Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

INFERCNMR: A ^{13}C NMR Interpretive Library Search System

Plamen N. Penchev, Klaus-Peter Schulz, and Morton E. Munk*

Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287, United States

ABSTRACT: INFERCNMR is an automated ^{13}C NMR spectrum interpretation aid for use either as a stand-alone program or as a component of a comprehensive, computer-based system for the characterization of chemical structure. The program is an *interpretive* library search which requires a database of assigned ^{13}C NMR spectra. An interpretive library search does not require *overall* structural similarity between an unknown and a library entry in order to retrieve a substructure common to both. Input consists of the chemical shift and one-bond proton-carbon multiplicity of each signal in the spectrum, and the molecular formula of the unknown. Program output is one or more substructures predicted to be present in the unknown, each of which is assigned an estimated prediction accuracy.

INTRODUCTION

The reduction of the spectral properties of a compound of unknown structure to their structural implications is a central component in modern structure elucidation. Although experienced spectroscopists and chemists are skilled in the interpretation of spectral data, considerable effort in recent years has been devoted to the development of computer software to serve as a competent "stand-in" for these scientists.[1−6] The increasing demand in solving structure elucidation problems and the availability of large, information-rich libraries of spectral data served as powerful incentives for such efforts.

^{13}C NMR spectrometry in particular has received considerable attention because of its power as a probe of the skeletal backbone of an organic compound, information not as readily accessible by other spectral methods. Current instrumentation is at an advanced state of development and routinely produces well resolved carbon signals, even in the case of complex molecular structures.

A variety of techniques have been utilized in developing computer-based programs for the interpretation of the spectroscopic data commonly used in organic chemistry: rule-based systems,[7−9] pattern recognition[10−17] and library search.[18−26] However, the boundaries between these classes are not always sharp. Rule-based systems rely on a knowledge base for the interpretation of the spectral data, e.g., a hierarchical set of rules, which could, but need not be derived from a spectral library. In contrast, pattern recognition programs and library search routines each require an appropriately constituted reference spectral library. In general, rule-based systems and pattern recognition reveal only those structural features for which they are explicitly programmed. However, library search systems can, but need not require preselection of substructures to be predicted.

In the application of the library search to ^{13}C NMR spectra, the database is a library of assigned spectra, including both spectral and structural information. The library search at its core is a spectrum matching procedure. Matched signals must be within an established tolerance and are usually required to be of the same multiplicity. The library search can be conducted in one of two ways. In the similarity library search, the spectrum of the unknown is compared to each of the entries in a reference library of assigned ^{13}C NMR spectra. The search retrieves spectra from the reference library which are "similar" overall to that of the unknown, as judged by the applied similarity measures. Operating on the premise that compounds with similar spectra are likely to be structurally similar, the similarity search may reveal the class of structure to which the unknown belongs, for example, steroid and nucleoside, and possibly more detailed structural information.[19] If a specific substructure is found to be present in many or all of the retrieved structures, its presence in the unknown may be inferred.[26]

Overall similarity between an unknown spectrum and reference spectrum is a requirement of the similarity library search. In contrast, overall similarity between the spectrum of the unknown and that of a reference compound is not a requirement of the interpretive library search. The interpretive library search is basically a *subspectrum* matching procedure. It serves to retrieve *substructures* from the reference compounds of a library of assigned ^{13}C NMR spectra which are predicted to be present in the unknown. The method takes advantage of the fact that each signal of a ^{13}C NMR spectrum contains information about the chemical environment of a single carbon atom (or atoms belonging to the same symmetry class). As a consequence, correlations between ^{13}C NMR *subspectra* and *substructures* can be more informative than those derived from other spectroscopic methods. The procedure is based on the premise that if an unknown and a reference library compound share a *subspectrum* in common, and if the subspectrum corresponds to a substructure, (i.e., a single unit of connected carbon atoms), the *substructure* assigned to the reference subspectrum is also present in the unknown.

Interpretive library searches are of two types: those limited to the retrieval of predefined substructures and those able to

Received: December 21, 2011
Published: May 17, 2012

retrieve any substructure contained in the reference compounds of the library. In practice, both require a large, structurally diverse, high-quality database of assigned ^{13}C NMR spectra. The former approach can be illustrated by work of Bremser,[27] one of the earliest to explore the merit of the method. A list of predefined substructures to be predicted is first compiled. A reference file is then created which correlates each predefined substructure with its subspectrum. The file includes the structure of the substructure and the chemical shift and signal multiplicity of each of its carbon atoms, information derived from the database. Using the set of signals of the unknown as input, the reference file is searched for sets of unknown signals that match (typically within 1−2 ppm) subspectra assigned to the predefined substructures. The match must meet a set of conditions to be considered a substructure prediction. A similarity number is calculated for each substructure to rank-order the predictions. The method forms the basis of an automated structure elucidation system, SpecSolv.[28] More recently, an interpretive library search, SISTEMAT, has been reported which is capable of inferring substructures whose carbon atoms are assigned ^{13}C NMR chemical shifts[29] (see Current Related Work below).

This paper describes the ongoing development of INFERCNMR, a ^{13}C NMR interpretive library search program designed to identify substructures of reference compounds in a library of assigned spectra that model substructures of a compound of unknown structure. As demonstrated by the results reported in this paper, the program can serve as a useful stand-alone tool for the interpretation of one-dimensional ^{13}C NMR spectra. However, it can also play a central role in the enhancement of the performance of SESAMI, an interactive, comprehensive, computer-based structure elucidation system.[30] SESAMI is built on a foundation of two major capabilities: spectrum interpretation (Program INTERPRET) and structure generation (Program HOUDINI). Therefore, its effectiveness is heavily dependent on in-depth spectrum interpretation. The pool of substructural inferences generated by INTERPRET from the collective spectral data must be sufficiently rich in information content to dramatically limit the number of compatible molecular structures, preferably to one, produced by HOUDINI. INFERCNMR can enrich that information pool. Such enrichment is especially important in solving the structure of unknowns where the ratio of hydrogen atoms to carbon atoms is low, specifically in compounds with a high degree of unsaturation such as aromatic compounds. SESAMI utilizes 2D NMR data in generating substructural inferences, however, in cases of low hydrogen−carbon ratios, the structural information derived from 2D NMR experiments can be limited. Without additional substructural inferences the effectiveness of SESAMI is reduced leading to an increased number of candidate structures.

The original version of INFERCNMR[24] focused on the implementation of several features important to an effective interpretive library search program. The intrinsic insensitivity of the program to the presence of nonmatching signals, that is, those signals in the unknown spectrum for which no matches were found in the reference spectrum, is one of its strongest assets. A subspectrum match is the only requirement for a substructure prediction. Predicted substructures need not be predefined. In fact, there are no restrictions placed on the nature or size of the predicted substructures. (The user can set the minimum number of signals to be matched.) A graphical output, in which the predicted substructures are embedded in the reference compounds from which they were retrieved, enhances the value of the information. However, this version of INFERCNMR failed to rise to the level of performance currently required in structure elucidation.

First, inferences were not assigned an estimated prediction accuracy. It is important for the user to have a sense of the reliability of a substructure prediction since if it is assumed to be correct, every proposed structure of the unknown will include the substructure. If in fact the substructure is incorrect (a false positive), every molecular structure proposed (or produced by SESAMI) will be incorrect. (Often in the case of a false positive, the user of SESAMI is alerted to the error because *no* molecular structures are produced by the program due to a contradiction between the incorrect substructure and information contained in the other inferences produced by INTERPRET.) Building on earlier work,[31,32] a more effective procedure for estimating prediction accuracy has been developed.

Second, with regard to its application in SESAMI, the information content of the predicted substructures needs to be enhanced. Although substructures predicted by the original version of INFERCNMR are already rich in information content (they contain at least six carbon atoms and are explicitly defined: atom type, hydrogen multiplicity, bond type), they can be further enriched by assigning the appropriate chemical shifts from the spectrum of the unknown to the carbon atoms of the predicted substructure. The performance of the structure generator HOUDINI is significantly improved by the addition of such information.

PROGRAM INPUT AND OUTPUT

INFERCNMR input consists of the signals of the spectrum of the compound of unknown structure—chemical shift and signal multiplicity—and its molecular formula. The output of INFERCNMR is one or more explicitly defined substructures. Heteroatoms attached directly to the carbon atoms of a retrieved substructure are considered to be part of the predicted substructure. With a large spectral library, the number of predicted substructures can be large, especially in the case of high molecular weight molecules. Two management tools are built into the program to facilitate application of the search results.

First, an estimated prediction accuracy is assigned to each substructure. In practice, users of the program limit their examination of the output to substructures with what for them is a minimum acceptable prediction accuracy, usually 90% or 95%.

Second, if the user so chooses, the selection of informative substructures to be considered can be further reduced, while retaining the structural diversity of the original set, based on the *concept of domination*.[33] The set of predicted substructures with the minimum acceptable estimated accuracy can be reduced to a smaller, *nondominated set* without loss of information based on two criteria: (1) information content and (2) prediction accuracy. Substructure A is said to dominate substructure B if A is better than B in at least one of the criteria and B is not better than A in either of the criteria. Structure B is *not* included as a member of the nondominated set.

In applying the information content criterion to two substructures, the goal is to favor the richer of the two substructures, while taking care not to eliminate the other if it is potentially useful. For example, if a substructure, 1, is a superstructure of substructure 2, it is richer in information content. If substructure 1 also has a higher estimated prediction accuracy than substructure 2, it is better than substructure 2 in at least one of the criteria and substructure 2 is not better than substructure 1 in either criteria.

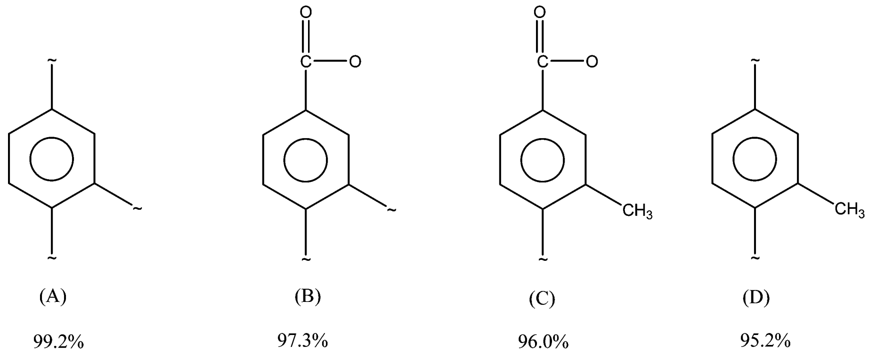

Figure 1. INFERCNMR search output.

Therefore, substructure 1 dominates substructure 2 and the latter is dropped from the nondominated set. If, however, the less information-rich substructure 2 has a higher estimated prediction accuracy than substructure 1, substructure 2 is better than substructure 1 in one of the two criteria. Therefore, neither substructure dominates the other: both 1 and 2 are retained in the nondominated set. Substructures not related as structure-superstructure cannot dominate one another.

It is important to note that for application to INFERCNMR, information content includes not only structure but also chemical shift assignments to each of the carbon atoms of the substructure. Thus, two substructures identical in structure, may differ in information content because of differences in chemical shift assignments. This concept is illustrated in the Independent Program Evaluation section.

Consider the substructures from an INFERCNMR search shown in Figure 1. (Note: The tilde (~) represents a free valence site.) The nondominated set consists of the first three substructures: A, B, and C. Although C is a superstructure of A and B, A and B are retained in the nondominated set because they each have a higher predicted accuracy than C. Substructure D is a substructure of substructure C, but it is assigned a lower prediction accuracy than substructure C. Substructure D is inferior with regard to each of the two criteria when compared to C and is therefore not included in the nondominated set.

INFERCNMR also assigns one or more signals from the observed spectrum to the carbon atoms of the predicted substructure. Where there is ambiguity in assignment, a set of alternative chemical shifts is assigned to a particular carbon atom of the substructure. In the graphic display of the output, each predicted substructure is highlighted and embedded in the reference compound from which it was retrieved.

■ THE SPECTRAL LIBRARY

The quality and diversity of the substructural inferences produced by INFERCNMR are dependent in part on the quality and diversity of the reference library of assigned ^{13}C NMR spectra. The entries in the reference library of 38,225 spectra used in this study meet three conditions: (1) they contain a minimum of six different signals in the spectrum; (2) each carbon atom of the reference compound has been assigned a single chemical shift; and (3) the chemical shift assigned to each carbon atom of the reference compound is consistent with the chemical shift range that has been independently assigned to such a carbon atom with the same first-layer nearest neighbors as those in the reference compound. Carbon chemical shifts in the library have been rounded to the nearest tenth of a part per million.

The structure representation utilized in the library allows for a very efficient execution of the tasks performed within INFERCNMR. Topological equivalence classes of the atoms of the reference compounds and their two-dimensional coordinates are included. Aromaticity and tautomerism play an important role in substructure matching, which is important in developing the probability function for estimating the accuracy of a predicted substructure. If, for example, the library would encode aromatic systems (e.g., a substituted phenyl group) as alternating single and double bonds, there exists the possibility that identical substructures from two different compounds could be erroneously reported as a mismatch. To address this problem, the program that created the reference library used in this study identified aromatic and tautomeric bonds, and flagged each accordingly. Only tautomers which do not involve a change in carbon multiplicity (e.g., HO—C═N ↔ O═C—NH) are considered since, in contrast to tautomers which do involve a carbon multiplicity change (e.g., HO—C═C—C ↔ O═C—CH—C), no simple distinction between the two tautomers is possible based on ^{13}C NMR. In defining tautomeric bonds in the library, consider the following generic representation of the tautomeric unit:

$$\mathrm{H{-}X_A{-}[A{=}A]_{\mathit{n}}{-}A{=}X_B}$$
$$\leftrightarrow \mathrm{X_A{=}A{-}[A{=}A]_{\mathit{n}}\cdots X_B\cdots H}$$

where X is a heteroatom such as oxygen or nitrogen and X_A and X_B can be the same or different element; A is any atom; and n is 0, 1, 2, ..., n. Although compounds in the reference library containing a tautomeric unit are encoded structurally in one tautomeric form or the other, the *bonds* in these units are flagged as tautomeric bonds.

The molecular formulas of the reference compounds fall within the range:

$$\mathrm{C_{6-77}H_{0-148}N_{0-15}O_{0-38}F_{0-15}S_{0-8}Cl_{0-12}Br_{0-8}I_{0-4}}$$

The *mean* molecular formula is

$$\mathrm{C_{14.37}H_{18.54}N_{0.905}O_{2.598}F_{0.072}S_{0.131}Cl_{0.213}Br_{0.049}I_{0.007}}$$

A structure-by-structure match in the library revealed 31 225 different structures. Multiple spectral entries of the same chemical compound are retained in the library only if they represent somewhat different chemical shift assignments, which can occur because of factors such as stereoisomerism, sampling method (e. g., solvent) and instrumentation.

■ PROGRAM OVERVIEW

INFERCNMR requires a diverse spectral library of assigned ^{13}C NMR spectra and consists of three major program modules which are seamlessly joined together: *substructure prediction*, retrieving substructures from the reference library; estimation of the *prediction accuracy* of each predicted substructure; and *chemical shift assignment*.

Substructure Prediction. *Subspectrum Matching.* Substructure prediction is a two-step process designed to retrieve substructures of six or more carbon atoms (default value) from reference compounds in the library. The initial step in processing the entered spectral data, chemical shift and signal multiplicity, is a subspectrum matching procedure whose function is to search every reference spectrum in the library for a set of six or more signals that match signals of the spectrum of the unknown. Each such set is a matching reference subspectrum. An unknown spectrum signal and a reference spectrum signal match if their multiplicities are the same and if the chemical shift difference between them is equal to or less than the designated matching tolerance in parts per million.

To facilitate the subspectrum search the reference library is indexed, separately for each signal multiplicity set (singlet, doublet, triplet, and quartet), according to ^{13}C NMR chemical shift in increments of 0.1 ppm in the range −80 to 350 ppm. For *each* signal in the unknown spectrum, the matching range to a signal in a reference spectrum is its chemical shift ± the designated tolerance. Depending on the tolerance, and the location and proximity of signals in a particular reference library spectrum, a signal in the unknown spectrum may match no signal, or one or more signals in the reference spectrum.

A given subspectrum match can therefore be expressed as a table which lists the chemical shifts of the unknown spectrum in a column on the left, and, on the right, in each row, one or more reference spectrum signals which match the corresponding unknown signal within the designated tolerance. Thus, a particular *reference subspectrum*, represented by the reference signals on the right side of the table, can be a "single" reference subspectrum (i.e., a clean one-to-one match between unknown spectrum signals and reference spectrum signals) or, in the case of multiple matches to one or more unknown spectrum signals, a representation of two or more reference subspectra. For operational purposes, note that the table described above can be transposed, that is, a column of the matched reference signals on the left, and in each row to the right, one or more unknown signals which match the corresponding reference signal within the designated tolerance.

Substructure Expression. In the second step, each retrieved *reference subspectrum* is expressed as corresponding *reference substructures*. (Note that a reference subspectrum of *n* signals can give rise to two or more substructures of less than *n* atoms, that is, a disjoint substructure.) In addition to the information in the above-described table, the connection table for the reference compound from which a substructure is retrieved is required in this second step. Thus, the specific atoms in the reference compound to which each signal of the reference subspectrum corresponds are accessible.

As currently programmed, INFERCNMR, given an entry spectrum, executes separate searches at tolerances from 0.0 to 2.0 ppm (default value) in increments of 0.1 ppm, for a total of 21 searches. For the purpose of describing the procedure of substructure expression, consider a subspectrum retrieved from a reference spectrum match at a *single tolerance*.

Substructure expression is a *breadth-first expansion* starting at selected root nodes. A root node initiating breadth-first expansion is any carbon atom in the *corresponding reference compound* whose chemical shift (1) is an entry in the *reference subspectrum*, and (2) differs in ppm from the chemical shift of a matched unknown signal by the *exact* tolerance being studied. In the course of the stepwise substructure expansion, an atom from the reference compound is added to the root node (or *evolving* predicted substructure) if three conditions are met: (1) the chemical shift of the added atom belongs to the reference subspectrum, (2) the added atom is not topologically equivalent to an atom in the reference compound already used in the expansion of the substructure, and (3) there is a one-to-one mapping possible between reference substructure atoms and unknown spectrum signals. One-to-one mapping is determined by finding the maximum matching in the bipartite graph formed by the signals of the unknown spectrum and the substructure atoms.[34] If an atom fails the third condition but is not part of a generated substructure from the current subspectrum match, it can be considered as a root node. However, an atom cannot be used as a root node more than once. If an atom qualifying as a root node ends up in a predicted substructure, it can no longer serve as an initiating root node.

The characteristics of substructure expression are again best clarified by considering a specific unknown spectrum-reference spectrum match at a single tolerance. The procedure described, although not theoretically exhaustive, has in practice, in the limited number of examples studied in detail, been exhaustive; all possible *different* substructures of six carbon atoms or more have been generated. However, in the event a valid substructure would fail to be generated, it would not be a fatal error; only the loss of one piece of information, generally in a pool of many.

The program produces the largest possible substructures, that is, none will be a substructure of another substructure (unless the reference compound contains two or more nonoverlapping instances of the same substructure). Since an atom cannot be used as a root node more than once, the maximum number of substructures that can be produced is equal to the number of matched signals. However, in practice, that upper limit was not reached in this study.

Bond types (single, double, triple) by which those atoms of the retrieved substructure are embedded in the reference compound (free valence sites) are included in the information content of the substructure (e.g., CH_3—CH<, CH_3—CH=).

The information content of a substructure includes heteroatoms (e.g., oxygen, nitrogen, sulfur) attached directly to a carbon atom. Additional heteroatoms attached directly to such a heteroatom are also included in the predicted substructure (e.g., NO_2). However, hydrogen atoms attached to the heteroatom are not included; they are treated as equivalent to a free valence instead. Thus, substructures with features as C-NH_2 and C−OH are reported as C−N and C−O, respectively. In these cases, the nature of the bond type at the free valence site of the heteroatom is *not* specified.

Prediction Accuracy. *Method.* In designing a procedure to estimate the probability that a substructure retrieved by INFERCNMR is correct, two approaches were considered. The principles of statistical analysis, specifically logistic regression analysis (LoRA), provide a basis for the development of a probability function.[35] Additionally, earlier studies[15] using the artificial neural network (ANN)[36] as a pattern recognition tool in spectrum interpretation suggested its application as a probability function. These approaches have two requirements in common: a large set of predicted substructures and a set of independent variables that characterize predicted substructures.

Substructure Set. The required large set of substructures is generated using the substructure prediction procedure described above. Each spectrum in the spectral library in turn serves as an inquiry which is searched against every spectrum in the library including itself. However, for each inquiry, a separate search is conducted at tolerances incremented by 0.1 ppm beginning at 0.0 ppm and terminating at 2.0 ppm, for a total of 21 searches. The search is conducted in a manner such that every substructure in a tolerance-specific set of substructures contains at least one carbon atom whose chemical shift difference is equal to the tolerance used, but no carbon atom with a greater chemical shift difference than that tolerance. Thus, each succeeding tolerance-specific set of substructures in the search sequence excludes those substructures generated at lower tolerances. The result is 21 separate sets of predicted substructures.

Incorrect predictions derived at a tolerance of 0.0 ppm (i.e., exact chemical shift matches) provided an opportunity to identify and delete reference library entries of questionable quality. In each instance of an incorrect prediction, the reference compound and its spectrum, and the "unknown" compound and its spectrum were manually examined. If, based on information in the original literature (if available) and/or accepted spectroscopic correlations, the quality of any entry was judged to be suspect, it was deleted from the library.

During the process of generating the substructure sets, the validity of each substructure prediction is determined. A substructure retrieved from a reference compound is considered to be a correct prediction if it meets two conditions. First, the predicted substructure must be superimposable in every structural detail on a set of atoms of the "unknown." This includes matching of all bonds, including the bond order of atom sites at which embedding of the substructure in the "unknown" occurs (free valence sites). Recall that predicted substructures include heteroatoms attached directly to carbon atoms. Therefore, in determining validity, a predicted substructure which possesses a free valence at a particular carbon atom, implying an unidentified carbon atom must be attached at that site since if there was an attached heteroatom in the reference compound it would be included in the predicted substructure, would *not* match a comparable substructure in the "unknown" if the corresponding carbon atom had an attached heteroatom.

The presence of aromatic and tautomeric bonds requires clarification of bond matching rules.

1. Localized bonds (neither aromatic nor tautomeric) match to
   a. Localized bonds which are of the same bond order.
   b. Tautomeric bonds (not aromatic bonds) only if the hybridization of the atoms joined by the bond matches the hybridization of the atoms of the delocalized bonds.
2. Aromatic bonds match to aromatic bonds *and* to bonds that are both aromatic and tautomeric.
3. Tautomeric bonds match to tautomeric bonds, to bonds that are both tautomeric and aromatic, and to localized bonds if condition 1b is met.

The second condition for establishing correctness of a predicted substructure pertains to chemical shift assignments. The set of chemical shifts assigned to the carbon atoms of the predicted substructure by the program (see the Chemical Shift Assignment section) must contain the actual chemical shifts assigned in the "unknown."

Independent Variables. The independent variables (Table 1) are selected for their perceived relevance to prediction accuracy. The goal is to develop a function that relates the variables characterizing a predicted substructure to its estimated prediction accuracy. A discrete probability function is developed at each of the 21 tolerances. The basis for this procedure is that some of the independent variables are tolerance-dependent (see below, e.g., sNOS$_t$ and NasS$_t$).

Table 1. Description of the 23 Independent Variables

1	uNA: number of atoms in the unknown compound
2	uNS: number of singlets in the unknown compound
3	uND: number of doublets in the unknown compound
4	uNT: number of triplets in the unknown compound
5	uNQ: number of quartets in the unknown compound
6	rNA: number of atoms in the reference compound
7	rNS: number of singlets in the reference compound
8	rND: number of doublets in the reference compound
9	rNT: number of triplets in the reference compound
10	rNQ: number of quartets in the reference compound
11	sNA: number of atoms in the substructure
12	sNS: number of singlets in the substructure
13	sND: number of doublets in the substructure
14	sNT: number of triplets in the substructure
15	sNQ: number of quartets in substructure
16	sNB: number of bonds in the substructure
17	sRMSD: minimum root mean standard deviation
18	sH: histogram signal density variable
19	sIH: inverse histogram signal density variable
20	sFV: number of free valences in the substructure
21	sNOS$_t$: the number of occurrences of a particular substructure produced in a search at tolerance t
22	sNRC: the number of reference compounds in the library containing the predicted substructure
23	NasS$_t$: the number of all substructures produced in a search at tolerance t

The variables are of three types, spectral, structural and statistical. The first five variables (1−5) describe five characteristics of the "unknown" compound (u) and its spectrum: specifically, the number (N) of atoms (A) in the unknown, and the number (N) of carbon signals of each possible multiplicity: (S, singlet; D, doublet; T, triplet; Q, quartet). The second five variables (6−10) describe the same characteristics for the reference compound (r) from which the substructure was derived. Similarity between these two sets of variables can increase the likelihood of a correct prediction. The next five variables (11−15) describe the same characteristics of the retrieved substructure (s).

In developing a probability function using a set of substructures of considerable structural and spectral diversity, scaling the variables can often enhance discrimination. That this is indeed the case in this study is suggested by the positive and statistically significant coefficients observed in logistic regression calculations upon scaling the variables listed in Table 1.

Scaled variables 1−5 (rRel), Table 2, describe characteristics of the reference compound/spectrum (variables 6−10, Table 1) relative to the corresponding characteristics of the "unknown" compound/spectrum (variables 1−5, Table 1). Scaled variables 6−10 (uRel), Table 2, are the inverse of these relationships, i.e., the characteristics of the "unknown" compound/spectrum (variables 1−5, Table 1) relative to the corresponding characteristics of the reference compound (variables, 6−10, Table 1). (The addition of "one" to the denominator of scaled variables 2−5 and 7−10 simply avoids division by "zero" in cases where either the unknown or reference spectrum lacks signals of one or more multiplicities.) Table 2 also describes the same five characteristics of the substructure relative to the unknown (s/uRel),

Table 2. Scaled Independent Variables

1	rRelAtoms = rNA/uNA
2	rRelSing = rNS/(uNS + 1)
3	rRelDoub = rND/(uND + 1)
4	rRelTrip = rNT/(uNT + 1)
5	rRelQuart = rNQ/(uNQ + 1)
6	uRelAtoms = uNA/rNA
7	uRelSing = uNS/(rNS + 1)
8	uRelDoub = uND/(rND + 1)
9	uRelTrip = uNT/(rNT + 1)
10	uRelQuart = uNQ/(rNQ + 1)
11	s/uRelSize = sNA/uNA
12	s/uRelSing = sNS/(uNS + 1)
13	s/uRelDoub = sND/(uND + 1)
14	s/uRelTrip = sNT/(uNT + 1)
15	s/uRelQuart = sNQ/(uNQ + 1)
16	s/rRelSize = sNA/rNA
17	s/rRelSing = sNS/(rNS + 1)
18	s/rRelDoub = sND/(rND + 1)
19	s/rRelTrip = sNT/(rNT + 1)
20	s/rRelQuart = sNQ/(uNQ + 1)
21	sRelH = sH/sNC
22	sIRelH = sIH/sNC
23	sRelFV = sFV/sNB
24	sFracRetr = $\mathrm{sNOS_t}$/sNRC
25	sSearchSel= $\mathrm{sNOS_t}$/$\mathrm{NasS_t}$

scaled variables 11−15, and relative to the reference compound (s/rRel), scaled variables 16–20.

Variable 16 (sNB), Table 1, describes the number (N) of bonds (B), independent of bond multiplicity, connecting atoms of the substructures. This information combined with that of variable 11, the number of atoms in the predicted substructure (sNA), provides information on the number of cycles in the substructure. In general, greater variations in chemical shift are expected among open chain fragments than in those with closed rings. Thus, the greater the ratio sNB/sNA, the more likely a prediction is to be correct.

For each predicted substructure (six or more carbon atoms), the *minimum* root-mean-square deviation (sRMSD) in chemical shift is determined (variable 17, Table 1). For the purpose of this calculation, a chemical shift from the "unknown" spectrum is assigned to each carbon atom of the retrieved substructure such that the sum of the squares of the chemical shift differences between the signals of the unknown and the signals assigned to the carbon atoms of the reference substructure is at a minimum and a one-to-one mapping of the atoms is achieved.[37] Often these chemical shift "assignments" do not correspond to the actual values. This variable is a measure of the fit between the unknown subspectrum and the matched reference subspectrum; the smaller the value of sRMSD, the better the match, and the more likely the corresponding predicted substructure is valid.

Signal density in regions of matched signals can be expected to influence reliability. Regions of high signal density are more likely to represent a diversity of structural features, leading to fortuitous signal matches, and therefore, possibly less reliable predictions. To account for this factor, signal density histograms, a plot of the *number* of signals versus chemical shift in intervals of 0.1 ppm, were prepared using the entire spectral library, one for signals of each of the four multiplicities. Two variables, sH and sIH (variables 18 and 19, respectively, Table 1), are based on this information.

Using chemical shifts "assigned" in the calculation of sRMSD (variable 17, Table 1), a value, h_i, is calculated for *each* carbon atom of the predicted substructure, where h_i is the number of signals that appear in the histogram (of the same multiplicity as the substructure atom) within the range set by the tolerance used in the search producing the substructure. $\Sigma\, h_i$ is the sum of these signal counts for the entire substructure. The histogram variable for a predicted substructure (sH, variable 18, Table 1) is defined in eq 1

$$\mathrm{sH} = \left(\sum h_i\right)/(2 \times 10 \times \mathrm{Tol} + 1) \tag{1}$$

The factor (2 × 10 × Tol +1) serves to normalize values of sH for comparison between the different tolerances without impacting the evaluation of the data. At the designated tolerance of the search, the factor is equal to the number of 0.1 ppm intervals in the histogram whose values must be summed to arrive at the appropriate value of h_i for each carbon atom in the substructure.

The scaled histogram variable sRelH (sH/sNC, variable 21, Table 2) describes the variable sH relative to the number of carbons atoms in the substructure (sNC). Signal matches occurring in regions of the spectrum of low signal density give rise to lower values of sH and predictions that are more likely to be valid. In contrast, the more matched signals occur in regions of high signal density, the greater the danger of an invalid substructure assignment. However, cases have been observed in which the majority of the unknown signals give small values of h_i but one or a few give large values. Thus, the scaled variable can be large even though the majority of matched signals occur in regions of low histogram signal density, misleadingly suggesting a less reliable prediction. To address such spurious values of sH, the sum of the *inverse* (I) of the signals counts is used in a complementary variable (sIH, variable 19, Table 1) as defined in eq 2.

$$\mathrm{sIH} = \left(\sum 1/h_i\right)/(2 \times 10 \times \mathrm{Tol} + 1) \tag{2}$$

Here, in contrast to low values of h_i, large value of h_i will add little to the sum ($\Sigma 1/h_i$). Thus, the scaled variable sIRelH (sIH/sNC, variable 22, Table 2) can clarify the value of sRelH. A high value of sH/sNC arising from a match in which a large number of signals matches occur in regions of high signal density will be accompanied by a very low value of sIH/sNC. However, a high value of sH/sNC arising from a match in which only a small number of signal matches occur in regions of high signal density will be accompanied by a "high" value of sIH/sNC.

The information content of a substructure predicted by INFERCNMR includes the exact nature of the bonding sites (free valence sites) at which the substructure is embedded in the reference compound (e.g., $\mathrm{CH_2{=}}$ is a substructure of $\mathrm{CH_2{=}CH_2}$, but $\mathrm{{-}CH_2{-}}$ is not). The number and nature of such bonding sites can influence the accuracy of a prediction. This influence is expressed as the variable sFV (variable 20, Table 1), the sum of all "free valences" in the substructure. For purposes of this calculation, a single "half-bond" counts as *one* free valence, a double "half-bond" as *two* free valences and any aromatic "half-bond" as two free valences. In general, as the value of sFV increases, prediction accuracy decreases. Matches of carbon atoms where the neighboring structural environment is well-defined, that is, distant from bonding sites, are expected to be more reliable. In contrast, carbon atoms at or near a bonding site have less well-defined immediate structural environments, and consequently matches of such atoms are less reliable. The situation is related to an observed characteristic of the interpretive library search: the predicted substructure is often a *subset* of the substructure that is actually common to unknown and reference because the outermost carbon atoms of the *actual* common substructure are in different chemical environments in the two compounds.[24] To normalize with regard to the size of predicted substructures, the sFV variable is scaled by dividing by the total of number of bonds in the substructure (sNB), that is, sRelFV = sFV/sNB (variable 23, Table 2). A low value of sRelFV favors increased prediction accuracy.

Three variables relate to "number of occurrences".

1. sNOS_t (variable 21, Table 1): the number (N) of occurrences (O) of a predicted substructure (s) produced in a search (S) at tolerance *t* (the count includes occurrences of *larger* substructures that contain the predicted substructure).
2. sNRC (variable 22, Table 1): the number (N) of reference compounds in the library (RC) containing the predicted substructures.
3. NasS_t (variable 23): the number (N) of *all substructures* (as) produced in a search (S) at tolerance *t*.

For a search at a particular tolerance *t*, the greater the fraction of reference library occurrences of a particular substructure retrieved, the stronger the case for a valid substructure prediction. Retrieval of only a few of the reference library occurrences of a particular substructure could suggest an "accidental" substructure match, i.e., a suspect substructure assignment. This property can be expressed by the variable sFracRetr (no. 24, Table 2) which is the ratio sNOS_t/sNRC, a larger value of which should contribute to enhanced prediction accuracy.

A related rationale suggests that a larger value for variable sSearchSel (no. 25, Table 2), the ratio of the number of occurrences of a particular substructure retrieved at tolerance *t* (sNOS_t) to the number of *all substructures* produced at tolerance *t* (NasS_t), indicates a more *selective* search outcome favoring a greater prediction accuracy. Conversely, a low value of sSearchSel indicates that the predicted substructure is in the minority among all predicted substructures, detracting from its reliability unless complemented by a high value for sFracRetr. Higher values of both of these variables reinforce the reliability of the predicted substructure.

Probability Function. The large set of predicted substructures required for development of the probability function consists of 21 separate sets of substructures each of which was collected at a specific tolerance (0.0, 0.1, 0.2, ..., 2.0 ppm). During the process of generating this set, the validity of each substructure prediction is determined and all 48 variables (Tables 1 and 2) are calculated for each substructure. Each of the 21 tolerance-specific sets of substructures is randomly divided into three approximately equally sized subsets. One of the three tolerance-specific subsets serves as the *learning set* for probability function training, the second serves as the *test set* for the purpose of estimating prediction accuracy, and the third serves as the *validation set* for evaluating the performance of the probability function.

The same *learning set* is used to calculate the corresponding logistic regression and to train the corresponding ANN. Calculated values for the independent variables selected (not all variables are used in each case, see Optimizing the ANN) are used as "input." The dependent variable *y* in this case is binary in nature; a prediction is either valid or invalid, and *y* is therefore assigned a value of one or zero, respectively.

Probability functions whose dependent variable has one of two values are usually curvilinear, a tilted S shape with asymptotes at 0 and 1 in the ideal case, thereby precluding solutions less than 0 and greater than one, and are referred to as logistic functions. The method used for finding such a function is called *logistic regression analysis* (LoRA).[35] The probability *p* of a prediction being valid is expressed by eq. 3, where $0 \leq p \leq 1$. Solving for *p* requires a corresponding value of *p*′ which can be determined from the independent variables *x* calculated for each predicted substructure. First-order and second-order functions have been studied. The latter is expressed by eq. 4 with cross terms, where *k*, *l*, *m* = 1, 2, ..., *N* (the number of independent variables), and *l* < *m*.

$$p = 1/(1 + e^{p'}) \qquad \text{(eq. 3)}$$

$$p' = \beta_0 + \sum \beta_k x_k + \sum \sum \beta_{l,m} x_l x_m \qquad \text{(eq. 4)}$$

Alternatively, an artificial neural network (ANN) can be used in place of the LoRA in estimating prediction accuracy. For this application, a feed-forward ANN with one hidden layer was studied. The independent variables served as input neurons. The output is a single neuron, the probability *p* that a substructure prediction is valid. A back-propagation-of-error algorithm[36] was employed in training the network. A sigmoidal squashing function is used as the transfer function, where Net_j is the net input of the *j*th neuron (eq. 5). The parameter α_j was set to unity without loss of generality,[36] and the threshold parameter θ_j and the

$$f(\text{Net}_j) = 1/(1 + \exp[-\alpha_j(\text{Net}_j + \theta_j)]) \qquad \text{(eq. 5)}$$

weights between neurons were optimized during the learning process. The ANN weights and offsets were initialized with pseudorandom values between −1 and +1.

Estimating Prediction Accuracy. In applying either LoRA or ANN, the calculated output *p* is neither 0 nor 1, but some value between the two. For real-world application, that probability *p* must be converted to estimated prediction accuracy. That requires relating *p* to an *observed* prediction accuracy in a model of the system to be studied.[38] The model in this case is based on the *test set* (the second of three subsets of predicted substructures retrieved at each tolerance). Each predicted substructure in the test set is treated as an object to be processed by either the logistic regression or the ANN developed with the training set. A probability value *p* is calculated for each predicted substructure, each of which is known to be correct or incorrect.

To relate *p* to *estimated prediction accuracy*, the interval between 0 and 1 is divided into 1001 equidistant units *i* (*i* = 0.000, 0.001, ..., 1.000). Each of the predicted substructures is assigned to the unit corresponding to its calculated probability *p*. For each of the units *i* to which substructures have been assigned, the *observed* accuracy *A*, expressed as a percentage, at *threshold i* is calculated using eq 6, where N_{i1} is the number of *all valid* predictions (hence the subscript 1) with probability values equal to and greater than *i* and N_{i0} is the number of *all invalid* predictions (denoted by the subscript 0) with probability values equal to and greater than *i*. The table of *p* versus *A* values derived from the test set serves as the

$$A_i = 100N_{i1}/(N_{i1} + N_{i0}) \qquad (6)$$

source of estimated prediction accuracies in the application of INFERCNMR to real world unknowns. The observed accuracy A that corresponds to the derived *p* value for a particular predicted substructure, using either of the two procedures described, is assigned as the *estimated prediction accuracy* for that substructure.

Chemical Shift Assignment. In an initial trial, each carbon atom of the predicted substructure was assigned one or more signals of the unknown of the appropriate multiplicity that matched the chemical shift of the corresponding atom of the reference substructure within the *particular* tolerance (between 0.0 and 2.0 ppm) used in retrieving the substructure. Since the tolerance ranges used are in fact less than the observed differences in chemical shift due to solvent variations, in practice the error rate due to chemical shift mismatches was found to be unacceptably high. Larger, but *informed* ranges were required.

Consider the predicted substructure shown in Figure 2. (Again, note that the sites designated with the tilde (~) represent sites of free valence, that is, atoms (and their bonds) at which embedding in the reference compounds occurs.) In an attempt to predict plausible chemical shift ranges for each carbon atom of the substructure, large ranges would be expected for carbon atoms with the free valence sites because of the uncertainty of the attached atom(s). It follows that, in general, the further removed a carbon atom is from a free valence site, the narrower the expected chemical shift range since the surrounding chemical environment is well-defined. This concept forms the basis of the method of chemical shift assignment in INFERCNMR.

The starting point is the *entire* set of predicted substructures (from all 21 tolerance-specific sets) which are *correct with regard to structure.* Recall that each member of the reference library serves as an "unknown" in generating the set of predicted substructures. Since the "unknowns" are in fact known in this study, the chemical shifts assigned to the carbon atoms of the predicted substructures are known. Thus, chemical shift differences between the actual chemical shifts assigned to the carbon atoms of the predicted substructure and those assigned to the corresponding reference substructure can be determined.

Using all carbon atoms of all of the predicted substructures, separate histograms were prepared of the chemical shift difference (in units of 0.1 ppm) between corresponding carbon atoms and the distance of that carbon atom from the nearest free valence site. i.e., a separate histogram was created for distance 0 (a free valence site), 1, 2, 3, ..., and 8.

For the substructure pictured in Figure 2, carbons atoms *a* and *h* are assigned to the histogram representing distance 0 and atoms *b*, *e*, and *g*, to the histogram representing distance 1. Atom *c* is assigned to distance 2 (two bonds removed from the closest free valence site, atom *a*); atom *d* is assigned to histogram distance 3 (it is three bonds removed from *two* free valence sites, atoms *a* and *f*). Note that a heteroatom (oxygen) with a free valence is used in measuring distance. In establishing the applicable chemical shift ranges per distance, one percent of the entries at either end of each histogram were considered outliers and deleted. The ± ppm ranges shown in Table 3 are derived from the larger of the two values at the ends of the histogram. However, a small increment was added to further broaden the applicable shift range since it is better to include an incorrect chemical shift assignment than to exclude the correct assignment. In the case of distance 7, an examination of the histogram indicated a larger complement of outliers at the high end; consequently, the value was reduced accordingly. Histograms for distances greater than 8 were determined, but the number of entries was low and the obtained ranges were less than differences expected because of solvent variation.

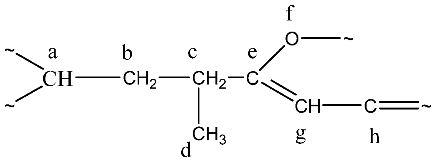

Figure 2. Chemical shift assignment.

The assignment of chemical shift to the carbon atoms of a predicted substructure derived from a compound of unknown structure begins with the known chemical shifts assigned the corresponding atoms of the reference substructure and the distance of each carbon atom of the reference substructure to the nearest free valence site. Then using the appropriate chemical shift range (Table 3) for each carbon atom, *all* signals of the spectrum of the unknown that match the reference spectrum signal within the range specified are assigned to the corresponding carbon atom of the predicted substructure. Given the breadth of the ranges shown in Table 3, multiple chemical shift assignments to carbons atoms do occur.

Table 3. Chemical Shift Assignment Windows

distance to free valence	1st percentile (ppm)	99th percentile (ppm)	number of data points	shift window (ppm)
0	−15.4	15.3	1 693 893	±15.5
1	−11.6	11.8	1 435 763	±12.0
2	−7.9	9.5	493 126	±9.7
3	−5.3	4.8	194 370	±5.5
4	−3.3	3.7	102 410	±4.0
5	−2.0	2.6	75 790	±3.0
6	−1.4	2.0	31 088	±2.2
7	−1.3	4.2	9 285	±2.2
8	−1.2	2.2	3 057	±2.2

■ RESULTS AND DISCUSSION

Criteria for Performance Evaluation. The performance of the probability functions for estimating prediction accuracy can be evaluated in terms of two characteristics: (1) *discrimination*, the ability to distinguish between valid and invalid predicted substructures, and (2) *recall*, at a given required level of accuracy, the fraction of valid substructures which are predicted to be correct, for example, recall at 95% accuracy. In most applications of INFERCNMR, retrieval of substructures with high estimated predicted accuracies will be of major interest. Therefore, in this study, performance is evaluated at estimated prediction accuracies of 90%, 95%, and 99%.

As indicated earlier, INFERCNMR executes an interpretive library search by collecting retrieved substructures separately at 21 tolerances ranging from 0.0 to 2.0 ppm. In addition to the tolerance-specific nature of some of the independent variables, this approach also serves to enhance discrimination since in estimating the accuracy of a predicted substructure, that probability function is used which was trained at the same tolerance as that at which the predicted substructure was produced. Data manageability is also improved since operating at multiple tolerances reduces the size of substructure sets to be processed and therefore the computation times.

As applied to INFERCNMR, the LoRA model demonstrated inferior performance to the ANN model. A linear LoRA model without cross terms resulted in poor discrimination and low recall. The addition of cross terms at low tolerances improved performance, but to a level significantly less than that of the ANN model. The addition of cross terms at higher tolerances proved to be impractical due to prohibitive computational times. Consequently, the ANN model was chosen for further development.

Table 4. Number of Substructures Generated As a Function of Tolerance (ppm)

Tol. (ppm)	no. of substrates	no. correct	percent correct	Tol. (ppm)	no. of substrate	no. correct	percent correct
0.0	74 199	74 055	99.8	1.1	704 196	349 045	49.6
0.1	78 930	77 896	98.7	1.2	820 156	368 910	45.0
0.2	160 802	156 008	97.0	1.3	941 369	386 046	41.0
0.3	219 580	206 963	94.3	1.4	1 071 969	402 392	37.5
0.4	264 050	239 861	90.8	1.5	1 212 608	422 370	34.8
0.5	309 536	265 738	85.9	1.6	1 366 983	441 641	32.3
0.6	349 269	280 182	80.2	1.7	1 532 637	464 198	30.3
0.7	392 613	290 308	73.9	1.8	1 695 074	483 547	28.5
0.8	455 005	306 455	67.4	1.9	1 868 455	500 226	26.8
0.9	523 833	318 392	60.8	2.0	2 044 446	524 550	25.7
1.0	605 186	332 017	54.9				

Optimizing the ANN Model. Back propagation neural networks have been shown to be capable of treating complex relationships,[36] however, optimal values of network parameters are required for optimal performance. Optimal values of three of these network parameters in particular proved to be dependent on a number of factors: (1) the tolerance, (2) the type of input neurons (i.e., the specific independent variables selected from Tables 1 and 2), (3) the number of input neurons, and (4) the number of hidden neurons.

The performance of back-propagation neural networks has been shown to be sensitive to the relationship between the number of "objects" (predicted substructures) in the learning set and the number of network weights (connections). Given the chemist's need for a model of high predictive reliability and the large number of substructures generated at each tolerance, the neural networks were constituted such that the *square* of the number of network weights was equal to or less than the number of substructures in the learning set. Since the network has a single *output neuron*, in practice, the number of network weights is dependent only on the number of input and hidden neurons. The number of hidden neurons considered in this study varied from 3 to 20.

Table 4 summarizes the number of *new* substructures retrieved at each of the 21 tolerances studied. As expected, the number of substructures predicted increases with increasing tolerance. Each of the three *subsets* of each tolerance-specific set of substructures, the *learning set*, the *test* set, and the *validation* set, consists of approximately one-third of the set of substructures, for example, approximately 24 700 substructures at 0.0 ppm and 681 500 substructures at 2.0 ppm.

For the purpose of optimizing the ANNs, the tolerance range (0.0−2.0 ppm) was divided into 5 segments (Table 5). For each segment, the average of the number of substructures was calculated. Next, for each of the five segments, a *number* of input neurons was selected such that the square the number of weights in a network of 20 hidden neurons was equal to or less than the average number of substructures. The results are summarized in Table 5. In the two smallest tolerance segments, 0.0 ppm and 0.1−0.2 ppm, the base set of 23 variables (with the exception of sRMSD in the 0.0 ppm segment since it has a value of zero for all substructures), which contains all of the basic information, and the two histogram variables (variables 21 and 22, Table 2, the most important of the scaled variables) are used. (Using all 48 variables and 3 hidden neurons with the substructures produces at 0.0 ppm (24,700) led to overtraining of the network after only one epoch.) The scaled variables 23, 24, and 25 (Table 2), variables derived from the base set and judged to be next in importance, are added in the 0.3−0.4 ppm tolerance segment. All variables of Table 2 with the exception of variables 6−10 and 16−20 (which provide parallel information to variables 1−5 and 11−15, respectively) are added to the 0.5−1.5 ppm tolerance segment for a total of 38 variables. All 48 variables are used in the 1.6−2.0 ppm tolerance segment.

Table 5. Selection of Variables as Input Neurons for ANN Training

group no.	tolerance range included (ppm)	number of input neurons	variables selected as input neurons
1	0.0	24	All variables from Table 1 except sRMSD and the two histogram variables (no. 21 and 22, Table 2).
2	0.1−0.2	25	All 23 variables from Table 1 and the two histogram variables (no. 21 and 22, Table 2).
3	0.3−0.4	28	All 23 variables from Table 1; the two histogram variables (no. 21 and 22, Table2); and variables 23, 24, and 25 (Table 2)
4	0.5−1.5	38	All 23 variables from Table and variables 1−5; 11−15 and 21−25 (Table 2).
5	1.6−2.0	48	All 48 variables from Table 1 and Table 2.

ANN network parameters were optimized during training, except for the learning rate η and the momentum factor μ. The latter two parameters were determined independently (0.7 and 0.3, respectively) by training with significantly smaller learning subsets. To determine the optimal number of hidden neurons for each network at each tolerance, each of the 21 tolerance-specific *learning sets* were trained through 30 epochs, first with 3 hidden neurons, and in sequence, up to 20 hidden neurons (for a total of 18 networks at each of the 21 tolerances). After each of the 30 epoch sessions and before the next session, the *test set* (at the same tolerance) was run in *predictive* mode using the ANN of that session. Two types of data were collected at each tolerance for each of the 18 networks trained with different numbers of hidden neurons: (1) root-mean-square error (a measure of the difference between calculated output value and target value) for both learning set and test set plotted as a function of epoch number; (2) recall at 99% accuracy for both sets as a function of epoch.

As expected, root-mean-square error decreases with increasing epochs until a plateau is reached. For each of the 21 tolerances studied, the 18 root-mean-square error plots for both learning set and test set were examined. Three of the best hidden neuron plots were selected as follows. For each of the 18 learning set and test set plots of root-mean-square error versus epoch, the *range* of small difference between root-mean-square errors is identified, since larger differences indicate nongeneralizing networks. Within those eighteen ranges, the three plots which displayed the smallest root-mean-square error of the learning set were selected.

The choice among the three surviving hidden neuron plots at each of the 21 tolerances was selected as follows. Each of the three plots of recall at 99% accuracy versus epoch for both learning set and test set were examined to identify the range of small difference in recall between the two. That hidden neuron plot was selected which displayed the maximum recall at 99% accuracy within the identified range. The optimal number of hidden neurons at each tolerance is summarized in Table 6. This study also suggested the optimal number of epochs in training each network.

Table 6. Optimum Number of Hidden Neurons at Each Tolerance

tolerance	no. of hidden neurons
0.0	3
0.1	3
0.2	4
0.3	4
0.4	4
0.5	5
0.6	6
0.7	6
0.8	6
0.9	7
1.0	9
1.1	10
1.2	9
1.3	10
1.4	10
1.5	12
1.6	10
1.7	14
1.8	15
1.9	16
2.0	17

In applying INFERCNMR in structure elucidation, a user could expect a substructure predicted with high estimated prediction accuracy to be present in the compound of unknown structure. If that substructure prediction is in fact invalid (a false positive), the molecular structure proposed for the unknown will be incorrect, a fatal error. In contrast, a valid substructure predicted with a low estimated prediction accuracy (a false negative) would not be expected to be present in the unknown by the user; a loss of valuable information, but not a fatal error.

To increase recall in using the ANN model, the *standard* back-propagation-of-error algorithm used in training the ANN was transformed into a *target-weighted* one. This required two minor changes. First, the equation for the correction, δ, of the output neuron during training is modified by multiplication with the user-defined weighting function, $W(T)$, which depends solely on the target value T, where T is either 0 or 1 (eq 7). Second, the calculated mean error, S_α, between T and Y for data set α (e.g., the learning set) must reflect the resulting target weights as in eq. 8 (where N_α is the number of objects in the data set and k designates a specific object).

$$\delta = (T - Y)Y(1 - Y)W(T) \text{ where } W(T) > 0 \; \forall \; T \qquad (7)$$

$$S_\alpha^{\ 2} = \sum (T_{\alpha,k} - Y_{\alpha,k})^2 W(T_{\alpha,k})/N_\alpha \; (k = 1, 2, ..., N_\alpha) \qquad (8)$$

The function $W(T)$ serves to bias the error: $W(0) > W(1)$ favors false negatives and $W(0) < W(1)$ favors false positives. The net effect of favoring false negatives is to decrease the probability of false positives and thereby increase recall. The value of $W(T)$ that maximizes recall was independently determined prior to full optimization of the ANNs. For this purpose, the value of $W(0)$ was set to 1.0 (it is the ratio of $W(0)$ to $W(1)$ that is important) and the value of $W(1)$ was varied from 0.05 to 1.20. (When $W(1) = 1$, no bias is introduced.) Studies were carried out at 90%, 95%, and 99% accuracy. Observed differences in recall with changes in the value of $W(1)$ were greatest at 99% with maximum recall at $W(1) = 0.1$. Since substantially smaller differences in recall were observed at 90% and 95%, $W(1)$ was set to that value thereby achieving close to optimal performance at all three levels of accuracy.

Neural Network Performance. Substructures of the *validation subsets* serve as the basis for evaluating the performance of the 21 optimized, tolerance-specific neural networks. The information required for the evaluation includes: (1) the validity and estimated prediction accuracy of each validation *substructure*; and (2) recall at 90%, 95%, and 99% accuracy for each validation *set*. The results at five selected tolerances in the range from ±0.3 ppm to ±2.0 ppm are summarized in Table 7. Data for the validation sets (VS) are presented along with the comparable data obtained for the corresponding learning sets (LS) and test sets (TS).

The third set of columns (predictions) records the number of retrieved substructures in each of the three sets at each of the five tolerances and the percentage of correct substructures. Again, the numbers represent new substructures not previously retrieved at lower tolerance. As expected, as the width of the tolerance increases, the number of substructures retrieved increases, but the percentage of correct substructures decreases. For example, in the validation sets, the number of substructures retrieved increases from 72 798 at ±0.3 ppm to 690 229 at ±2.0 ppm, while the percentage of correct substructures decreases from 93.3% to 25.3%. Each succeeding set of three columns (90% accuracy; 95% accuracy; and 99% Accuracy) records the number of substructures predicted at *estimated* accuracies of 90%, 95%, and 99%; the *actual* prediction accuracy for each set; and the percentage of valid substructures captured (recall). Note that in the case of the test sets (TS), *estimated prediction accuracy* equals *actual prediction accuracy* (see Estimating Prediction Accuracy) except at ±0.3 ppm where the procedure gives an accuracy of 98.9% instead of 99%.

The data reveal some noteworthy trends regarding recall. At any tolerance, recall decreases with an increasing requirement for accuracy. For example, at a tolerance of ±0.5 ppm, recall for the validation set decreases from 98.2% at 90% accuracy to 64.2% at 99% accuracy. Thus, the price for demanding higher accuracy at any given tolerance is a loss of information, i.e., fewer predicted substructures for a given unknown compound. At all three prediction accuracies, recall also decreases with an increase in tolerance from 0.3 ppm to 1.5 ppm. At 90% accuracy, a further increase in tolerance to ±2.0 ppm results in no further decrease in recall, while at accuracies of 95% and 99%, recall actually improves with the increase to ±2.0 ppm. The apparent anomaly is explicable in terms of the relative influence of the variables used in training the ANNs. With the large number of substructures retrieved at the high tolerances,

Table 7. Neural Network Performance at Estimated Accuracy of 90%, 95%, and 99%

		predictions		90% accuracy[b]			95% accuracy			99% accuracy		
Tol. (ppm)	set	no. of substructures	percent correct	no. predicted	accuracy (%)	recall (%)	no. predicted	accur acy (%)	recall (%)	no. predicted	accuracy (%)	recall (%)
±0.3	LS[a]	78 176	95.0				77 581	95.5	99.8	56 070	99.4	75.1
	TS[a]	68 606	94.4				68 044	95.0	99.8	47 986	98.9	73.2
	VS[a]	72 798	93.3				72 202	93.9	99.8	52 976	99.0	77.2
±0.5	LS	107 206	85.5	101 744	88.5	98.3	89 537	94.7	92.5	60 948	99.5	66.2
	TS	99 521	86.8	94 183	90.0	98.1	83 515	95.0	91.9	57 615	99.0	66.0
	VS	102 809	85.3	97 485	88.4	98.2	86 021	93.4	91.6	57 030	98.7	64.2
±1.0	LS	206 457	55.3	93 451	90.7	74.3	71 397	96.5	60.4	29 855	99.5	26.0
	TS	198 158	54.8	88 052	90.0	73.0	66 204	95.0	57.9	26 219	99.0	23.9
	VS	200 571	54.5	88 071	89.9	72.5	66 513	94.9	57.8	27 977	98.9	25.3
±1.5	LS	407 991	35.2	69 336	92.6	44.7	36 609	97.7	24.9	12 541	99.7	8.7
	TS	399 873	34.7	65 796	90.0	42.7	34 844	95.0	23.9	11 930	99.0	8.5
	VS	404 744	34.6	66 611	90.9	43.2	35 346	96.4	24.3	11 513	99.4	8.2
±2.0	LS	684 443	26.1	81 928	93.5	42.8	60 276	97.5	32.9	31 651	99.6	17.6
	TS	669 774	25.6	76 093	90.0	40.0	53 956	95.0	29.9	25 752	99.0	14.9
	VS	690 229	25.3	79 030	89.4	40.5	56 991	94.0	30.7	28 713	98.2	16.2

[a] LS = learning set; TS = test set; VS = validation set. [b] At a tolerance of ±0.3 ppm all predictions receive an estimated accuracy greater than 90%.

Table 8. Results of Independent Evaluation

	substructure output						
	no. substructures	≥90%	≥90%	nondominated			
compound	total	no. substructures	% correct	no. substructures	% correct	tau	tau max
1	299	47	100	6	100		
2	1366	40	93	13	77	0.47	0.64
3	47	5	100	1	100		
4	1522	20	95	9	89	0.3	0.48
5	267	168	98	5	60	−0.77	0.77
6	666	58	95	6	50	0	0.8
7	715	435	97	15	60	0.4	0.72
8	37	21	100	5	100		
9	8092	3609	88	324	52	0.38	0.71
10	716	7	57	6	67	0.73	0.73
11	349	39	95	11	91	0.43	0.43
12	154	20	0	8	0		

those variables which are statistical in nature (e.g., $sNOS_t$/sNRC) play a more influential role in training networks at higher tolerances, leading to enhanced discrimination between valid and invalid substructures.

Network performance is evaluated in terms of two measures: (1) how closely the *estimated* prediction accuracies of 90%, 95%, and 99% match *actual* accuracy for each validation set; and (2) how closely recall at 90%, 95%, and 99% accuracy for each validation set matches recall values for the corresponding test sets. At an estimated prediction accuracy of 90%, the observed difference in accuracy between validation and test sets is small; a maximum of 1.6 percentage points at 0.5 ppm. At 90% accuracy, the observed deviations at tolerances 1.0 ppm, 1.5 ppm, and 2.0 ppm are less, and range from 0.1% to 0.9%. (At a tolerance of 0.3 ppm, all substructures retrieved have prediction accuracies above 90%.) At estimated prediction accuracies of 95% and 99%, deviations from test set accuracies are equally small, varying from a low of 0.1 percentage points to a high of 1.6 percentage points. A comparison of test sets with the learning sets reveals comparably small differences ranging from 0.3% to 3.5%.

A comparison of the recall values of validation sets and test sets at each of the three estimated prediction accuracies and five tolerances (Table 7) leads to a similar result, small differences. The difference varies from 0.0 to 0.8 percentage points, except at an estimated accuracy of 99% where somewhat larger differences are observed, 4.0%, 1.8%, 1.4%, and 1.3%, at tolerances of 0.3, 0.5, 1.0, and 2.0 ppm, respectively. Comparable differences, varying from 0.0% to 3.0%, are observed between the learning sets and the test sets.

■ INDEPENDENT PROGRAM EVALUATION

To further evaluate performance, the program was tested using a structurally diverse set of 12 complex natural products[39−50] as "unknowns", none of which are included in the reference library used in this study (Figure 3). Specifically, the substructures derived from these compounds were not used in training the ANN for estimating prediction accuracy. The results are described in Table 8.

The first three columns (substructure output) report information about the substructures retrieved by INFERCNMR with an *estimated* accuracy equal to or greater than 90%. In examples 9, 10 and 12, the *actual* accuracy is less than 90%. Note that Table 8 reports results for the substructures retrieved from each individual compound collected over 21 different tolerances. Table 7, in contrast, reports average results for all substructures produced from the entire validation set at a specific tolerance.

The data in Table 8 further attests to the quality of the measure of estimated prediction accuracy. However, the particular estimated prediction accuracy procedure described herein will have limitations given the limited reference library used in developing the neural network. Neural networks are effective in interpolating, but not as effective in extrapolating. Thus, a network trained with a set of substructures derived from a set of compounds of limited size and diversity will in general be most effective in estimating substructure prediction accuracy for unknowns within the range of substructure diversity in the training set.

Next, turn to the data for the nondominated set, the remaining four columns in Table 8. Consider the second entry (compound 2). Of the total of 1366 substructures retrieved, 40 are reported with an estimated accuracy equal to or greater than 90%. Of the 40, 37 (93%) are correct predictions. The nondominated set of predictions with an estimated accuracy equal to or greater than 90% includes 13 substructures of which 10 (77%) are correct predictions. These results elicit two questions.

First, note that the percentage of correct substructures is less in the nondominated set (77%) than in the original set from which it was derived (93%). (This same observation obtains in the case of all but one compound, compound 10, in Table 8.) The "concentration" of incorrect substructures in the nondominated set is to be expected since a correct substructure can never dominate (eliminate) an incorrect substructure. Only an incorrect substructure can dominate an incorrect substructure and since there are few incorrect substructures relative to correct substructures, the elimination of an incorrect substructure, although possible (and occurs in the case of compound 10), is usually unlikely. Thus, while the number of correct substructures is reduced in arriving at the nondominated set, there is little or no change in the number of incorrect substructures.

Second, the percentage of correct substructures in the nondominated set by itself is not sufficiently informative. A better sense of the discriminating power of the method is required. In the output, substructures of the nondominated list are arranged in decreasing order of estimated accuracy. The *ideal* arrangement would have all correct predictions at the top of the list, and the incorrect predictions at the bottom. If, in practice, this is not obtained, what is a measure of the "goodness" of the mix? (The worst case is where all incorrect substructures have a higher estimated accuracy than the correct substructures.) One applicable measure is Kendall's tau value (τ).[51] Tau is a measure of the degree of concordance between two rankings; in this case between the rankings of estimated accuracy and correctness of the predicted substructures. In this method, all possible combinations of *two* predicted substructures in the set of nondominated substructures which differ both in predicted accuracy and correctness are examined. Each pair is counted either as *concordant* if the substructure with the higher reliability is correct or as discordant if it is not. The difference in the numbers of concordant and discordant pairs is then normalized by the geometric mean of the number of pairs that differ in predicted accuracy (N_x) and the number of pairs that differ in correctness (N_y) (eq 9). Kendall's τ lies

$$\tau = (\text{concordant} - \text{discordant})(N_x N_y)^{-0.5} \tag{9}$$

in the range from −1 to 1. A value of +1 indicates perfect order, that is, all correct substructures appear before the incorrect ones, while the value −1 characterizes the worst case. Due to "ties" (i.e., where the two substructures of a pair have the same correctness or reliability), a perfect order will normally result in a value of less than one in cases of more than two substructures. (A pair of substructures in which both correctness and reliability are the same is not considered in the determination of τ.) To facilitate interpretation in such cases, the maximum value that τ can assume, given the number of correct and incorrect substructures and conserving the ties in predicted accuracy, is also computed. Thus in the case of compound 2, the order of correctness within the nondominated set is reasonable, but not perfect. It is perfect in compounds 10 and 11. In compound 5, all incorrect substructures lie above correct substructures. In compounds 1, 3 and 8, all members of the nondominated set are correct and therefore ranking does not play a role. The same is true for compound 12 where all substructures are incorrect.

A closer look at the results of compound 11, Velloquercetin, serves to amplify the nature of the information in the INFERCNMR output and illustrate the application of that information. The role of INFERCNMR is to enhance the elucidation of complex structures, either in a standalone mode or as an addition to a comprehensive, computer-based structure elucidation system such as SESAMI,[30] by increasing the information pool. INFERCNMR is at its most powerful when the information is used in SESAMI.

The structure of Velloquercetin was elucidated in 1998[49] using a traditional approach based heavily on NMR spectral properties: 1D ^{13}C NMR, 1D ^{1}H NMR, HMQC (one-bond carbon−hydrogen correlations), and HMBC (long-range carbon−hydrogen correlations). No COSY data were provided. Using just these published NMR data (Table 9) as input to SESAMI resulted in a session that was aborted after 30 min and 5000 generated structures (Table 10, row 1). Clearly, the information content of the collective NMR data was insufficient to narrow the plausible structures to a very small number, that is, to a solution useful to the chemist. Can INFERCNMR provide additional substructural information content, which does not duplicate that produced by the spectrum interpretation program (INTERPRET) of SESAMI, to produce a solution that *is useful* to the chemist?

Using the published 1D ^{13}C NMR data for Velloquercetin,[49] INFERCNMR generates 31 substructures with estimated prediction accuracies greater than or equal to 90%. Of the 31 substructures, 11 substructures 1, 2, 3, 7, 12, 16, 20, 24, 28, 30 and 31, in order of decreasing estimated prediction accuracy, are nondominated (Figures 4 and 5). Of the 11 nondominated substructures, only structure 31, the structure with the lowest estimated prediction accuracy of the group, is a false positive. Thus, in this case, INFERCNMR performance is excellent; in the list of substructures, all correct ones lie above the incorrect one (the Kendall's τ value is at a maximum in this case, Table 8).

With the exception of substructures 16 and 30, the substructures are related as superstructure-substructure. These latter relationships are illustrated in Figure 4 by means of a HASSE diagram.[52] For example, substructures 24 and 28 (identical) are richest in structural content. They are the root node of the HASSE tree. Viewing the leftmost branch of the tree, substructures 24 and 28 are superstructures of substructure 2.

Recall, correct chemical structure *and* correct chemical shift assignments are required to designate a substructure prediction as valid. Substructure 1 is structurally identical to substructure 31, yet substructure 31 is incorrect. Thus, the difference must lie in chemical shift assignments. An examination of chemical

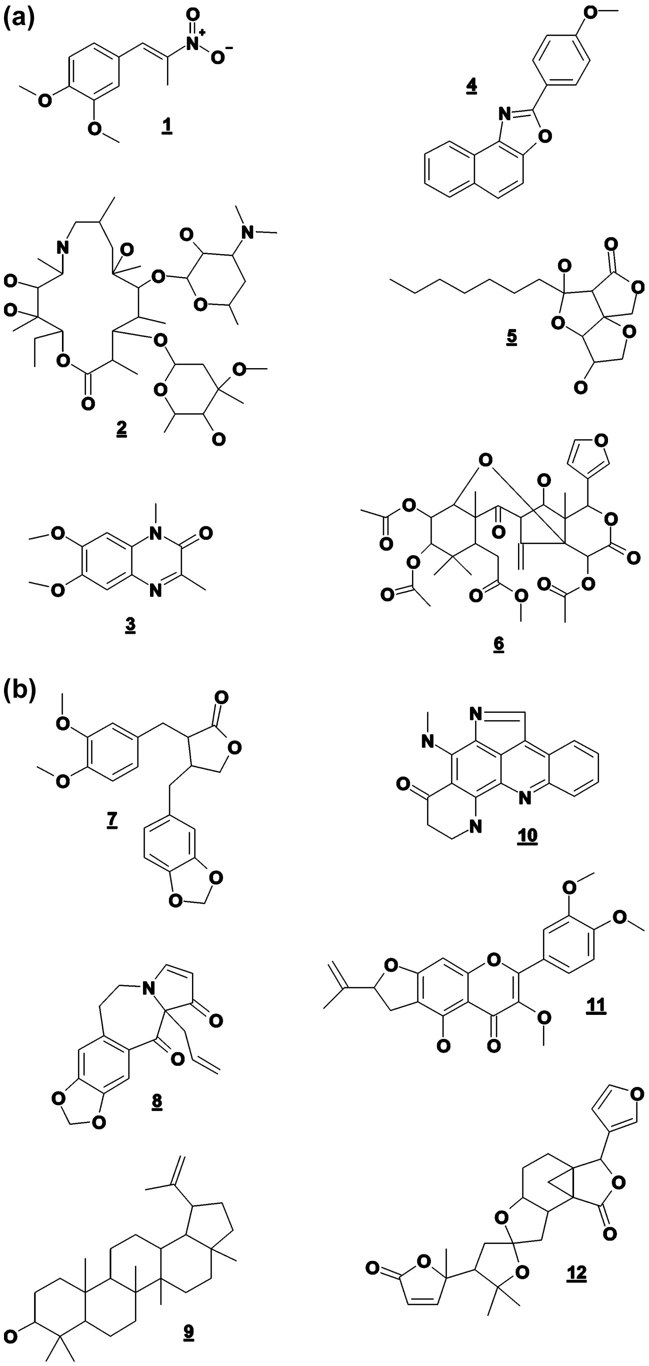

Figure 3. (a) Independent evaluation set. Structures shown are complete. Terminal lines represent methyl groups. (b) Independent evaluation set.

shift assignments of substructure 31 reveals one carbon atom lacks a correct chemical shift assignment. A given carbon atom of a substructure may have multiple assignments, but if the correct assignment is missing, the substructure is incorrect. Regardless of how many incorrect chemical shifts are included, the assignment is *correct* as long as the correct chemical shift assignment is present. It is true that if Velloquercetin were an actual unknown, it would not be known that substructure 31 is incorrect. However, given its lower estimated prediction accuracy and lower information content (Figure 4), it is not likely to be considered for initial study.

Substructures 24 and 28 are likewise identical in chemical structure. Since both are correct in terms of chemical structure *and* chemical shift assignment, the substructure with the higher estimated prediction accuracy (substructure 24) is selected for input to SESAMI. To elaborate on the role of chemical shift assignment, it can be observed that given the parameters used in estimating prediction accuracy (Section Prediction Accuracy), it is conceivable that of two *correct*, structurally identical substructures, the one with the lower estimated prediction accuracy could be richer in information content by virtue of the following. A substructure whose carbon atoms have fewer chemical shift assignments is less ambiguous and in that sense richer in information. All it takes for a substructure to be designated *correct* in terms of chemical shift assignment is for each carbon atom to include the correct chemical shift assignment. It matters not how many chemical shifts are assigned.

In selecting INFERCNMR-generated substructures to use as input to SESAMI, the set of nondominated substructures whose estimated prediction accuracy equals or exceeds that specified by the user provides the starting point. The selection criterion is simple: highest information content at highest available estimated

Table 9. Velloquercetin NMR Spectral Data

δ_C	DEPT	HMQC	HMBC
155.73	S		7.69, 7.73
139.04	S		3.86
179.00	S		
156.65	S		12.89
108.39	S		3.02, 3.38, 6.43, 12.89
166.41	S		3.02, 3.38, 6.48
157.25	S		6.48
106.44	S		6.48, 12.89
123.14	S		7.00
148.91	S		7.00, 7.69
151.45	S		7.79, 7.73
143.33	S		1.78, 3.02, 3.38
89.00	D	6.43 (s, 1H)	
111.33	D	7.69 (d, 1H)	7.73
110.0	D	7.00 (d, 1H)	
122.26	D	7.73 (dd, 1H)	7.69
88.36	D	5.35 (1H)	1.78, 3.02, 4.95, 5.10
112.92	T	4.95 (s, 1H)	1.78
	T	5.10 (s, 1H)	
30.67	T	3.02 (dd, 1H)	
	T	3.38 (dd, 1H)	
17.13	Q	1.78 (s, 3H)	4.95, 5.10
60.36	Q	3.86 (s, 3H)	
56.18	Q	3.96 (s, 3H)	
56.19	Q	3.97 (s, 3H)	

Table 10. SESAMI Results for Velloquercetin

	INFERCNMR substructure	assigned	execution time (min)	no. structures
1			30.0	5000[a]
2	2	yes	5.5	1784
3	16	yes	10.0	5000[a]
4	24	yes	1.5	859
5	30	yes	12.5	3047
6	24, 30	no	611.0	4
7	24, 30	yes	1.5	2
8	2, 30	yes	1.5	14

[a] Aborted after the generation of 5000 structures.

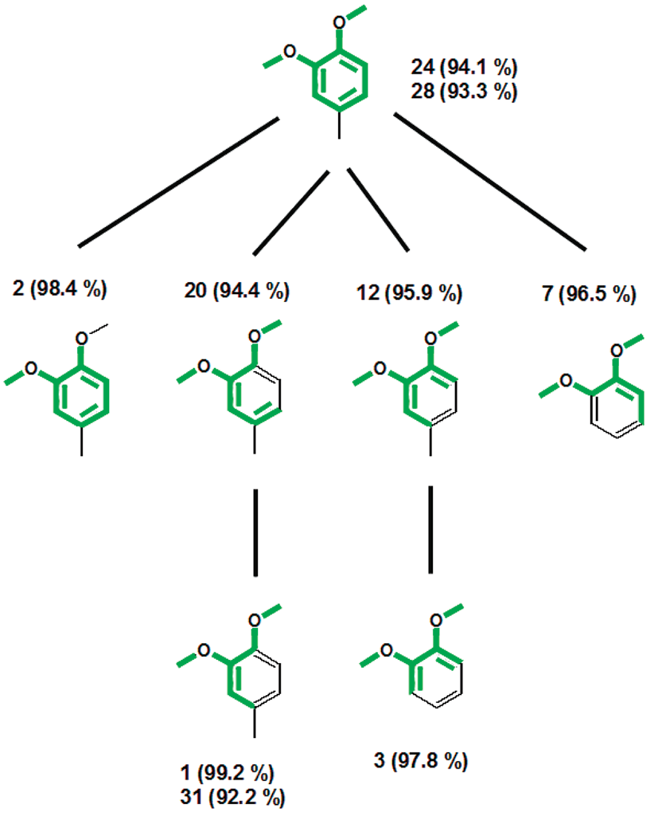

Figure 4. Hasse diagram of structurally related INFERCNMR substructures.

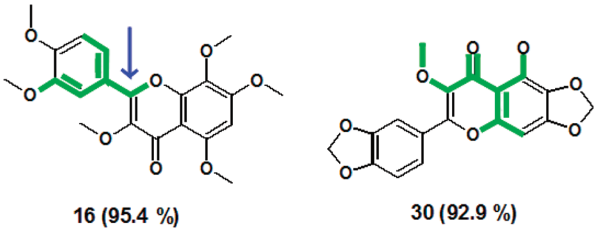

Figure 5. INFERCNMR Substructures 16 and 30 for Velloquercetin.

prediction accuracy. That approach suggests the application of substructures 16, 24 (substructure 28 has a lower estimated prediction accuracy) and 30 since they each contain some unique structural information. (Note the selection of substructure 24 over substructure 2 (a user decision) even though substructure 2 has a significantly higher estimated prediction accuracy, a difference of over 4%. Substructure 24 still has a high estimated prediction accuracy (94.1%) and is richer in information content—it has one more methoxy methyl group than substructure 2.)

The structure-reducing impact of adding a single *assigned* substructure, 16, 24, or 30, to the original SESAMI session (Table 10, row 1) is shown in Table 10, rows 3, 4 and 5, respectively. Substructures 24 and 30 separately significantly reduce the number of candidate structures generated, from greater than 5000 to 859 and 3047, respectively. The outcome does suggest that substructure 24 is the richer in information that does not duplicate information from the other spectral sources used by SESAMI. However, at 859 generated structures, the collective information pool is still not sufficiently rich enough to provide a useful result for the chemist. The number of candidate structures generated by adding substructure 16 to the original input still exceeds 5000, not a useful result. Thus, substructure 16 adds the least useful information in elucidating the structure. The result is not surprising. An examination of substructures 16 and 24, reveals a complete aromatic ring in 24, but incomplete in 16. Substructure 16 does include a carbon atom (see arrow in substructure 16) not present in substructure 24, but that appears to be insufficient to offset the absence of information in the incomplete aromatic ring.

Table 10 also reveals the significant impact of the additional information content of substructure 24 relative to substructure 2, one additional methyl group. The former (row 4) produces substantially fewer structures than the latter (row 2); 859 and 1784, respectively. Lower information content also increases execution time.

The addition of *assigned* substructures 24 *and* 30 to the SESAMI input (Table 10, row 7) leads to a dramatic reduction in the reduction of the structures generated and a very useful outcome for the chemist. Two structures are produced. It was stated earlier that adding chemical shifts assignments to the carbon atoms of the substructure, even ambiguous assignments, that is, more than one chemical shift to one or more carbon atoms, increases information content and enhances the performance of SESAMI. This is consistent with the result of a SESAMI session using both substructures 24 and 30, but without chemical shift assignments (Table 10, row 6). Four structures rather than two are produced and the execution time required is increased by more than 2 orders of magnitude.

A session utilizing *assigned* substructures 2 and 30 as input was run to compare the result to the session using *assigned* substructures 24 and 30 to determine if the lower information content of substructure 2 relative to substructure 24 would impact the outcome. The combination of 2 and 30 did substantially reduce the numbers of generated structures compared to using each individual substructure separately (Table 10, row 8), but not as effectively as the combination of 24 and 30 (Table 10, row 7).

Clearly, in the case of the Velloquercetin problem, the information content produced by INFERCNMR complements rather than duplicates the information content derived from the current INTERPRET program in SESAMI. Therein lies the power of INFERCNMR.

The two structures of Velloquercetin proposed by SESAMI in the session described in Table 10, row 7 are shown in Figure 6.

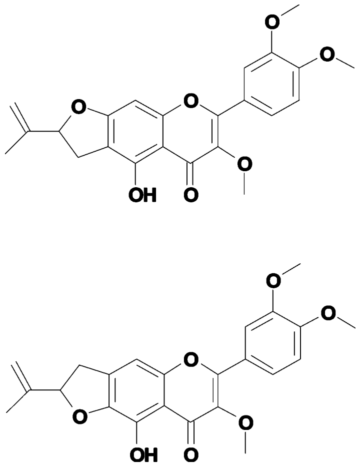

Figure 6. SESAMI output for velloquercetin.

The structure at the top is the same as that assigned by the authors of the original paper.[49] Interestingly, it would appear that the collective data reported in that paper does not rule out the second structure generated by SESAMI. In reality, a structure assignment should not be considered final until it can be demonstrated that there is no other structure equally compatible with the available information.

■ CURRENT RELATED WORK

The application of ^{13}C NMR in the elucidation of the structure of complex natural products has been extensively studied by a Emerenciano and co-workers.[53,54] They have described an interpretive library search system which predicts the presence of substructures in an unknown.[55] The program is based on an earlier algorithm[24] with important modifications to address the problem of combinatorial explosion which can occur in the case of large molecules. Prediction reliability is based on the size of the substructure. Chemical shifts are assigned to each carbon atom of the substructure drawing on the work of Robien.[56]

■ CONCLUSIONS

A ^{13}C NMR interpretive library search as a tool in the elucidation of the structure of complex compounds is of value only if its predicted substructures are reliable. Therefore, in designing INFERCNMR, great emphasis was placed on the ability of the program to estimate the accuracy of a predicted substructure. The user can then make an informed decision whether to consider it as a required substructure. The results reported in Table 7 indicate an ANN model of substantial predictive capabilities, providing sufficient discrimination between valid and invalid retrieved substructures to meet the needs of an interpretive library search system which can function either as a standalone tool or as a component of a comprehensive structure elucidation system. The development of INFERCNMR was facilitated by a new approach to the encoding of tautomeric structural units and by the application of a target-weighted neural network.

In application to real-world, complex unknown structures, where high prediction accuracy is important, an estimated accuracy of 99% may be preferred by the user. Running INFERCNMR at low tolerance leads to high recall, but a considerably smaller number of retrieved substructures. At higher tolerances, more substructures are retrieved, but recall will be less. In solving complex structure elucidation problems, even one correct substructure of six or more carbon atoms can provide substantial information content provided it does not duplicate information from other sources. In our experience, an examination of the nondominated list of substructures produced at an accuracy of 90% and above can be an informative exercise. Recall that INFERCNMR, like any reference library-based tool, can only retrieve substructures which are present in the reference compounds of the library.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mem@asu.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The financial support of this research by the National Institutes of Health and Ash Stevens, Inc. is gratefully acknowledged. The coordinates for the structures were generated using the CACTVS browser.[57] The ongoing technical assistance of Frank Davis, E-Jay Christopher, and Mary Rushton in hardware and software support was indispensable to the completion of this research.

■ REFERENCES

(1) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley: New York, 1986.

(2) Munk, M. E. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997−1009.

(3) Munk, M. E.; Madison, M. S. Structure Determination via Computer-Based Spectrum Interpretation. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 2785−2811.

(4) Jaspars, M. Computer-Assisted Structure Elucidation of natural Products Using Two-dimensional NMR Spectroscopy. *Nat. Prod. Rep.* **1999**, *16*, 241−248.

(5) Steinbeck, C. Recent Developments in Automated Structure Elucidation of Natural Products. *Nat. Prod. Rep.* **2004**, *21*, 512−518.

(6) Elyashberg, M. E.; Williams, A. J.; Martin, G. E. *Prog. Nucl. Mag. Res. Sp.* **2004**, *53*, 1−104.

(7) Luinge, H. J.; Kleywegt, G. J.; Van't Klooster, H. A.; Van Der Mas, J. H. Artificial Intelligence Used for thr Interpretation of Combined Spectral Data′ 3. Automated Generation of Interpretation Rules for Infrared Spectral Data. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 95−99.

(8) Woodruff, H. B.; Munk, M. E. A Computerized Infrared Spectral Interpreter as a Tool in Structure Elucidation of Natural Products. *J. Org. Chem.* **1977**, *42*, 1761−1767.

(9) Trulson, M. O.; Munk, M. E. Table-Driven Procedure for Infrared Spectrum Interpretation. *Anal. Chem.* **1983**, *55*, 2137−2142.

(10) Zupan, J.; Munk, M. E. Hierarchical Tree Based Storage, Retrieval and Interpretation of Infrared Spectra. *Anal. Chem.* **1985**, *57*, 1609−1616.

(11) Jurs, P. C. *Computer Software Applications in Chemistry*; Wiley Interscience: New York, 1996.

(12) Varmuza, K. *Pattern Recognition in Chemistry*; Wiley: New York, 1980.

(13) Meisel, W. S.; Jolley, M.; Heller, S. R.; Milne, G. W. A. The Role of Pattern Recognition in the Computer-Aided Classification of Mass Spectra. *Anal. Chim. Acta* **1979**, *112*, 407−416.

(14) Woodruff, H. B.; Snelling, C. R.; Shelley, C. A.; Munk, M. E. Computer-assisted Interpretation of C-13 Nuclear Magnetic Resonance Spectra Applied to Structure Elucidation of Natural Products. *Anal. Chem.* **1977**, *49*, 2075−2080.

(15) Munk, M. E.; Madison, M. S.; Robb, E. W. The Neural Network as a Tool for Multispectral Interpretation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−238.

(16) Klawun, C.; Wilkins, C. L. Joint Neural Network Interpretation of Infrared and Mass Spectra. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.

(17) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, England, 1987.

(18) Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. Computer-Aided Interpretation of Mass Spectra. III. A Self-Training Interpretive and Retrieval System. *J. Am. Chem. Soc.* **1973**, *95*, 4185−4210.

(19) Schwarzenbach, R.; Meilli, J.; Könitzer, H.; Clerc, J.-T. A Computer System for the Identification of Organic Compounds from C-13 NMR Data. *Org. Magn. Reson.* **1976**, *8*, 11−16.

(20) Farkas, M.; Bendl, J.; Welti, D. H.; Pretsch, E.; Dütsch, S.; Portmann, P.; Zürcher, M.; Clerc, J.-T. Similarity Search for a Proton-NMR Spectroscopic Database. *Anal. Chim. Acta* **1988**, *206*, 173−187.

(21) Naegeli, P. R.; Clerc, J.-T. Computer System for Structural Identification of Organic Compounds from Spectroscopic Data. *Anal. Chem.* **1974**, *45*, 739A−744A.

(22) Zupan, J.; Penca, M.; Hadzi, D.; Marsel, J. Combined Retrieval System for Infrared, Mass, and C-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1977**, *49*, 2141−2146.

(23) Bremser, W.; Klier, M.; Meyer, E. Mutual Assignment of Subspectra and Substructures: A Way to Structure Elucidation by C-13 NMR-Spectroscopy. *Org. Magn. Reson.* **1975**, *7*, 97−106.

(24) Shelley, C. A.; Munk, M. E. Computer Prediction of Substructures from C-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1982**, *54*, 516−521.

(25) Bremser, W. Importance of Multiplicities and Substructures for Evaluation of Relevant Spectral Similarities for Computer-Aided Interpretation of C-13 NMR-Spectra. *Fresenius Z. Anal. Chem.* **1977**, *286*, 1−13.

(26) Chen, L.; Robien, W. Application of the Maximum Common Substructure Algorithm to Automatic Interpretation of ^{13}C-NMR Spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 34−941.

(27) Bremser, W. Structure Elucidation and Artificial Intelligence. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 247−260.

(28) Will, M.; Fachinger, W.; Richert, J. R. Fully Automated Structure Elucidation-A Spectroscopist's Dream Come True. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221.

(29) Plainchont, B.; Nuzillard, J.-M.; Rodrigues, G. V.; Ferreira, M. J. P.; Scotti, M. T.; Emerenciano, V. P. New Improvements in Automatic Structure Elucidation Using the LSD (Logic for Structure Determination) and the SISTEMAT Expert Systems. *Nat. Prod. Comm.* **2010**, *5*, 763−770.

(30) Schulz, K.-P.; Korytko, A.; Munk, M. E. Applications of a HOUDINI-Based Structure Elucidation System. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1447−1456.

(31) Munk, M. E.; Velu, V. K.; Madison, M. S.; Robb, E. W.; Badertscher, M.; Christie, B. D.; Razinger, M. Chemical Information Processing in Structure Elucidation. In *Recent Advances in Chemical Information II*; Collier, H, Ed.; Royal Society of Chemistry: Cambridge, U.K.,1993, pp 247−263.

(32) Munk, M. E.; Christie, B. D.; Velu, V. K. The Role of NMR Spectra in Computer-Enhanced Structure Elucidation. In *Computer-Enhanced Analytical Spectroscopy*, Vol. 3; Jurs, P., Ed.; Plenum: New York, 1992, Chapter 5.

(33) Deb, K. *Multi-Objective Optimization using Evolutionary Algorithms*; Wiley: Chichester, U.K., 2001; p 28.

(34) Chartrand, G.; Oellermann, O. R. *Applied and Algorithmic Graph Theory*; McGraw Hill: New York, 1993; pp 25−31.

(35) Nater, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985; pp 361−367.

(36) Zupan, J.; Gasteiger, J. *Neural Networks for Chemist: An Introduction*; VCH Publishers: Weinheim, Germany, 1993.

(37) Papadimitriou, C. H.; Steiglitz, K. *Combinatorial Optimization: Algorithms and Complexity*; Dover: Mineola, New York, 1998; pp 248−255.

(38) Penchev, P. N.; Andreev, G. N.; Varmuza, K. Automatic Classification of Infrared Spectra Using a Set of Improved Expert-Based Features. *Anal. Chim. Acta* **1999**, *388*, 145−159.

(39) Breitmeier, E. *Structure Elucidation by NMR in Organic Chemistry*; Wiley: Chichester, U.K., 2002; p 28.

(40) Bayod-Jasanada, M.; Carbajo, R. J.; Lopez-Ortiz, F. Assignment of the ^{1}H and ^{13}C Spectra of 9-Deoxo-9a-aza-9a-homoerythromycin A, 9-Deoxo-9a-aza-9a-homoerythromycin A 11,12-hydrogenborate and Azithromycin 11,12-hydrogenborate. *Magn. Reson. Chem.* **1998**, *36*, 217−225.

(41) Rampont-Placidi, V.; Cossec, B.; Brondeau, M.-T.; Mutzenhardt, P. Complete ^{13}C and ^{1}H Spectral Assignments of Certain Substituted Quinoxalinones. *Magn. Reson. Chem.* **1998**, *36*, 300−302.

(42) Marquez, A.; Saitz, C.; Canete, A.; Rodriguez, H.; Jullian, C. Complete Assignment of the ^{1}H and ^{13}C NMR Spectra of 2-Phenyl-3H-naphtho[2,1b][1,4]oxazin-3-one, 2-p-Methoxyphenylnaphtho[1,2-d]-oxazole and 2-Phenylnaphtho[1,2-d]oxazole. Concerted Use of One- and Two-Dimensional NMR Techniques. *Magn. Reson. Chem.* **1998**, *36*, 449−453.

(43) Smith, M. J.; Mazzola, E. P.; Sims, J. J.; Midland, S. L.; Keen, N. T.; Burton, V.; Stayton, M. M. The Syringolides: Bacterial C-Glycosyl Lipids That Trigger Plant Disease Resistance. *Tetrahedron Lett.* **1993**, *34*, 223−226.

(44) Ismail, I. S.; Ito, H.; Hatano, T.; Taniguchi, S.; Yoshida, T. Modified limonoids from the leaves of *Sandoricum koetjape*. *Phytochemistry* **2003**, *64*, 1345−1349.

(45) Chang, C.-C.; Lien, Y.-C.; Chen Liu, K. C. S.; Lee, S. S. Lignans from *Phyllanthus urinaria*. *Phytochemistry* **2003**, *63*, 825−833.

(46) Wasserman, H. H. Personal Communication, Department of Chemistry, Yale University, New Haven CT 06511.

(47) Wenkert, E.; Baddeley, G. V.; Burfitt, I. R.; Moreno, L. Carbon-13 Nuclear Magnetic Resonance Spectroscopy of Naturally-occurring Substances; LVII. Triterpenes Related to Lupane and Hopane. *Org. Mag. Res.* **1978**, *11*, 337−343.

(48) Inman, W. D.; O'Neill-Johnson, M.; Crews, P. Novel Marine Sponge Alkaloids. Plakinidine A and B, Anthelmintic Active Alkaloids from a Plakortis Sponge. *J. Am. Chem. Soc.* **1990**, *112*, 1−4.

(49) Branco, A.; Braz-Filho, R.; Kaiser, C. R.; Pinto, A. C. Two Monoisoprenylated Flavonoids from Vellozia Graminifolia. *Phytochemistry* **1998**, *47*, 471−474.

(50) Mulholland, D. A.; MacFarland, K.; Randrianarivelojosia, M.; Rabarison, H. Cedkathryns A and B, pentanortriterponoids from *Cedrelopsis gracilis* (Ptaeroxylaceae). *Phytochemistry* **2004**, *65*, 2929−2934.

(51) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing*; Cambridge University Press: Cambridge, U.K., 1992; pp 642−645.

(52) Brightwell, G.; West, D. B. Partially Ordered Sets. In *Handbook of Discrete and Combinatorial Mathematics*;, Rosen, K. H., Ed.; CRC Press: Boca Raton, FL, 2000; pp 717−752.

(53) Ferreira, M. J. P.; Oliveira, F. C.; Alvarenga, S. A. V.; Macari, P. A. T.; Rodrigues, G. V.; Emerenciano, V. P. Automatic Identification of C-13 NMR of Substituent Groups Bonded in Natural Products. *Comput. Chem. (Oxford, U. K.)* **2002**, *26*, 601−632.

(54) Ferreira, M. J. P.; Brant, A. J. C.; Rodrigues, G. V.; Emerenciano, V. P. Automatic Identification of Terpenoid Skeletons Through C-13 Nuclear Magnetic Resonance Data Disfunctionalization. *Anal. Chim. Acta* **2001**, *429*, 151−170.

(55) Fromanteau, D. l. G.; Gastmans, J. P.; Vestri, S. A.; Emerenciano, J. P.; Borges, J. H. G. A Constraints Generator in Structural Determination by Microcomputer. *Comput. Chem. (Oxford, U. K.)* **1993**, *17*, 369−378.

(56) Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for the Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103−108.

(57) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki., S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 109−116.