The Impact of Artificial Intelligence on Natural Products Research: An Assessment of Small Molecule Accurate Recognition Technology (SMART)

Julius Adam V. Lopez[1,2,*]
[1] *School of Innovation and Sustainability, De La Salle University Laguna Campus*
[2] *Department of Chemistry, De La Salle University Manila*
**Corresponding Author: julius.adam.lopez@dlsu.edu.ph*

Abstract: Natural products are a promising source of new drugs and drug leads. However, determining the chemical structure of a natural product is laborious and time-consuming via traditional methods. The emergence of AI-based structure elucidation tools is expected to change this landscape by streamlining the dereplication of known compounds and aiding in the structure elucidation of new and novel compounds. The objective of this study is to assess the Small Molecule Accurate Recognition Technology or SMART tool. It utilizes a combination of deep Convolutional Neural Networks (CNNs) and a training set of heteronuclear single quantum coherence (HSQC) NMR data to automatically identify the structure of a compound. To test the software, 11 datasets were used consisting of HSQC data from published and unpublished compounds as well as a random and an outlier dataset. Match results from SMART were gauged as exact, close, or far from the actual compound structure. Results revealed a total of 9 out of 11 had exact (3) or close (6) structural matches which is viewed as a great advantage. Even if the match was not exact, one is still led to a closely related structural family which is key for structure elucidation. More importantly, the SMART analysis only took seconds. The same step could take days/hours/weeks if done via manual interpretation of the NMR data. Additionally, the program also recommends other useful external databases that may aid the user such as GNPS, NPATLAS, and MIBiG. As shown with the use of the SMART tool in this study, the analysis was almost instantaneous and seamless which led to a more efficient dereplication process. Points for improvement include better differentiation for small molecules, and recognition for compounds with repeating units and symmetry. Nevertheless, natural products research has received a huge boost with the emergence of these AI-based structure elucidation tools.

Key Words: natural products chemistry; structure elucidation; nuclear magnetic resonance; dereplication; artificial intelligence

1. INTRODUCTION

Natural products are chemical compounds or substances produced naturally by living organisms, including plants, animals, and microorganisms (Mali, 2023). Natural products research has made constructive contributions to drug discovery, nutrition, agriculture, and other disciplines (Mounier et al., 2022). In terms of approved drugs, natural products have been shown to be a rich source for therapeutics against various diseases, with approximately 60% of drugs in the market (Jena et al., 2019). While natural products offer structural diversity and potential for novel lead compounds, challenges such as access to biological resources and technical barriers exist (Conrado et al., 2024).

In particular, the process of collecting a compound from nature, isolating the active ingredient, and determining the chemical structure is time-consuming (Cooper and Nicola, 2014). Moreover, the increasing chance of re-isolating known compounds has slowed down drug discovery from natural products, emphasizing the need for automated dereplication processes using computational resources (Mohamed et al., 2016). Dereplication is a crucial process in natural product screening that aims to quickly identify known compounds, thereby streamlining the discovery of novel or new compounds (Ito and Masubuchi, 2014). To ameliorate the dereplication and structure determination processes, computer-assisted structure elucidation strategies have been developed. These methods automatically propose a list of possible chemical structures in samples by utilizing chromatographic and spectroscopic techniques such as mass spectrometry and nuclear magnetic resonance (NMR) (Su et al., 2017).

More recently, new technologies and methods employing artificial intelligence (AI) have enhanced the screening of natural products, improving efficiency and precision (Cao, 2016). Advanced machine learning and AI algorithms have simplified the search for novel natural products, analyzing their chemical structure and predicting biological function (Manochkumar and Ramamoorthy, 2024). Examples of these AI-ready tools are ACD/Structure Elucidator (Elyashberg and Williams, 2021), DP4-AI (Howarth et al., 2020), IMPRESSION (Gerrard et al., 2020), ANN-PRA method and quantum NMR calculations (Marcarino et al., 2020), and SMART 2.0 (Reher et al, 2020), which is the main focus of this paper.

SMART or Small Molecule Accurate Recognition Technology (Zhang et al., 2017) utilizes a combination of deep Convolutional Neural Networks (CNNs) and a training set of heteronuclear single quantum coherence (HSQC) NMR data to automatically identify the structure of a sample. It provides rapid dereplication and categorize into molecular structural classes. Through this study, the author aims to test the latest iteration of the program, SMART 2.0, using organic compounds that the author discovered and have completed structure elucidation using traditional approaches. Based on the results, insights regarding the advantages and disadvantages of the tool will be provided.

2. METHODOLOGY

The required HSQC NMR data were retrieved from the publications reporting the following compounds: 2-methylthio-*N*^{7}-methyl-*cis*-zeatin (Lopez et al., 2021), columbamides D and E (Lopez et al, 2017), nocardamin glucuronide and bisucaberin (Lopez et al., 2019), wewakazole (Nogle et al., 2003), wewakazole B (Lopez et al., 2016), *N*-acetyl-ß-oxotryptamine and *N*-acetyl-α-hydroxy-ß-oxotryptamine (Lopez et al., 2021). To generate the suitable HSQC NMR data, ^{1}H and ^{13}C were tabulated in Microsoft® Excel for Mac version 16.16.27 according to the suggested format. The data for unpublished novel and new compounds were also included. In addition, "dummy data" and "outrageous data" were also created to represent random NMR data and illogical NMR data, respectively. The data were then saved as CSV UTF-8 file format (.csv) for each compound. One by one, these files were subsequently uploaded to the website, http://smart.ucsd.edu/classic, and subjected to SMART analysis.

Screenshots of the results were acquired. Next, the results were summarized in a table using the same Microsoft® Excel software. To rank the results obtained, the following words were used to judge the compound matching: exact, close, far, N/A. Please note that only the top result considered. "Exact" means the software predicted exactly the same structure as reported in literature. "Close" means the structure generated is closely related to or is within the same compound family. "Far" means the structure shown by SMART is different or not structurally related to the actual compound. And "N/A" represents not applicable and means no

matching structure/s were presented or there was an error in the analysis brought about by inadequate data.

3. RESULTS AND DISCUSSION

The classical structure elucidation of compounds is an arduous process due to the high level of knowledge required for the analysis and time needed to complete the task. The undertaking entails gathering the ^{1}H and ^{13}C NMR data, which are one dimensional (1D) and provide initial impressions of the functional groups in the compound, followed by establishing connections using 2D NMR data. As an example of an actual experience, the author took several months to crack the structure of his first new compound, wewakazole B (Lopez et al., 2016), a relatively large compound at 1127 g/mol. However, it is not always the compound size that dictates the difficulty of structure elucidation. Rather, it is more in the complexity of the structure and the experience of the chemist. For example, in the case of nocardamin glucuronide (Lopez et al., 2019) having a moderate weight of 777 g/mol, it took just a day to come up with the planar structure because it only consists of a sugar moiety and repeating units of *N*-hydroxy-*N'*-succinylcadaverine. Still, spending a day or even a few hours for structure elucidation sounds inefficient especially with the advent of AI-based software such as SMART which can provide structure prediction in seconds.

To put this technology to the test, a small sample set consisting of 13 compounds and pseudo-compounds were subjected to SMART (Table 1). Various molecular weights (MW) and compound classes were included in the sample set. Actual results from SMART show the predicted HSQC spectrum (Figure 1a) and the table of structure matches (Figure 1b and 1c). Table 1 shows the summary of SMART results including the compound name with the corresponding MW, the match score (exact, close, far, N/A) as described in the methodology, the top compound hit with its MW, and the cosine score which indicates good structure similarity as the value nears 1 (Liu, 2014). For reference, the structures of the published compounds that were incorporated in the sample set are illustrated in Figure 2.

Table 1. Summary of SMART results.

Compound	MW, g/mol	Match in SMART	Cosine Score	Matched Compound	MW, g/mol
2-Methylthio-*N*^{6}-methyl-*cis*-zeatin	280	close	0.80875	*cis*-Zeatin	219
Columbamide D	451	close	0.96388	Columbamide C	423
Columbamide E	485	close	0.97160	Columbamide C	423
Nocardamin glucuronide	777	far	0.87809	Stolonidiol	336
Bisucaberin	400	close	0.94351	Homocitrulline	189
Wewakazole	1141	exact	0.99443	Wewakazole	1141
Wewakazole B	1127	exact	0.89324	Wewakazole B	1127
N-acetyl-8-oxotryptamine	214	exact	0.96493	*N*-acetyl-8-oxotryptamine	216
N-acetyl-α-hydroxy-8-oxotryptamine	232	close	0.87777	3-Acetylindole	159
Unpublished new compound (pyr_novel)	265	far	0.86175	Arbusculidine A	241
Unpublished new compound (5-25-4_new)	1033	close	0.90989	Misakinolide A	668
dummy data	N/A	N/A	0.88786	Hypalocrinin E	435
outrageous data	N/A	N/A	0.99999	Bromoiodoacetamide	263

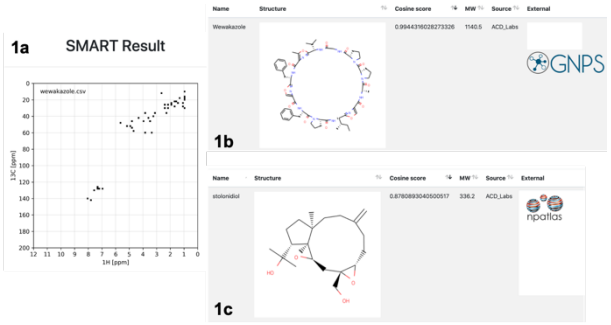

Figure 1. Results from SMART analysis: a) predicted HSQC spectrum for wewakazole, b) table of structure match for wewakazole, and c) table of structure match for nocardamin glucuronide.

Results showed that the SMART tool is suitable for large compounds (> 1000 g/mol) as with the case of wewakazole and wewakazole B. Larger compounds have more data points leading to better matching, and differentiation from other compounds. For smaller compounds, the matching scores are mostly close which means predicted structure is of the same compound class or related to the actual structure. This is evident in columbamides D and E where the top hit for both cases showed columbamide C. Columbamide D has two methylenes (-CH_2CH_2-) more than columbamide C, and these are features that are tricky to detect since the NMR signals usually overlap with each other. On the other hand, columbamide E has an extra chlorine atom over columbamide D, which unfortunately was not recognized by the software. The same can be said

between *N*-acetyl-β-oxotryptamine and *N*-acetyl-α-hydroxy-β-oxotryptamine where the obvious hydroxy group in the latter was not distinguished. This is in contrast to the accuracy observed in differentiating wewakazole B from wewakazole. Although their general structure is similar, there are still variations in the position and amino acid type. This may suggest that SMART has some difficulty in detecting single point differences between smaller molecules.

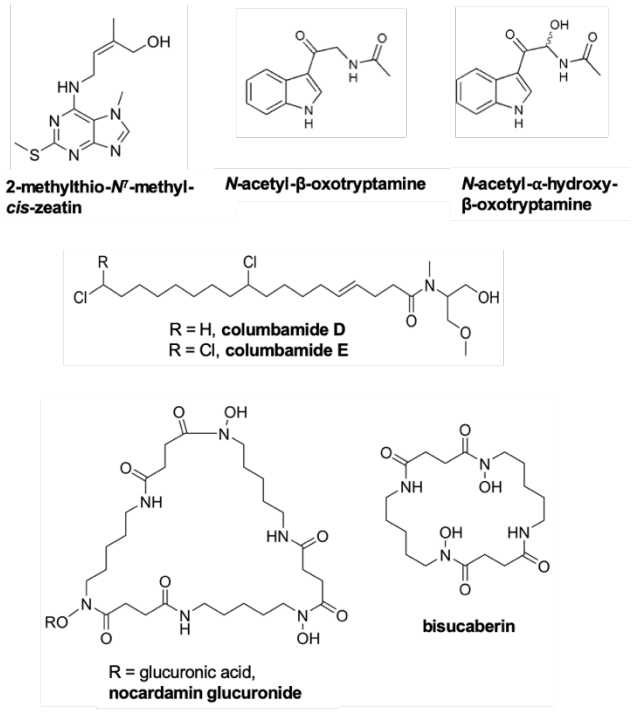

Figure 2. Structures of previously published compounds used in this study.

Likewise, the software failed to identify a good match with nocardamin glucuronide which is relatively moderate in size. On the other hand, better results were obtained for its smaller counterpart, bisucaberin. However, the structure match is just a smaller portion of the bisucaberin which is similar to a single *N*-hydroxy-*N*′-succinylcadaverine.group. Consequently, bisucaberin is a dimer of *N*-hydroxy-*N*′-succinylcadaverine while nocardamin glucuronide is a trimer with an additional glucuronide. This may suggest that the presence of repeating units and the glucuronide component confused SMART leading to subpar matching.

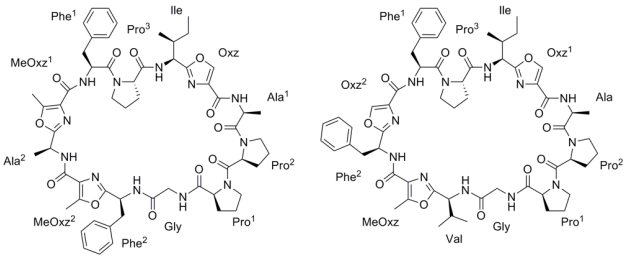

Figure 2 *continued*. Structures of previously published compounds used in this study.

Data for unpublished compounds were also analyzed to check whether SMART can predict compound structures that are presumably not part of its database yet. For analogs or new compounds belonging to a known structural class like compound 5-25-4, a close fit was found using SMART suggesting that the particular compound family is present in its database. Also, since it's a large compound (1033 g/mol), more data points are available for matching as mentioned above. For the putative compound, pyr_novel, a poor match was expected and confirmed with the tool. This may support the idea that it really belongs to a novel structural class.

Finally, "dummy" and "outrageous" data were run in the software to simulate random and illogical data, respectively. For the outrageous data, the expected result was achieved where the default cosine score of 0.99999 and structure match with bromoiodoacetamide were retrieved signifying an error with the data. The same results will show for blank data. Surprisingly, for the dummy data, a structure match with hypalocrinin E was observed indicating that as long as the values are within the accepted threshold for ^{1}H and ^{13}C NMR chemical shifts, a structural match may still be possible.

Overall, having a total of 9 out of 11 exact/close structural matches using the SMART tool is already a great advantage because even if the match was not exact, one is still led to a closely related structural family which is key for structure

elucidation. More importantly, the SMART analysis only takes seconds. In contrast, the same step could take days/hours/weeks using the traditional way of manually reading the NMR data to eliminate several possibilities and pinpoint a specific compound class. Additionally, the program also recommends other useful external databases (Figures 1b and 1c) that may aid the user such as GNPS (Wang et al., 2016), NPATLAS (van Santen et al., 2022), and MIBiG (Terlouw et al., 2023). A caveat remains as the sample set for this study was small and this may not represent the optimum performance of the software. Thus, a full assessment of the SMART tool using a larger and diverse dataset will be considered for future studies.

4. CONCLUSIONS

Natural products research has received a huge boost with the emergence of AI-based structure elucidation tools by streamlining the dereplication of known compounds and aiding in the structure elucidation of new and novel compounds. As shown with the use of the SMART tool in this study, the analysis was almost instantaneous and seamless which led to a more efficient dereplication process. Points for improvement include better differentiation for small molecules, and recognition for compounds with repeating units and symmetry.

5. ACKNOWLEDGMENTS

The author would like to thank his former mentors, Dr. Tatsufumi Okino and Dr. Hiroyuki Osada, for their guidance and supervision throughout his natural products chemistry journey in Japan.

6. REFERENCES

Cao, Q. (2017). Study on the Screening of Active Components of Natural Products Based on Artificial intelligence. 606–609.

Conrado, G. G., da Rosa, R., Reis, R. D., & Pessa, L. R. (2024). Building Natural Product–Based Libraries for Drug Discovery: Challenges and Opportunities from a Brazilian Pharmaceutical Industry Perspective. Revista Brasileira de Farmacognosia.

Cooper, R., & Nicola, G. (2014). Natural Products Chemistry: Sources, Separations and Structures (0 ed.). CRC Press.

Elyashberg, M., & Williams, A. (2021). ACD/Structure Elucidator: 20 Years in the History of Development. Molecules, 26, 6623.

Gerrard, W., Bratholm, L. A., Packer, M. J., Mulholland, A. J., Glowacki, D. R., & Butts, C. P. (2020). IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. Chemical Science, 11, 508–515.

Howarth, A., Ermanis, K., & Goodman, J. M. (2020). DP4-AI automated NMR data analysis: Straight from spectrometer to structure. Chemical Science, 11, 4351–4359.

Ito, T., & Masubuchi, M. (2014). Dereplication of microbial extracts and related analytical technologies. Journal of Antibiotics, 67, 353–360.

Jena, N., Bal, C., & Sharon, A. (2019). Plant and marine products: A promising hope in the search of therapeutics against dengue. In Discovery and Development of Therapeutics from Natural Products Against Neglected Tropical Diseases (pp. 385–405).

Liu, C. (2014). Discriminant analysis and similarity measure. Pattern Recognition, 47, 359–367.

Lopez, J. A. V., Al-Lihaibi, S. S., Alarif, W. M., Abdel-Lateff, A., Nogata, Y., Washio, K., Morikawa, M., & Okino, T. (2016). Wewakazole B, a Cytotoxic Cyanobactin from the Cyanobacterium *Moorea*

producens Collected in the Red Sea. Journal of Natural Products, 79, 1213–1218.

Lopez, J. A. V., Nogawa, T., Futamura, Y., Aono, H., Hashizume, D., & Osada, H. (2021). *N*-Acetyl-α-hydroxy-ß-oxotryptamine, a racemic natural product isolated from *Streptomyces* sp. 80H647. The Journal of Antibiotics, 74, 477–479.

Lopez, J. A. V., Nogawa, T., Futamura, Y., Shimizu, T., & Osada, H. (2019). Nocardamin glucuronide, a new member of the ferrioxamine siderophores isolated from the ascamycin-producing strain *Streptomyces* sp. 80H647. The Journal of Antibiotics, 72, 991–995.

Lopez, J. A. V., Nogawa, T., Yoshida, K., Futamura, Y., & Osada, H. (2021). 2-Methylthio- N 7-methyl- cis- zeatin, a new antimalarial natural product isolated from a *Streptomyces* culture. Bioscience, Biotechnology, and Biochemistry, 86, 31–36.

Lopez, J. A. V., Petitbois, J. G., Vairappan, C. S., Umezawa, T., Matsuda, F., & Okino, T. (2017). Columbamides D and E: Chlorinated Fatty Acid Amides from the Marine Cyanobacterium *Moorea bouillonii* Collected in Malaysia. Organic Letters, 19, 4231–4234.

Mali, S. B. (2023). Cancer treatment: Role of natural products. Time to have a serious rethink. Oral Oncology Reports, 6.

Manochkumar, J., & Ramamoorthy, S. (2024). Artificial intelligence in the 21st century: The treasure hunt for systematic mining of natural products. Current Science, 126, 19–35.

Marcarino, M. O., Zanardi, M. M., Cicetti, S., & Sarotti, A. M. (2020). NMR Calculations with Quantum Methods: Development of New Tools for Structural Elucidation and Beyond. Accounts of Chemical Research, 53, 1922–1932.

Mohamed, A., Nguyen, C. H., & Mamitsuka, H. (2016). Current status and prospects of computational resources for natural product dereplication: A review. Briefings in Bioinformatics, 17, 309–321.

Mouneir, S. M., El-Hagrassi, A. M., & El-Shamy, A. M. (2022). A Review on the Chemical Compositions of Natural Products and Their Role in Setting Current Trends and Future Goals. Egyptian Journal of Chemistry, 65, 491–506.

Nogle, L. M., Marquez, B. L., & Gerwick, W. H. (2003). Wewakazole, a Novel Cyclic Dodecapeptide from a Papua New Guinea *Lyngbya majuscula*. Organic Letters, 5, 3–6.

Reher, R., Kim, H. W., Zhang, C., Mao, H. H., Wang, M., Nothias, L.-F., Caraballo-Rodriguez, A. M., Glukhov, E., Teke, B., Leao, T., Alexander, K. L., Duggan, B. M., Van Everbroeck, E. L., Dorrestein, P. C., Cottrell, G. W., & Gerwick, W. H. (2020). A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. Journal of the American Chemical Society, 142, 4114–4120.

Su, B.-H., Shen, M.-Y., Harn, Y.-C., Wang, S.-Y., Schurz, A., Lin, C., Lin, O. A., & Tseng, Y. J. (2017). An efficient computer-aided structural elucidation strategy for mixtures using an iterative dynamic programming algorithm. Journal of Cheminformatics, 9.

Terlouw, B. R., Blin, K., Navarro-Muñoz, J. C., Avalon, N. E., Chevrette, M. G., Egbert, S., Lee, S., Meijer, D., Recchia, M. J. J., Reitz, Z. L., van Santen, J. A., Selem-Mojica, N., Tørring, T., Zaroubi, L., Alanjary, M., Aleti, G., Aguilar, C., Al-Salihi, S. A. A., Augustijn, H. E., … Medema, M. H. (2023). MIBiG 3.0: A community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Research, 51, D603–D610.

van Santen, J. A., Poynton, E. F., Iskakova, D., McMann, E., Alsup, T. A., Clark, T. N.,

Fergusson, C. H., Fewer, D. P., Hughes, A. H., McCadden, C. A., Parra, J., Soldatou, S., Rudolf, J. D., Janssen, E. M.-L., Duncan, K. R., & Linington, R. G. (2022). The Natural Products Atlas 2.0: A database of microbially-derived natural products. Nucleic Acids Research, 50, D1317–D1323.

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., … Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology, 34, 828–837.

Zhang, C., Idelbayev, Y., Roberts, N., Tao, Y., Nannapaneni, Y., Duggan, B. M., Min, J., Lin, E. C., Gerwick, E. C., Cottrell, G. W., & Gerwick, W. H. (2017). Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. Scientific Reports, 7, 14243.

Zhang, L., Song, J., Kong, L., Yuan, T., Li, W., Zhang, W., Hou, B., Lu, Y., & Du, G. (2020). The strategies and techniques of drug discovery from natural products. *Pharmacology and Therapeutics*, *216*.