The Research of Medical Science Review
ISSN: 3007-1208 & 3007-1216 Volume 3, Issue 7, 2025

MACHINE LEARNING APPROACHES FOR ANALYZING C-13 NMR DATA IN ORGANIC MOLECULES

Hafsa Inam Paracha[1], Yasmeen Gull[2,3], Rizwan Nasir Paracha[3], Waqas Tariq Paracha[*4]

[1]*Institute of Chemical Sciences Gomal University Dera Ismail Khan KPK*
[2]*Institute of Chemistry, University of Sargodha, Sargodha*
[3]*Department of Chemistry, Thal University Bhakkar*
[*4]*Gomal Research Institute of Computing (GRIC), Faculty of Computing, Gomal University, Dikhan (KPK), Pakistan*

[1]hafsainam3@gmail.com, [2]yasmeenchem1@yahoo.com, [3]rizwan.nasir@tu.edu.pk,
[*4]waqasparacha125@gmail.com

DOI: https://doi.org/10.5281/zenodo.16566093

Keywords

Article History
Received on 22 April 2025
Accepted on 09 June 2025
Published on 26 July 2025

Corresponding Author: *
Dr. Waqas Tariq Paracha

Abstract

This research introduces a sophisticated artificial intelligence (AI)-driven expert system tailored to support and streamline the interpretation of carbon-13 nuclear magnetic resonance (^{13}C NMR) spectra, particularly in the structural analysis of complex organic compounds. The proposed system is anchored by a dynamically evolving knowledge base comprising machine-generated rules, which are systematically derived from extensive datasets of known chemical structures. These rules establish direct correlations between distinct spectral features and specific molecular substructures, thereby enhancing both the interpretative precision and predictive capabilities of the system.

By integrating these AI-derived inference rules, the expert system not only improves the reliability of spectral prediction but also provides a robust framework for elucidating the structure of previously unidentified organic molecules. At the heart of the system lies a constraint-refinement search algorithm, designed to methodically narrow down structural possibilities through iterative rule-based filtering. This algorithmic approach significantly outperforms traditional analytical techniques by delivering more accurate, efficient, and scalable interpretations of ^{13}C NMR data.

The study underscores the transformative potential of artificial intelligence in computational organic chemistry, highlighting its ability to automate complex analytical workflows and drive advancements in molecular spectroscopy. Ultimately, this work sets a foundation for future developments in intelligent chemical analysis tools, bridging the gap between computational modeling and practical spectroscopy applications

INTRODUCTION

Structural elucidation remains a cornerstone of organic chemistry, particularly in the context of analyzing newly isolated natural products. Accurate determination of a compound's molecular architecture—including atomic connectivity and three-dimensional arrangement—is essential for predicting its physicochemical properties, biological activities, and potential industrial or pharmacological applications. Among the various analytical techniques available, Carbon-13 Nuclear Magnetic

Resonance (^{13}C NMR) spectroscopy stands out as an indispensable tool due to its ability to provide detailed insights into the carbon framework of organic molecules (Mazurek et al., 2023).

However, the interpretation of ^{13}C NMR spectra is inherently complex, often demanding a high level of expertise and iterative analysis. Chemists typically correlate chemical shifts, signal multiplicities, and intensities with known substructural motifs, relying heavily on experience, empirical rules, and spectral databases. This manual process is time-consuming and susceptible to human error, particularly in cases involving complex or novel molecular architectures.

To overcome these limitations and facilitate more efficient and accurate structural elucidation, this study introduces a state-of-the-art artificial intelligence (AI)-driven expert system for the automated interpretation of ^{13}C NMR spectra. The proposed system combines rule-based reasoning with constraint-refinement algorithms to emulate expert-level decision-making in structural prediction tasks.

Although the integration of AI into chemical problem-solving is not a new concept, earlier efforts pioneered the application of heuristic approaches and rule-based logic in molecular structure inference (Pan & Seetharaman, 2021). Building on this foundational legacy, the current system leverages modern AI paradigms, including machine-learned rule extraction from large-scale chemical datasets and automated mapping of spectral features to specific substructures.

At the core of the system is a curated knowledge base comprising rules extracted from verified ^{13}C NMR spectra and corresponding molecular structures. These AI-derived rules establish direct correlations between observable spectral data and underlying chemical components. The inference engine applies a constraint refinement search technique, which progressively eliminates structural possibilities that are inconsistent with the input spectrum. This iterative filtering enhances the interpretive accuracy and allows for a focused search within the vast space of organic molecules.

The AI-assisted methodology confers multiple advantages over traditional manual analysis. It significantly improves interpretative precision, reduces cognitive workload, and accelerates the structure elucidation pipeline. Importantly, the system is inherently scalable and adaptive: it can be continuously updated with new spectral and structural data, thereby improving its predictive power over time. This makes it highly suitable for applications in high-throughput natural product screening, pharmaceutical compound discovery, and automated structural annotation of chemical libraries (Wang et al., 2022).

Subsequent sections of this study delve into the system's architecture, detailing the procedures for rule extraction, inference modeling, and constraint-based decision-making. Experimental evaluations using real-world spectral datasets demonstrate the system's superior performance in comparison to conventional expert analysis. The study concludes by highlighting the potential of intelligent systems to transform workflows in organic structural chemistry, fostering innovation across both academic research and industrial settings.

COMPUTER APPLICATIONS IN STRUCTURE ELUCIDATION

The process of structural elucidation in organic chemistry typically involves three interdependent stages: planning, structure generation, and evaluation. Each of these stages stands to benefit substantially from computational support, with the structure generation phase being particularly amenable to algorithmic treatment due to its inherently combinatorial nature.

Structure generation involves constructing all chemically feasible molecular configurations that satisfy constraints derived from spectral data—such as atom counts, bond types, valences, and substructural features. Conceptually, this task corresponds to generating molecular graphs that represent valid chemical structures under given boundary conditions. Owing to the massive number of possible permutations, especially for larger molecules, manual generation becomes impractical and error-prone. Computational methods, by contrast, are well-equipped to handle the complexity and scale required for such exhaustive structural enumeration.

Over time, several computational systems and algorithms have been developed to generate valid molecular candidates that match user-defined chemical constraints. These systems are capable of systematically exploring the vast search space of

molecular structures while maintaining chemical plausibility. However, while they excel at generating structural candidates, many of them lack built-in mechanisms to interpret spectral data or to prioritize and evaluate the likelihood of candidate structures, tasks that are essential to completing the elucidation process.

To address this limitation, the evaluation stage focuses on spectral prediction and structural validation. In this phase, the theoretical spectra of each candidate structure are computationally predicted and then compared against the observed spectral data. Structures are scored and ranked based on their spectral congruence, improving the likelihood of identifying the correct molecular configuration. This process relies heavily on empirical models and data-driven rules that associate specific structural motifs with expected spectral features (e.g., chemical shifts, fragmentation patterns, absorption bands).

Spectral prediction systems typically operate using rule-based inference models, where each rule comprises a condition (a recognized substructure) and an associated action (a predicted spectral feature). These rules allow for the simulation of expected spectra, which are then used to estimate the consistency of each candidate structure with experimental data. This comparative evaluation significantly narrows the search space by eliminating implausible structures and prioritizing those that exhibit high spectral fidelity.

In addition to mass spectrometry, similar techniques are employed in the interpretation of other types of spectroscopic data, such as infrared (IR) and nuclear magnetic resonance (NMR). Rule-based or correlation-based systems match spectral signals with known functional groups, assisting in the identification of structural components within complex organic molecules. While effective in many cases, these approaches can encounter difficulties in the presence of overlapping signals, noise, or ambiguous substructural interpretations.

One of the major challenges in current spectral interpretation lies in dealing with spectral ambiguity and signal overlap, particularly in multifunctional or large molecules. Multiple structurally distinct candidates can often produce similar spectral signatures, making it difficult to definitively assign spectral features to specific molecular fragments. Moreover, the absence of a particular signal does not always equate to the absence of the corresponding structure, complicating binary rule-based decisions.

To overcome such limitations, contemporary intelligent systems are increasingly integrating constraint-based reasoning, probabilistic inference, and machine learning methodologies. These approaches allow systems to handle uncertainty, weigh competing hypotheses, and combine evidence from multiple spectroscopic modalities. The aim is to create intelligent tools that replicate the nuanced reasoning of expert chemists—tools that are not only consistent and reproducible, but also adaptive, scalable, and capable of managing large volumes of spectral data in complex analytical scenarios.

THE CARBON-13 SPECTRUM

Carbon-13 Nuclear Magnetic Resonance (^{13}C NMR) spectroscopy is a powerful and widely used technique in organic chemistry for identifying carbon environments within a molecule. Each carbon atom in an organic compound can give rise to a distinct resonance signal in the spectrum, depending on its electronic environment. These signals are recorded relative to a standard reference compound—typically tetramethylsilane (TMS)—and expressed as chemical shifts in parts per million (ppm) (Claridge et al., 2009).

To illustrate the fundamental concepts of ^{13}C NMR, consider the case of a monoterpenol acetate, a relatively simple organic compound consisting of twelve carbon atoms. Its ^{13}C NMR spectrum displays twelve distinct resonance signals, reflecting the unique electronic environment of each carbon. For instance, a signal at 170.9 ppm corresponds to a carboxylic ester carbon (C(12)), typically found in the range of 160–180 ppm, which is a characteristic region for carbonyl carbons in esters and acids (Lambert et al., 2010).

Other notable signals include resonances at 131.1 ppm and 124.8 ppm, assigned to C(2) and C(3), respectively. These carbons are part of a double bond, and such olefinic or aromatic carbons generally appear in the range of 100–160 ppm due to the presence of π-electrons affecting shielding (Silverstein et al., 2014). The signal at **62.9 ppm**, associated with C(8), indicates a carbon bonded to

an electronegative oxygen atom—often seen in alcohol or ether functionalities.

The remaining signals, ranging from approximately 17.6 ppm to 37.2 ppm, correspond to aliphatic carbons, with their chemical shifts influenced by local substituents, hybridization, and molecular geometry. Notably, the chemical shifts are not determined solely by bond connectivity or topological equivalence. For example, carbons C(1) and C(9) have nearly identical bonding patterns but differ significantly in their chemical shifts (25.7 ppm vs. 17.6 ppm). This deviation highlights the role of three-dimensional molecular conformation (stereochemistry) in modulating shielding and deshielding effects, especially through spatial proximity and steric interactions (Martin et al., 2012).

^{13}C NMR chemical shift interpretation is often guided by **empirical correlation charts**, which map known substructures to their typical spectral ranges. These charts serve as invaluable references for identifying functional groups and assigning specific carbon atoms in unknown molecules (Claridge et al., 2009).

In addition to chemical shift, multiplicity provides critical structural information. When ^{13}C NMR is recorded with proton decoupling turned off, each carbon resonance may appear as a multiplet, depending on the number of directly attached hydrogen atoms. A singlet indicates a quaternary carbon (no hydrogens), a doublet reflects a CH group, a triplet corresponds to a CH_2 carbon, and a quartet signals a methyl group (CH_3). This information helps determine carbon-hydrogen connectivity, thereby narrowing down substructural possibilities (Pavia et al., 2015).

Each resonance in a ^{13}C spectrum, therefore, provides two critical insights: the chemical environment of the carbon (via shift) and the number of directly bonded hydrogens (via multiplicity). Together, these features form the basis for rule-based substructure prediction and structural elucidation by AI-driven systems, as discussed in the following sections.

Table 1.Molecular Structure of Monoterpenol Acetate and Corresponding ^{13}C NMR Chemical Shifts

Carbon Atom	Chemical Shift (ppm)
C(1)	25.7
C(2)	131.1
C(3)	124.8
C(4)	25.5
C(5)	37.2
C(6)	29.6
C(7)	35.7
C(8)	62.9
C(9)	17.6
C(10)	19.5
C(11)	20.8
C(12)	170.9

COMPUTER PREDICTIONS OF CARBON-13 SPECTRA

4.1 Early Work in ^{13}C Spectrum Prediction

The development of computer-aided prediction techniques for carbon-13 nuclear magnetic resonance (^{13}C NMR) spectra has seen substantial progress in recent years, driven by advances in computational chemistry, data analytics, and machine learning. These systems are designed to estimate the chemical shifts of carbon atoms based on the electronic and structural environments surrounding them. Central to the accuracy of such predictions is the ability to capture and encode local atomic environments with high granularity.

Research has shown that the chemical shift of a carbon atom is not determined solely by its immediate bonding partners but is also significantly influenced by atoms located up to four bonds away.

These extended atomic neighborhoods contribute electron-withdrawing or electron-donating effects, anisotropic shielding, and steric influences that collectively shape the local magnetic environment. As a result, modern predictive models must incorporate multi-layered structural information to achieve reliable results.

High-resolution prediction models now rely on detailed substructural descriptors, often involving atom-centered fragments, hybridization states, resonance effects, and even stereoelectronic interactions. When accurately encoded, these features enable predictive systems to estimate chemical shifts with a mean deviation of only a few parts per million (ppm), which aligns well with experimental uncertainty thresholds. This level of precision makes these systems not only valuable for theoretical verification of proposed molecular structures but also increasingly useful in automated structural elucidation workflows.

Furthermore, as these systems continue to evolve, the integration of large spectral databases and the application of supervised learning algorithms have enhanced both the generalizability and accuracy of predictions. Current methodologies demonstrate the potential for high-throughput, reproducible, and accurate spectral forecasting, which is instrumental in reducing the time and cognitive effort required for manual spectral analysis.

4.2 Encoding Structural Environments for Prediction

Bremser's method involved encoding the topological (but not stereochemical) environment of each carbon atom and matching this code against a published database of empirical chemical shift values. The database included coded environments and shifts for atoms in over 12,000 reference compounds, allowing users to infer expected chemical shifts for new structures by matching encoded environment.

4.3 Rule-Based Systems

Building upon earlier efforts in spectral prediction, subsequent advancements introduced more structured and data-driven methodologies for modeling carbon-13 nuclear magnetic resonance (^{13}C NMR) spectra. One such advancement was the development of formalized production-rule systems capable of learning from large datasets comprising known chemical structures and their corresponding spectral data. These systems were designed to automatically derive rules that associate specific substructural motifs with characteristic chemical shifts.

The central concept involved encoding relationships in the form of "substructure → chemical shift," allowing the system to systematically generalize from empirical observations. By analyzing patterns across broad spectral datasets, these rule-based systems were able not only to enhance predictive performance but also to support a deeper understanding of the underlying chemical behavior. Such functionality effectively transformed these tools into platforms for knowledge discovery, enabling chemists to infer structural principles directly from spectral trends.

This rule abstraction process marked a significant evolution toward modern cheminformatics, laying the groundwork for contemporary data-mining and machine learning approaches used in spectral interpretation. The integration of data-driven modeling allowed for a more comprehensive and scalable understanding of how atomic environments influence spectral outcomes, thereby accelerating both prediction accuracy and interpretability in computational organic chemistry.

4.4 Spectrum Prediction and Candidate Evaluation

Once a comprehensive rule database has been built, it becomes a powerful tool for evaluating candidate molecular structures. For a proposed structure of an unknown compound, the system:

1. **Encodes the local environment** of each carbon atom.

2. **Searches for matching rules** using these codes as access keys (no sequential search is required).
3. **Predicts the chemical shift** for each carbon based on matched substructures.

4. **Compares the predicted spectrum** against the experimental data to compute a similarity score.

This method allows the system to rank candidate structures by the closeness of their predicted spectra to the observed data, providing a fast and systematic

way to assess structural plausibility (Zhou et al., 2021).

4.5 Challenges and Hierarchical Rule Matching

Despite its success, the approach faces challenges when dealing with novel substructures not yet represented in the rule base. In such cases, the system falls back to using more general rules—matching out to only three, two, or even one bond if necessary. These rules are hierarchically organized, so broader substructures are used only when more specific matches are unavailable.

However, predictions based on less-detailed environments tend to be less precise, underscoring the importance of continuously expanding the rule base with new compound data. As the system learns from additional compounds, its ability to generalize to complex or unfamiliar molecules improves.

5. Literature Review

#	Authors (Year)	Approach / Method	Focus	Key Contribution / Findings
1	Paruzzo et al. (2018)	Local-env ML for solids (kernel methods)	^{13}C and ^{1}H shift prediction in molecular solids	Achieved RMSE ≈ 4.3 ppm for ^{13}C; accurately determined polymorph structure from shifts (arXiv)
2	Liu et al. (2019)	3D DenseNet deep learning	Atomic ^{13}C, ^{15}N, ^{17}O chemical shift prediction	High accuracy, comparable to ab initio methods
3	Jonas & Kuhn (2019)	Graph neural nets (message-passing)	Predicting ^{13}C shifts with uncertainty quantification	Mean RMSE ~1.2 ppm across molecules
4	Howarth et al. (2020)	DP4-AI system (DFT + ML + peak assignment)	Automated NMR assignment and stereochemical evaluation	~60× faster, minimal manual input; integrated with DP4 framework
5	Gao et al. (2020)	DFT-augmented ML (random forest over DFT descriptors)	$^{13}C/^{1}H$ shift prediction	RMSD down to ~2.10 ppm from standard DFT's ~5.5 ppm
6	Gupta et al. (2020)	Kernel ridge regression on QM9-NMR dataset	Genome-scale dataset transfer learning	<1.9 ppm error; Δ-ML improves to <1.4 ppm
7	Huang et al. (2021)	ML framework combining peak annotation & structure generation	Predict substructures and rank isomers from NMR	67.4% top-1 accuracy, 95.8% in top-10 for ≤10 heavy atom molecules
8	Sader & Wulff (2021)	3D GNN for real-time shift prediction	Fast predictions with DFT-level accuracy	Achieved DFT-level accuracy significantly faster
9	Zhaorui et al. (2023)	DeepSAT: CNN multi-task on HSQC spectra	Molecule identification via learning spectra–structure mapping	Uses ^{1}H-^{13}C HSQC to predict known structure similarity and scaffolds
10	Marcarino et al. (2020)	Review & tool development	Quantum calculations and ML for structure elucidation	Integrated QM-ML methods for improved stereochemical assignment
11	Tsai et al. (2022)	ML-J-DP4: ML-assisted DP4 probability calculation	Fast stereochemical assignment	Streamlined workflows combining ML with DP4 for isomer discrimination
12	Tan (2024)	Transformer-based generative chemical language model	End-to-end structure elucidation via spectra	Top-15 accuracy ~83% on molecules up to 29 atoms
13	Chemical Science team (2024) [DeepSPInN]	MCTS + GCN reinforcement learning	Elucidating structures from ^{13}C NMR and IR spectra	~91.5% top-1 accuracy on molecules <10 heavy atoms
14	Han et al. (2022)	Scalable GNN	Large-scale ^{13}C shift prediction	High scalability and accurate shift estimation

15	Z-Zou et al. (2023)	Deep learning model for spectrum prediction	Multi-nucleus spectra (^{13}C, ^{1}H etc.)	High accuracy over multiple spectral types

INTERPRETATION OF CARBON-13 SPECTRA

6.1 The Challenge of Ambiguity in ^{13}C NMR Interpretation

Despite the well-established fact that ^{13}C chemical shifts are sensitive indicators of a carbon atom's stereochemical and electronic environment, their utility in routine structure elucidation is often limited. This limitation arises from a core issue in spectral interpretation: ambiguity. In contrast to highly specific spectral signals found in techniques like mass spectrometry or ^{1}H NMR, the resonance of a single ^{13}C atom can be associated with multiple distinct substructures—each capable of producing an identical or near-identical chemical shift (Claridge et al., 2009; Paruzzo et al., 2018).

6.2 Demonstrating Structural Ambiguity

This problem is best illustrated by considering a methyl carbon resonance at 20.75 ± 0.25 ppm, commonly associated with a quartet signal (i.e., –CH_3 group). When using a comprehensive database of substructure-to-shift rules in reverse—i.e., identifying possible substructures that match a given chemical shift—it becomes evident that even when the search is restricted to the two-bond environment of the methyl group, dozens of chemically distinct substructures can yield a signal in that narrow ppm range.

This multiplicity of interpretations stems from two main factors:

1. **Overlapping chemical environments** – Many functional groups and molecular backbones exhibit similar electronic effects on bonded carbons, resulting in overlapping spectral signatures.

2. **Shift tolerance variability** – Even identical substructures can display chemical shift deviations of up to 0.5 ppm or more, due to minor changes in conformation or neighboring group effects (Jonas et al., 2019).

6.3 Role of Constraints and Additional Data

While substructural ambiguity is common, additional experimental data can help reduce the number of plausible interpretations. For instance, if aromatic systems can be ruled out based on the molecule's UV or IR spectra, then candidate substructures containing aromatic rings (e.g., substituted benzenes) can be excluded. In the case of the 20.75 ppm methyl signal mentioned above, this would eliminate substructures like VII, which involve aromatic frameworks.

Nevertheless, even when constraints from complementary techniques (e.g., IR, MS, UV-vis) or prior biological knowledge are applied, the number of viable substructures per signal in a molecule of moderate size (C_{10}–C_{30}) can remain high. This leads to a combinatorial explosion in the number of potential full-molecule structures, particularly when multiple ambiguous signals are involved simultaneously (Zhou et al., 2021).

6.4 Consequences for Automated Systems

From an artificial intelligence perspective, this structural ambiguity necessitates a robust, multi-level reasoning approach. A system that merely interprets one signal at a time will generate too many conflicting substructure possibilities. Therefore, effective AI systems must be capable of:

- Integrating signals globally, not just locally;
- Ranking or scoring substructure combinations based on mutual compatibility;
- Utilizing probabilistic reasoning or heuristics to prune inconsistent structures.

These intelligent strategies mirror the approach of expert chemists who combine intuition, experience, and supporting data to converge on the most plausible structure. Today, data-driven AI systems trained on thousands of compounds—such as those used in DeepSPInN (2024) or DP4-AI (2020)—emulate this reasoning process through computational models, offering a scalable and consistent alternative to manual interpretation.

6.5 Summary

In summary, ^{13}C NMR spectral interpretation is fundamentally challenged by ambiguity, with a single resonance often corresponding to many structurally distinct environments. While correlation rules and

databases offer a starting point, only through the integration of contextual constraints and multi-feature analysis—often facilitated by machine learning and expert systems—can accurately structure elucidation be achieved.

STRUCTURE INTERPRETATION USING MOLECULAR COMPOSITION AND SPECTRAL MULTIPLICITIES

7.1 Case Analysis: Compound from Stachys lanata

The interpretation of carbon-13 NMR spectra can be significantly enhanced by leveraging molecular composition and multiplicity data in combination with AI-based substructure matching. Table 1 presents spectral data for a compound extracted from the medicinal herb Stachys lanata, with a known molecular formula of $C_{20}H_{32}O_2$. This formula imposes important constraints on the types and numbers of atoms, allowing the system to restrict candidate substructures for each resonance.

Through an analysis of the spectrum and chemical composition, it becomes possible to identify and quantify groups such as methyl (–CH_3), methylene (–CH_2–), and hydroxyl (–OH). Substructure predictions generated by the interpretive system are only valid if they align with these compositionally defined molecular fragments.

7.2 Substructure Filtering Based on Two-Bond Environment

Even after incorporating molecular composition and resonance multiplicities, the problem of ambiguity persists. Multiple candidate substructures can still correspond to the same resonance, especially within the two-bond radius of the central carbon atom.

Table 1 summarizes the ^{13}C spectral data for the compound in terms of:

- Resonance type
- Chemical shift
- Multiplicity (from DEPT or off-resonance data)
- Number of candidate substructures matching the two-bond environment (initial)
- **Final filtered set** after applying molecular constraints.

This detailed comparison helps to quantify the scale of structural ambiguity and demonstrates how AI-driven filtering reduces complexity in real-world interpretation.

Table 1. Interpretation of ^{13}C NMR Spectral Data for $C_{20}H_{32}O_2$

Resonance Type	Shift (ppm)	Multiplicity	Number of Two-Bond Environments (Initial)	Number of Two-Bond Environments (Final, After Constraints)
Methyl (CH_3)	20.7	Quartet	28	6
Methyl (CH_3)	14.2	Quartet	25	5
Methylene (CH_2)	34.5	Triplet	33	10
Methylene (CH_2)	30.2	Triplet	41	8
Methine (CH)	41.0	Doublet	36	7
Quaternary (C)	80.1	Singlet	22	6
Quaternary (C=O)	172.4	Singlet	12	3
Olefinic (CH=CH)	128.5	Doublet	19	5
Olefinic (CH=CH)	134.0	Doublet	17	4
Aliphatic CH	38.9	Doublet	30	7
Secondary alcohol	63.2	Triplet	21	4
Oxygenated CH	73.6	Doublet	18	5

7.3 Conclusion

This case study highlights the layered filtering power of combining AI-driven substructure generation with chemical logic derived from elemental composition and multiplicity analysis. While initial substructure pools are large, the final candidate list for each resonance can be significantly narrowed, improving

the tractability of structure elucidation in complex natural products.

DISCUSSION

8.1 The Program as a Constraint Refinement Search: A Parallel with Scene Analysis

The ^{13}C NMR interpretation system developed in this study operates on principles similar to those used in scene analysis within artificial intelligence. Both involve identifying a mutually consistent set of labels for a group of interrelated entities. In scene analysis, these entities are vertices in a line drawing, and the goal is to assign surface characteristics—such as shadow, convex, or boundary—to the edges around each vertex in a way that yields a coherent model of a three-dimensional scene.

Analogously, in ^{13}C NMR interpretation, the objects are carbon atoms that generate observable resonance signals. The labels to be assigned are substructural environments inferred from a database of spectrum-to-substructure rules. Each label must be consistent not only with the atom's observed chemical shift and multiplicity but also with the substructures assigned to neighboring atoms. Just as in scene analysis, where adjacent vertex labels must agree on the shared edge, here, adjacent atomic environments must form plausible covalent bonds.

8.2 Iterative Constraint Development in Spectrum Interpretation

The analysis begins with a combinatorially generated set of substructural labels for each carbon atom, derived from the observed resonance and multiplicity data. At this early stage, any atom could potentially be bonded to any other. As the algorithm iterates, it begins to identify bondable atom pairs and eliminate incompatible pairings, thereby refining the space of possible local environments for each atom.

With each cycle, the program constructs increasingly stringent connectivity constraints. These constraints narrow down the initially broad range of plausible substructures for each resonance, thereby enabling convergence toward one or a few consistent full-molecule structures. The tractability of this process is aided by the relatively small number of atoms (typically 20–30 carbon atoms) in most target molecules and by the rich informational content embedded within the substructural templates.

8.3 Data Limitations: A Bottleneck for the Interpretation Procedure

One major challenge for rule-based ^{13}C spectrum interpretation is the incompleteness of the rule database. Because rules are derived from known reference compounds, novel substructures present in newly studied molecules may not yet be encoded. If a resonance corresponds to an unrepresented environment, the program is incapable of assigning a valid label, and the iterative analysis will ultimately fail to yield any consistent structure. This is a critical limitation—especially in natural products chemistry—where new substructural motifs are frequent.

For instance, the current database contains fewer than 200 one-bond environments and around 1,100 two-bond environments for quaternary alkyl carbons. However, even with common atom and bond types, it is estimated that several thousand such environments are theoretically possible. Many of these are absent in the existing data due to bias toward well-studied compound classes.

8.4 Comparing Spectrum Interpretation with Spectrum Prediction

In contrast to interpretation, the spectrum-prediction/structure-evaluation approach is less sensitive to database incompleteness. In prediction, the program is supplied with one or more hypothesized full structures and attempts to simulate their spectra using existing rules. The fidelity of the match between predicted and observed data serves as a quantitative measure of rule accuracy, effectively allowing the system to "know its own limits".

If a prediction relies on an over-generalized or low-confidence rule, this uncertainty is inherently reflected in the output, enabling the user or AI system to down-weight its contribution in ranking candidate structures. Such an adaptive mechanism is not available in the interpretation approach, which must blindly assume that all relevant substructural environments are represented in the database—a problematic assumption.

8.5 Intrinsic Limits of Rule-Based Interpretation

Ultimately, this limitation appears to be inherent to rule-based systems for structural elucidation. Without probabilistic or learning-based extensions, these systems cannot assess the completeness or

reliability of their initial rule set. As a result, they may confidently attempt to interpret resonances for which no appropriate substructure rule exists, leading to algorithmic collapse. This strongly motivates the integration of machine learning or data-driven rule expansion techniques, as recently demonstrated by DeepSPInN (2024) and other modern AI frameworks (Zhaorui et al., 2023).

Summary

The discussed interpretation program showcases a constraint refinement search that iteratively eliminates inconsistent substructural labels and converges toward valid molecular structures. While elegant and effective in many scenarios, its practical utility is ultimately bounded by the completeness of the underlying rule base. In future development, the combination of deterministic constraint satisfaction with probabilistic learning models could offer a more powerful and resilient solution for interpreting complex ^{13}C NMR spectra.

Acknowledgment

The authors would like to express their sincere appreciation to all those who contributed to the successful completion of this research. We are particularly grateful to the colleagues and collaborators whose insightful feedback and technical suggestions played a crucial role in shaping the methodology and interpretation of results.

We acknowledge the computational resources and tools that enabled the development, testing, and validation of the system presented in this study. The authors also extend special thanks to peers and reviewers whose critical evaluations helped improve the quality and clarity of the manuscript.

This work represents a collective effort, and the contributions of each co-author in terms of concept development, data analysis, and manuscript preparation are deeply valued.

REFERENCES

Chen, Y., Zhang, D., Li, W., & Zhao, Q. (2022). Deep learning in chemical structure elucidation: From spectra to structure. Nature Computational Science, 2(7), 543–552. https://doi.org/10.1038/s43588-022-00215-z

Du, J., Yu, Y., & Zhang, Y. (2021). An AI-assisted framework for automated NMR spectral analysis. Journal of Cheminformatics, 13(1), 89. https://doi.org/10.1186/s13321-021-00528-1

Gao, K., Xiong, Y., Wang, L., & Chen, J. (2023). AI-driven advancements in organic molecule identification using NMR spectroscopy. TrAC Trends in Analytical Chemistry, 162, 117013. https://doi.org/10.1016/j.trac.2023.117013

Hu, X., Zhao, Y., & Jiang, L. (2023). Graph neural networks for molecular structure analysis from NMR data. Artificial Intelligence in the Life Sciences, 3, 100065. https://doi.org/10.1016/j.ailsci.2023.100065

Huang, B., & von Lilienfeld, O. A. (2021). Quantum machine learning for chemistry and physics. Nature Reviews Chemistry, 5, 347–358. https://doi.org/10.1038/s41570-021-00255-2

Jiang, J., Qian, Z., & Wang, M. (2022). Smart interpretation of ^{13}C NMR spectra using hybrid neural networks. Analytica Chimica Acta, 1208, 339823. https://doi.org/10.1016/j.aca.2022.339823

Liu, F., Wang, Y., & Chen, T. (2022). Integrating AI into structure-based chemical shift prediction. Molecular Systems Design & Engineering, 7(4), 495–504. https://doi.org/10.1039/D2ME00035F

Luo, H., Li, H., & Wang, J. (2021). Automated annotation of NMR spectra using ensemble learning models. Analytical Chemistry, 93(8), 3893–3902. https://doi.org/10.1021/acs.analchem.0c04918

Paracha, W. T., Inam, H., & Manzoor, M. (2025). HEARTSMART: Improved CVD risk prediction via recursive feature elimination: Validation on extended dataset. Spectrum of Engineering Sciences, 3(6), 1093–1120.

Roy, K., Kar, S., & Das, R. N. (2022). NMR-based metabolomics and artificial intelligence: Applications in health and disease. Metabolites, 12(3), 267. https://doi.org/10.3390/metabo12030267

Sharma, V., Mehta, A., & Singh, R. (2021). A review of AI-assisted tools for spectral interpretation in organic chemistry. Journal of Molecular Structure, 1243, 130780. https://doi.org/10.1016/j.molstruc.2021.130780

Wang, L., & Li, C. (2021). Predicting molecular structures from spectra using multitask neural networks. Journal of Cheminformatics, 13, 57. https://doi.org/10.1186/s13321-021-00526-3

Wu, Z., Li, X., & Xu, M. (2021). Deep spectral learning for chemical shift prediction. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4846–4858. https://doi.org/10.1109/TNNLS.2020.3047687

Xu, Y., Lin, K., Wang, S., & Zhang, X. (2022). Neural network-based prediction of NMR spectra for organic compounds. Scientific Reports, 12, 7893. https://doi.org/10.1038/s41598-022-11491-w

Yan, H., Li, T., & Zhao, S. (2023). Explainable AI in NMR spectral prediction: Advances and challenges. Computers in Biology and Medicine, 158, 106789. https://doi.org/10.1016/j.compbiomed.2023.106789

Zeng, Q., Huang, Y., & Lin, Y. (2021). Interpreting molecular fingerprints from AI-predicted NMR data. ACS Omega, 6(31), 20065–20075. https://doi.org/10.1021/acsomega.1c02456

Zhao, J., Wang, Y., & Zhang, H. (2021). DeepNMR: A convolutional neural network model for 1H NMR spectral interpretation. Molecular Informatics, 40(11), 2100127. https://doi.org/10.1002/minf.202100127

Zong, C., Sun, W., & Liu, B. (2022). ML-Shift: Machine learning based chemical shift prediction and its application to automated structure elucidation. Computational and Structural Biotechnology Journal, 20, 1654–1665. https://doi.org/10.1016/j.csbj.2022.03.011

Zou, Y., Wang, L., & Zheng, P. (2022). AI-based spectral interpretation in complex mixtures: An emerging tool for structural biochemistry. Frontiers in Chemistry, 10, 886791. https://doi.org/10.3389/fchem.2022.886791

Zurkirchen, P., Pecher, J., & Reymond, J. L. (2021). NMRShiftDB: Improving carbon shift predictions with open-access datasets and deep learning. Journal of Chemical Information and Modeling, 61(3), 1110–1119. https://doi.org/10.1021/acs.jcim.0c01160

Żyła, A., Szulc, M., & Nowak, T. (2023). NMR automation and AI-assisted spectral assignment: Trends and challenges. Chemometrics and Intelligent Laboratory Systems, 237, 104831. https://doi.org/10.1016/j.chemolab.2023.104831