in Corvina and Primitivo Wines with CMC-Se. *Magnetic Resonance in Chemistry***2018**, *56* (6), 480–492.

(15)Troche-Pesqueira, E.; Clemens Anklin; Gil, R. R.; Navarro-Vázquez, A. Computer-Assisted 3D Structure Elucidation of Natural Products Using Residual Dipolar Couplings. *Angew. Chem. ,Int. Ed.,* **2017**, *56* (13), 3660–3664. https://doi.org/10.1002/anie.201612454.

(16)Nuzillard, J.-M.; Plainchont, B. Tutorial for the Structure Elucidation of Small Molecules by Means of the LSD Software. *Magnetic Resonance in Chemistry***2017**, *56* (6), 458–468. https://doi.org/10.1002/mrc.4612.

(17)Navarro-Vázquez, A.; Gil, R. R.; Blinov, K. A. Computer-Assisted 3D Structure Elucidation (CASE-3D) of Natural Products Combining Isotropic and Anisotropic NMR Parameters. *J. Nat. Prod.,***2018**, *81* (1), 203–210. https://doi.org/10.1021/acs.jnatprod.7b00926.

(18)Castro, S. J.; García, M. E.; Padrón, J. M.; Navarro-Vázquez, A.; Gil, R. R.; Nicotra, V. E. Phytochemical Study of *Senecio Volckmannii* Assisted by CASE-3D with Residual Dipolar Couplings and Isotropic $^{1}H/^{13}c$ NMR Chemical Shifts. *J. Nat. Prod.,***2018**, *81* (11), 2329–2337. https://doi.org/10.1021/acs.jnatprod.8b00162.

(19)Plainchont, B.; de Paulo Emerenciano, V.; Nuzillard, J.-M. Recent Advances in the Structure Elucidation of Small Organic Molecules by the LSD Software. *Magnetic Resonance in Chemistry***2013**, *51* (8), 447–453. https://doi.org/10.1002/mrc.3965.

(20)Bakiri, A.; Bertrand Plainchont; Vicente; Reynaud, R.; Hubert, J.; Renault, J.-H.; Jean-Marc Nuzillard. Computer-Aided Dereplication and Structure Elucidation of Natural Products at the University of Reims. *Mol. Inf.,***2017**, *36* (10), 1700027–1700027. https://doi.org/10.1002/minf.201700027.

(21)*CSEARCH-NMR-Server*. nmrpredict.orc.univie.ac.at. http://nmrpredict.orc.univie.ac.at (accessed 2023-07-13).

(22)Howarth, A.; Kristaps Ermanis; Goodman, J. M. DP4-AI Automated NMR Data Analysis: Straight from Spectrometer to Structure. *Chem. Sci.,***2020**, *11* (17), 4351–4359. https://doi.org/10.1039/d0sc00442a.

(23)Wenk, M.; Jean-Marc Nuzillard; Steinbeck, C. Sherlock—a Free and Open-Source System for the Computer-Assisted Structure Elucidation of Organic Compounds from NMR Data. *Molecules***2023**, *28* (3), 1448–1448. https://doi.org/10.3390/molecules28031448.

(24)*Structure Elucidation | Small Molecule Elucidation*. www.bruker.com. https://www.bruker.com/en/products-and-solutions/mr/nmr-software/cmc-se.html (accessed 2023-07-14).

(25) The ^{1}H NMR data was compiled from AIST (National Institute of Advanced Industrial Science and Technology (AIST)) and compiled into excel sheet.

PLAGIARISM REPORT

Thesis of Vanshika

ORIGINALITY REPORT

6%	2%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

Mikhail E. Elyashberg, Kirill A. Blinov, Antony J. Williams, Eduard R. Martirosian, Sergey G. Molodtsov. "Application of a New ExpertSystem for the Structure Elucidation of Natural Products from Their 1D and 2D NMR Data", Journal of Natural Products, 2002 Publication	2%
Bruce A. Johnson, Richard A. Blevins. "NMR View: A computer program for the visualization and analysis of NMR data",Journal oygf Biomolecular NMR, 1994Publication	1%
www.wattpad.com Internet Source	1%
Handbook of Chemoinformatics", Wiley, 2003 Publication	1%
www.ibb.cnr.it Internet Source	<1%
Mikhail E. Elyashberg, Kirill A. Blinov, Antony J. Williams, Sergey G. Molodtsov, Gary E. Martin, Eduard R. Martirosian. " A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments ", Journal of Chemical Information and Computer Sciences, 2004 Publication	<1%
pubmed.ncbi.nlm.nih.gov Internet Source	<1%
www.researchgate.net Internet Source	<1%

Gary E. Martin. "Identification of degradants of a complex alkaloid using NMR cryoprobe technology and ACD/structure elucidator", Journal of Heterocyclic Chemistry, 11/2002 Publication	<1%
hrcak.srce.hr Internet Source	<1%
Alexei V. Buevich, Mikhail E. Elyashberg. "Synergistic Combination of CASE Algorithms and DFT Chemical Shift Predictions: A Powerful Approach for Structure Elucidation, Verification, and Revision", Journal of Natural Products, 2016 Publication	<1%
Lecture Notes in Chemistry, 2015.Publication	<1%
Exclude quotes	ON
Exclude Bibliography	ON
Exclude matches	<7 Words