# Using Predicted $^{13}C$ NMR Spectra with Open Resources for Structure Dereplication

Dimitris Argyropoulos, Sergey Golotvin, Rostislav Pol, Joe DiMartino, Arvin Moser and Brent Pautler

*Advanced Chemistry Development, Inc. (ACD/Labs), 8 King Street East, Toronto, ON, M5C 1B5, Canada*

## Introduction

Over the past two decades, market pressure has increased demands for development of New Molecular Entities (NME's)[1]. In response, the pharmaceutical industry has made efforts to accelerate this development by implementing more efficient, higher volume techniques into development procedures. These include, High-Throughput Screening, Parallel Synthesis, and Absorption, Distribution, Metabolism, and Excretion Toxicology (ADMET) Predictions. Natural product discovery programs reveal chemical diversity that can complement high-throughput screening efforts. However, this is only worthwhile if the active components in natural product mixtures can be reliably separated and quickly identified. The practice of screening active compounds early in the development process for recognizing and eliminating known compounds is called dereplication. This enables scientists to focus on testing truly 'unknown' compounds.

There are two conditions that must be fulfilled for efficient dereplication:

1. One must be able to easily identify characteristic spectral 'fingerprints' of 'unknown' compounds.
2. One must have access to databases containing spectra of known structures.

NMR and MS spectra are typically used for dereplication. High resolution MS is the simplest and fastest to record, but it lacks the structural information that NMR provides. $^{1}H$ NMR is a fast and simple method that gives structural information. However, $^{1}H$ NMR spectra is not a reliable fingerprint because of its limited resolution and the fact that measured spectra can be affected by factors like pH, concentration, and solvent effects. The $^{13}C$ NMR spectrum of a compound, on the other hand, can be considered an effective fingerprint since it is virtually unaffected by the aforementioned conditions. It is also largely magnetic field independent, since there are no couplings that could cause variations in stronger or weaker fields. As a result, it is very easy and accurate to predict.

To satisfy the second condition, one can consider using databases of real spectra or predicted spectra. Databases of real spectra usually contain a limited number of structures, and their spectra may not be ideal. On the other hand, there are several "open" databases with millions of chemical structures that could be used to predict $^{13}C$ spectra. Two examples of this are PubChem[2] and ChemSpider[3]. Predicted spectra benefit from being magnetic field independent, can be adjusted for solvents, and can be very accurate if the correct algorithms are used [4].

Here we propose an efficient dereplication strategy, which treats the experimental $^{13}C$ NMR spectra of an 'unknown' as a fingerprint. The 'unknown' is identified by finding a match for its fingerprint in a database of predicted $^{13}C$ NMR spectra of known compounds. We explore the possibilities and limitations of using predicted $^{13}C$ spectra for structures from open databases, describe the workflow, and critically evaluate the usefulness of the technique.

### Open-Source Structure Databases

PubChem

- Maintained by the National Center for Biotechnology Information (NCBI)
- Part of the National Library of Medicine (NLM) and the National Institutes of Health (NIH)
- Contains ~95 million structures (March 2018)
- All stored structures can be downloaded through ftp

ChemSpider
Search and Share Chemistry

- Maintained by the Royal Society of Chemistry (RSC)
- Contains ~64 million structures (March 2018)
- Structures can be browsed individually but not batch-downloaded without an agreement

## References

1. Earm, K., Earm, Y. E., Integrative Medicine Research, **2014**, 3, 211-216.
2. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. **2016** Jan 4; 44(D1):D1202-13. Epub **2015** Sep 22 [PubMed PMID: 26400175] doi: 10.1093/nar/gkv951.
3. Pence, H. E., Williams, A., J. Chem. Educ., **2010**, 87, 1123-1124.
4. Data presented by Burkhard Kirste, FU Berlin, 38th FGNMR Meeting, Sept. **2016**, Dusseldorf
5. ACD/Labs, "ACD/NMR Predictors," **2017**. [Online]. Available: www.acdlabs.com/nmrpredictors
6. G.A. Dziwornu, N.R. Toorabally, M.G. Bhowon, S. Jhaumeer-Laulloo, S.N. Sunassee, A. Moser, D. Argyropoulos, poster presented at 58th ENC, Pacific Grove, **2017**.
7. Jingjing Wu, Paula Lorenzo, Siying Zhong, Muhammad Ali, Craig P. Butts, Eddie L. Myers & Varinder K. Aggarwal, Nature, **2017** *547*, 436–440.

## Method

In order to evaluate the proposed dereplication strategy, one must consider the quality of the predicted $^{13}C$ spectra. To ensure that the spectra of the PubChem and ChemSpider structures are predicted as accurately as possible, we used ACD/Labs NMR Predictors – the industry standard for NMR prediction [4][5].

The following parameters must be defined to ensure that spectral matches are not falsely excluded from the search results:

1. The mean maximum difference between experimental and theoretical chemical shifts. This is usually set as ≤ 2 ppm.
2. The number of known peaks in the experimental spectrum that are not visible due to a low signal to noise ratio (number of missing peaks). This value can be set to 0 or 1 for good quality spectra or higher otherwise.
3. The number of peaks resulting from impurities (number of extra peaks). Can be low if only the appropriate peaks are picked.
4. Optional – one can filter by molecular formula to expedite the search time. For example, this can reduce a ten minute search to 30 seconds.

## Experimental and Results

To illustrate the benefits and limitations of this dereplication strategy, we show 4 examples of the results that may be obtained when attempting to identify an unknown substance.

### Case 1 | Identification of a Known Compound

A natural product was isolated from the leaves of *Ochna Mauritiana*[6], The $^{13}C$ spectrum is shown in figure 1(a). Using our dereplication strategy, this compound was identified as 2',3'-Dihydroochnaflavone (PubChem entry 10437291) in 30 sec. This match was confirmed within a few minutes by examining additional spectra.

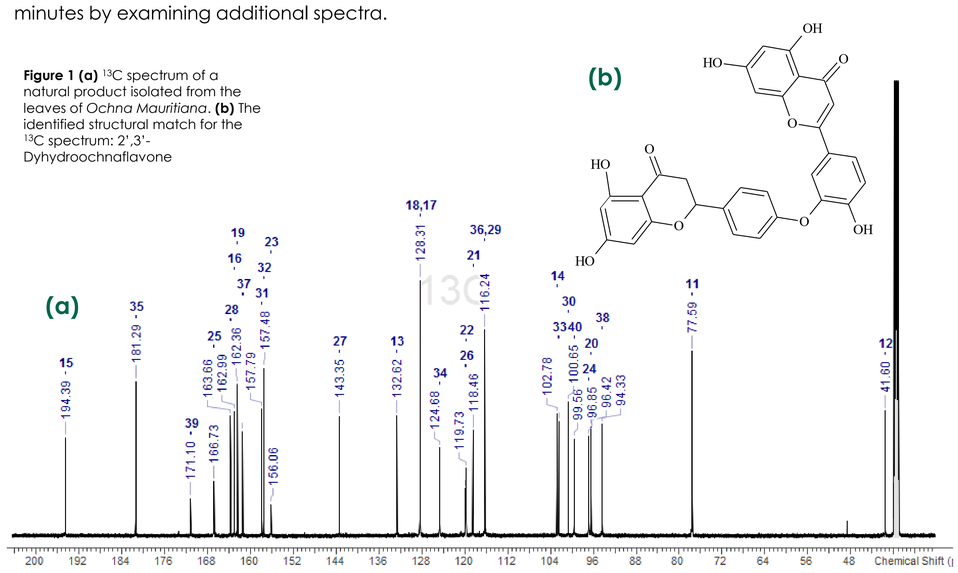


**Figure 1 (a)** $^{13}C$ spectrum of a natural product isolated from the leaves of *Ochna Mauritiana*. **(b)** The identified structural match for the $^{13}C$ spectrum: 2',3'-Dyhydroochnaflavone

### Case 2 | Identification of an Unknown Compound

Another natural product was isolated from the leaves of *Ochna Mauritiana*[6], seen below in figure 2(a). A search did not produce any structures from PubChem or ChemSpider with a similar predicted $^{13}C$ spectrum. As a result, we increased the tolerances for additional peaks in both the experimental and predicted spectra, but this only produced structures similar to that shown in Figure 1(b) with very low match factors (high mean chemical shift deviations).

Since the correct structure could not be found in either the PubChem nor ChemSpider databases, a full structure elucidation was performed. This revealed the correct structure, shown in Figure 2(b). In this example a minor difference in the structure (an additional methyl group) was sufficient to differentiate it from other structures in the two databases and provide evidence of the novelty of the structure.

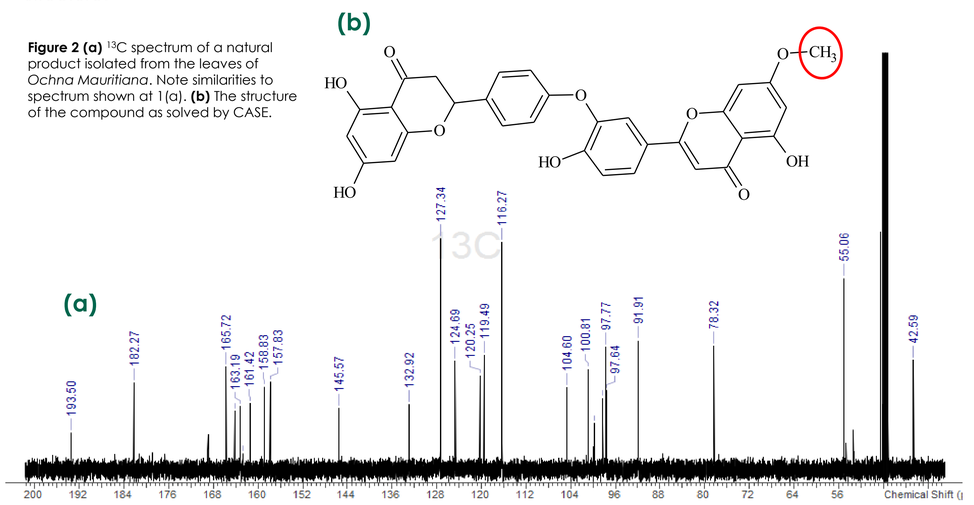


**Figure 2 (a)** $^{13}C$ spectrum of a natural product isolated from the leaves of *Ochna Mauritiana*. Note similarities to spectrum shown at 1(a). **(b)** The structure of the compound as solved by CASE.

### Case 3 | Identification of a Well-Known Compound

In this example, the dereplication strategy is used to identify a very well-known compound, retrorsine. The experimental $^{13}C$ spectrum used (not shown) was well resolved and free of impurity peaks, so searching the known structures database with minimal tolerances is expected to produce a single (or very few) matching structures. However the results produced 40 isomers of the correct structure, with differences in stereochemistry and explicit hydrogen atoms displayed. The first 15 results are shown in Figure 3.

For the purpose of dereplication these results are sufficient; the compound was identified as already known. However, to discern the stereochemistry of the structure, further analysis must be conducted.

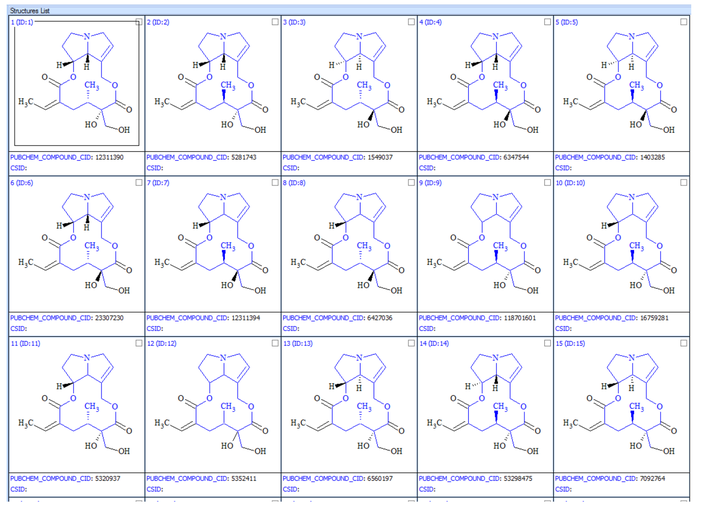


**Figure 3** The first 15 of 40 isomers of the matching structure selected by our dereplication method, the first one is actually retrorsine

### Case 4 | Identification of a Famously Misidentified Compound

Baulamycin-A is famously known for having its 3D structure and stereochemistry misidentified multiple times. Only recently, Butts *et al* (2017) identified its true structure using a collection of spectroscopic, DFT, and synthetic methods [7].

Searching the predicted spectra for PubChem structures produces two results: IDs 100951617 and 74223134. Neither is named Baulamycin. A search of the PubChem database for Baulamycin-A results in the entry shown in Figure 4(a), which is an incorrect structure. The correct structure has a 3-methyl butyl group rather than a 2-methyl butyl as shown in PubChem. Interestingly, the correct structure with the right name can be found in ChemSpider under ID 32674678.

Even if one is able to identify a structure as known, some discretion is required since the dereplication method is not immune to mislabeled structures in the open source databases. As a result, one must critically consider the structural matches, particularly when multiple structural results are given for a single spectrum.

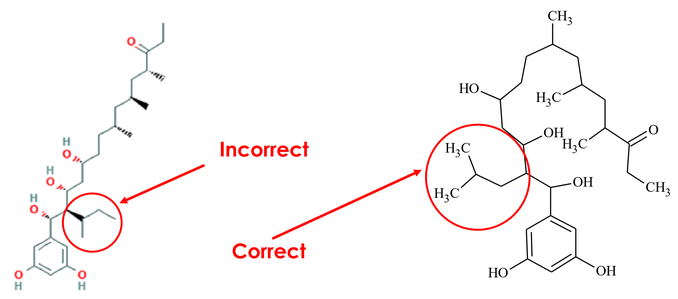


**Figure 4(a)** The incorrect structure shown in PubChem for Baulamycin-A with a 2-methyl butyl group. **(b)** The correct Baulamycin-A structure with a 3-methyl butyl group.

## Discussion and Conclusions

This dereplication strategy using $^{13}C$ spectra is a very powerful method since it is able to significantly reduce the time spent determining if an 'unknown' has previously been identified. One can see that databases with well predicted $^{13}C$ spectra are a very strong resource for dereplication, and that PubChem and ChemSpider are invaluable sources of reported structures. However, there is always a possibility for error, so some caution is needed when interpreting the search results.