JOURNAL OF NATURAL PRODUCTS

pubs.acs.org/jnp Article

DP4+App: Finding the Best Balance between Computational Cost and Predictive Capacity in the Structure Elucidation Process by DP4+. Factors Analysis and Automation

Bruno A. Franco, Ezequiel R. Luciano, Ariel M. Sarotti,* and María M. Zanardi*

Cite This: https://doi.org/10.1021/acs.jnatprod.3c00566 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: DP4+ is one of the most popular methods for the structure elucidation of natural products using NMR calculations. While the method is simple and easy to implement, it requires a series of procedures that can be tedious, coupled with the fact that its computational demand can be high in certain cases. In this work, we made a substantial improvement to these limitations. First, we deeply explored the effect of molecular mechanics architecture on the DP4+ formalism (MM-DP4+). In addition, a Python applet (DP4+App) was developed to automate the entire process, requiring only the Gaussian NMR output files and a spreadsheet containing the experimental NMR data and labels. The script is designed to use the statistical parameters from the original 24 levels of theory (employing B3LYP/6-31G* geometries) and the new 36 levels explored in this work (over MMFF geometries). Furthermore, it enables the development of customizable methods using any desired level of theory, allowing for a free choice of test molecules.

The search for new bioactive compounds remains one of the leading interests for chemistry research, with molecular structure characterization being a key step. Natural products have long been recognized as a valuable source of new chemotherapeutic agents. However, the structural complexity of many natural products poses challenges for chemists in accurately assigning their structures. The intricate nature of natural product structures, often containing multiple stereocenters, fused ring systems, and diverse functional groups, makes their elucidation a complex task. Hence, even with the advancements in analytical techniques (including NMR spectroscopy), there might be ambiguities and uncertainties in the structure determination process that could lead to structural misassignments.[1−3] While total synthesis has played a crucial role in rectifying misassigned structures, it is important to acknowledge that synthesizing all possible stereoisomers becomes impractical and time-consuming, especially as the number of stereogenic centers increases. The continuous evolution of related theories and technologies has turned quantum chemical calculation methods into a standard to guide the structure identification of new complex natural and synthetic compounds.[4] These methods are mostly based on the simulation of NMR parameters coupled with sophisticated techniques for data treatment.[5−8] One of the leading approaches involves the calculation of the probability associated with each structure from a set of two or more candidates. Inspired in Bayes's theory, different statistical-based methods have been developed, including DP4,[9] *J*-DP4,[10] ML-*J*-DP4,[11] DP4+,[12] DiCE,[13] and MESSI.[14] In addition, they were widely explored to determine the structure of both rigid cyclic compounds and conformationally flexible acyclic compounds, with variable levels of confidence. Among them, DP4+ stands out for its exceptional predictive capability across a broad range of molecules.[15]

In the original formulation DP4+ was parametrized for 24 levels of theory, though considering the balance between performance and computational cost the PCM/mPW1PW91/6-31+G**//B3LYP/6-31G* level was recommended for broad applications.[16] However, it was observed that research groups often choose other levels for different reasons.[15] Correlating experimental and calculated values with improperly parametrized levels could potentially have a negative impact on the quality of the results. It is crucial to ensure that the chosen levels of theory and associated parameters are properly validated to accurately represent the molecular systems under investigation. To address these situations, a customizable DP4+ method was developed,[17] which allows preliminary calculations at any desired level of theory using a small set of training molecules. Despite this development enabling the

Received: June 30, 2023

implementation of as many levels as imaginable, the limitation lies in the uncertainty of the new method's predictive capacity, unless the levels have been thoroughly validated. Different factors are linked with the accuracy of the prediction, such as molecular geometry quality; level of theory in NMR calculations, including DFT functional, basis set, or solvation model; and mathematical formalism of probability calculations. In this regard, it is desirable to find the optimal combination of elements that maximize predictive capacity while minimizing computational costs.

The main source of computational cost is related to the geometry optimization step (*vide infra*) of all possible conformations at the QM level. Although there are new tools aimed at mitigating the cost of geometry optimizations, either by narrowing the conformational space with *J* coupling analysis[10] or by enhancing conformational analysis through CONFPASS or CREST, we opted to investigate the DP4+ formalism starting with MM geometries.[18,19] This decision was taken to enable a more straightforward assessment of the impact of the geometry on the DP4+ outcome, as it has been shown that DFT-optimized geometries tend to exhibit overall superiority.[12] However, we considered that this limitation could be compensated by the proper choice of the level of theory for the GIAO NMR calculation step. Another time-consuming stage is related to data processing and file management. Taking this into account, we decided to concurrently develop software capable of automating the DP4+ calculation process and its parametrization.

■ RESULTS AND DISCUSSION

The influence of the main stages required for DP4+ calculations at the recommended level of theory (PCM/mPW1PW91/6-31+G**//B3LYP/6-31G*) in the overall computational cost is illustrated in Figure 1. On average, around 75% of the CPU time is associated with the geometry optimization step, whereas the single-point NMR calculation requires the remaining 25%. However, since the computational cost scales exponentially with the system size, running DFT optimizations might be overwhelming for highly flexible and large molecules for which a large number of conformations should be expected.

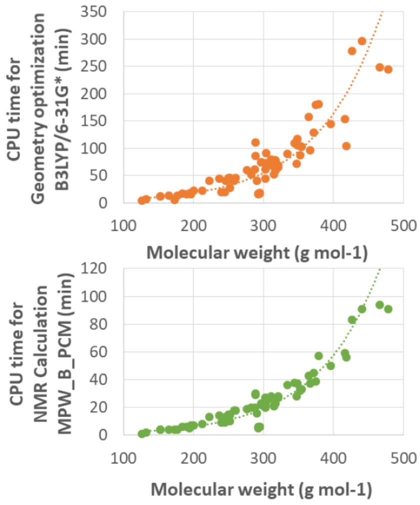


Figure 1. Average time dependence between the quantum calculations (optimization and NMR) and the molecular weight calculated for each conformer of the MM-DP4+ test set (compounds **1**−**72**, see the Supporting Information).

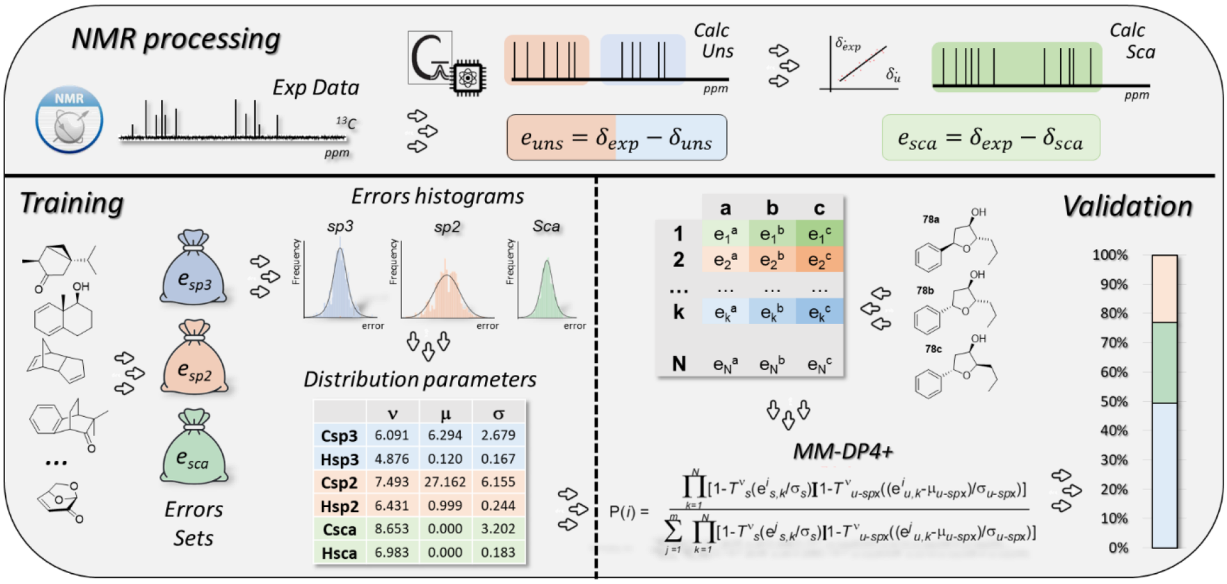
Although the use of DFT-optimized geometries is a key factor accounting for the improved performance of DP4+ over DP4, there are other variables that could be adjusted to maximize the predictive capacity of DP4+ when employing molecular mechanics geometries, hence avoiding the most time-consuming step. In this regard, understanding the effects of the quality of the molecular geometry, the level of theory used in the NMR calculation, and the mathematical formalism on the confidence of Bayesian methods is crucial for improving the accuracy of structure assignments. Our analysis was focused on assessing how each factor could influence the ability to discriminate between diastereomers, aiming to find an optimal balance between computational cost and predictive capacity in the DP4+ structure elucidation process.

Quality of the Molecular Geometry Analysis. Although less refined geometries could be detrimental to the quality of the prediction of the calculated chemical shifts, it is important to recognize that the Bayesian methods are relative in nature, as for a good assignment it is only required that the correct isomer has a better fit with the experimental data than the other candidates. The use of the DP4+ formalism (including both scaled and unscaled data, see below) could compensate for this limitation by providing more spectroscopic information. In this regard, it is important to mention that the DP4+ scheme has never been thoroughly explored in geometries optimized at the molecular mechanics level.

The DP4+ probability is built with a set of 16 statistical parameters $[\mu, \sigma, \nu]$ that describe the distribution of errors (e) between calculated and experimental chemical shifts (δ) for ^{1}H and ^{13}C nuclei, respectively ($e = \delta_{\text{calc}} - \delta_{\text{exp}}$). Those $[\mu, \sigma, \nu]$ values, in turn, show a strong dependence on the level of theory used in the calculation procedure. We have recently shown that incorporating improper values (i.e., correlating calculated values at one theoretical level with statistical parameters from a different level) can exert a profound influence on the outcomes, potentially leading to a complete shift in the assignment's interpretation.[17] Therefore, when exploring a new level of theory, it is crucial to determine its associated statistical parameters through the utilization of a well-defined set of molecules during the training stage. Subsequently, the performance of the approach can be assessed by applying it to a complex set of diastereoisomers during the validation stage. The complete workflow is outlined in Figure 2.

For DP4+ at the molecular mechanics level (MM-DP4+), the parameters were obtained by modeling the same training set of 72 molecules (**1**−**72** in the Supporting Information) used in the original work and calculating the isotropic shielding constant at the PCM/mPW1PW91/6-31+G** level over geometries obtained directly from the conformational search at MMFF without further optimization at the DFT level. After the NMR calculations, 1219 and 1123 individual errors for ^{13}C and ^{1}H were obtained, respectively, for both scaled and unscaled data. Once the randomness and independence of the errors were verified, the statistical adjustment was carried out to determine the parameters $[\mu, \sigma, \nu]$ that describe each *t* probability distribution.

To assess the performance of MM-DP4+, it was tested in 84 real cases of configurational assignment that had been previously determined correctly by DP4+.[12] Analyzing the performance of both methods on the studied systems, a


NMR processing
NMR
Exp Data
^{13}C
ppm
Calc
Uns
ppm
$e_{uns} = \delta_{exp} - \delta_{uns}$
δ_{exp}
δ_{u}
Calc
Sca
ppm
$e_{sca} = \delta_{exp} - \delta_{sca}$
Training
OH
e_{sp3}
e_{sp2}
e_{sca}
Errors
Sets
Errors histograms
sp3
sp2
Sca
Frequency
error
Distribution parameters
a b c
1 e_1^a e_1^b e_1^c
2 e_2^a e_2^b e_2^c
k e_k^a e_k^b e_k^c
N e_N^a e_N^b e_N^c
78a
78b
78c
OH
Validation
MM-DP4+
$P(i) = \frac{\prod_{k=1}^{N}[1-T^{\nu}_{s}(e^{i}_{s,k}/\sigma_s)][1-T^{\nu}_{u\text{-}spx}((e^{i}_{u,k}-\mu_{u\text{-}spx})/\sigma_{u\text{-}spx})]}{\sum_{j=1}^{m}\prod_{k=1}^{N}[1-T^{\nu}_{s}(e^{j}_{s,k}/\sigma_s)][1-T^{\nu}_{u\text{-}spx}((e^{j}_{u,k}-\mu_{u\text{-}spx})/\sigma_{u\text{-}spx})]}$
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

	ν	μ	σ
Csp3	6.091	6.294	2.679
Hsp3	4.876	0.120	0.167
Csp2	7.493	27.162	6.155
Hsp2	6.431	0.999	0.244
Csca	8.653	0.000	3.202
Hsca	6.983	0.000	0.183

Figure 2. Process of training and validation of MM-DP4+ at a new level. **NMR processing**: involves the correlation of the experimental NMR data with the calculated unscaled (uns) and scaled (sca) chemical shifts. **Training**: using a large set of known molecules, the errors (differences between experimental and calculated data) are computed and classified depending on the calculated source (unscaled or scaled) and the hybridization of the atom (sp^2 or sp^3). The corresponding histograms of each series are fitted to *t*-Student′s distributions to estimate the [μ, σ, ν] statistical parameters. **Validation**: the [μ, σ, ν] parameters are used to build new DP4+ probabilities, whose performance is evaluated in a complex set of diastereoisomers.

decrease of 17% in predictive accuracy (ability to assign the right isomer as the most likely candidate) was observed when MM geometries were utilized instead of DFT geometries for the calculations (Figure 3). This result highlights the importance of the geometry optimization step in DP4+. Nonetheless, the fact that 83% of the cases were correctly assigned at a remarkably reduced computational cost paved the way for further exploration into new levels of theory for the NMR calculation step.

Level of Theory in the NMR Calculation. DP4+ was trained and evaluated at 24 levels of theory for the NMR calculation combining two functionals (B3LYP and mPW1PW91) and six basis sets (6-31G*, 6-31G**, 6-31+G**, 6-311G*, 6-311G**, and 6-311+G**) and using two solvation modes: gas phase or PCM. The method's performance demonstrates a dependence on the level of theory, with PCM/mPW1PW91/6-31+G**//B3LYP/6-31G* showing greater accuracy.

There are antecedents where some change in components of the theory level impact positively the predictive capacity of the Bayesian tools. For instance, in the second version of DP4 (DP4.2),[20] a better performance than the original method was found by employing the level mPW1PW91/6-311G* for the GIAO NMR calculation combined with energies obtained at M06-2X/6-31G**. Another important change in performance was observed in polyhydroxylated compounds with a biased description of the conformational landscape caused by intramolecular hydrogen bonding interactions. The use of levels with SMD solvation mode for the energy calculation led to improved Boltzmann descriptions increasing the predictive capacity of DP4+.[21−23]

Due to the fact that geometry quality affects the NMR calculation, it is necessary to establish the most suitable level when MM geometries are used. Therefore, 35 additional levels of theory for the NMR calculation step (over MM geometries) were studied, combining four functionals (B3LYP, mPW1PW91, M06-2X, and ωB97XD), three basis sets (6-31G**, 6-31+G**, and 6-311+G**) and calculations in the gas phase or by using two solvation modes PCM (*Polarizable Continuum Model*)[24,25] and SMD (*Solvation Model Based on Density*)[26] both in chloroform.

For the training step, the set of 72 molecules was calculated for each level under study, obtaining the corresponding error sets which were statistically adjusted to the *t*-Student distributions. Once the six sets of parameters ([μ, σ, ν] for scaled and unscaled ^{1}H and ^{13}C) were determined, each trained level was tested in a set of 34 examples (Supporting Information), and the selection was made considering a representative group of molecules with similar performance as the complete set analyzed at the PCM/mPW1PW91/6-31+G**//MMFF level.

The performance of MM-DP4+ at the different levels was measured by two metrics: percentage of success (examples correctly assigned) and by scoring with a punctuation system. The results are summarized in Figure 4. For a better analysis, they are ordered according to their performance, taking into account the percentage of correct assignments and the total score obtained.

After analyzing the comprehensive ranking, several important observations can be drawn. The majority of levels designed for gas-phase calculations (GAS) were positioned lower in the ranking. This outcome is justified by the inaccuracies that arise due to the absence of solvent, leading to the neglect of crucial stabilizing influences from the surrounding medium. Therefore, it is advisible to refrain from utilizing gas-phase NMR calculations for correlation methods. Conversely, the incorporation of semiempirical solvation models like PCM and SMD showcased enhanced performance, although no model displayed significant superiority over the other. Instead, their effectiveness was contingent upon the specific combination of the employed functional and basis set.

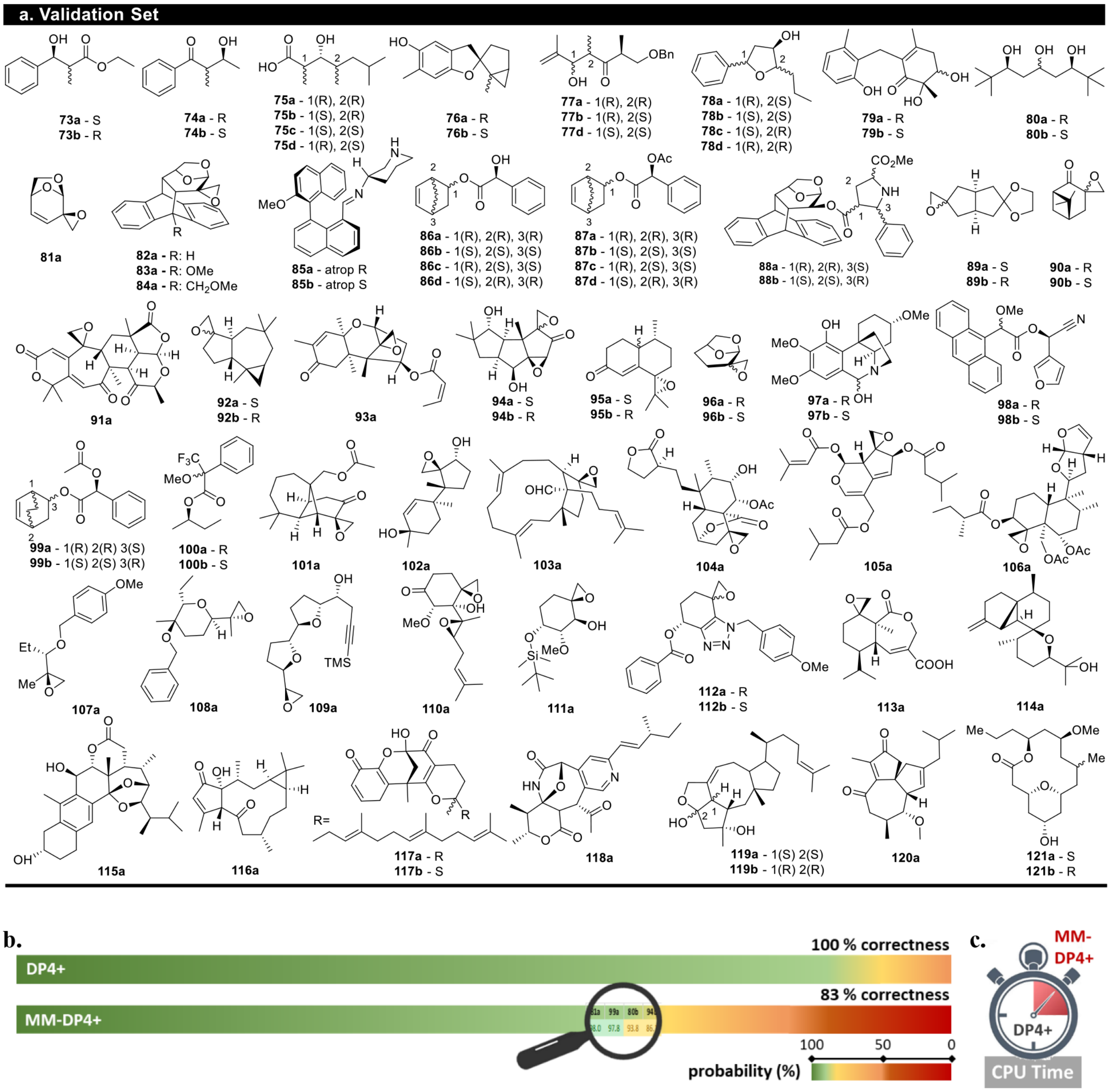


Figure 3. (a) Complete validation set of compounds. (b) Results of DP4+ and MM-DP4+ for compounds **73**−**121** in descending order of probability. The color gradient indicates the success of each method, with green denoting a highly probable and accurate assignment (>97%), while red indicates an incorrect assignment (<50%). (c) Schematic representation of the average time demand of DP4+ and MM-DP4+ calculations at the PCM/mPW1PW91/6-31+G** level.

Regarding the basis set, an increase in precision would be expected with the complexity of the basis functions employed. Therefore, the anticipated trend would suggest that triple-ζ sets (C) should outperform polarized double-ζ sets (B), and in turn, the latter should outperform nonpolarized double-ζ sets (A). However, this theoretical prediction did not align with the observed results. No significant trend was observed in the performance of the method among the tested bases. Instead, the outcome depended on the accompanying functional in each case. Additionally, it is worth highlighting that the most suitable base (6-31+G**) coincides with the one demonstrated for DP4+, which can be considered a suitable basis set for this kind of formalism.

Finally, the performance of the new functional ωB97XD (WB9) stands out, which was located in the five first positions of the ranking. In addition to the previous work by the Kutatelazde group,[27] where it is employed in the assignment of halogenated compounds, it proves to be an excellent choice for NMR calculations. It was followed by the well-established functionals mPW1PW91 and B3LYP, those used in DP4+ and DP4, respectively. In agreement with other reports,[28] the worst performance was assessed to M06-2X, indicating a low capacity for the estimation of magnetic properties despite being recognized for improving energy calculations.

The two levels with the best score were WB9_B_SMD and WB9_C_PCM, both with hit percentages close to 90% and a slightly higher score for the first. To select the best level, a good balance between predictive capacity and computational cost must be taken into account. That is why, having similar results, WB9_B_SMD is preferred for having a better assignment efficiency and the less demanding base function, which includes polarized and diffuse functions, together with a

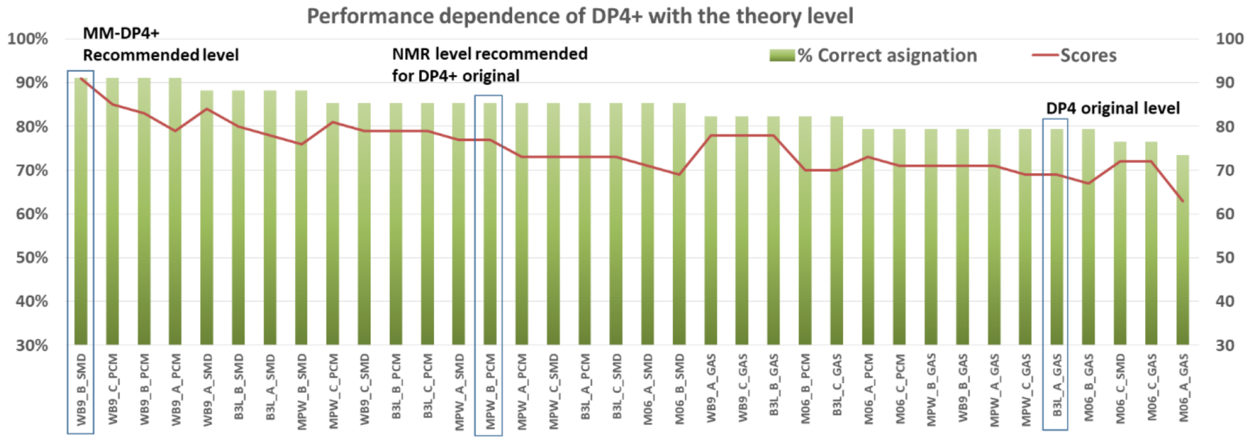


Figure 4. Performance of the MM-DP4+ probability calculated in a reduced set of 34 molecules at the 36 theory levels under study. It was analyzed by two metrics: % of molecules correctly assigned and by a scoring system (3 points if $P > 95\%$, 1 point if $50\% < P < 95\%$, and 0 if the compound is incorrectly assigned). The levels are codified by functional: B3L (B3LYP), MPW (mPW1PW91), M06 (M06-2X), and WB9 (ωB97XD), basis sets: A, 6-31G**; B, 6-31+G**; and C, 6-311+G**, and media: GAS, PCM, and SMD.

Table 1. Mathematical Formalism for the Probability Calculation[a]

	DP4 Original Formalism
Sc*TMS* or **Sc***MSTD*	$P(i) = \dfrac{\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{STD}}{\sum_{j=1}^{m}\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{STD}}$
	where STD could be TMS or MSTD used as reference standard
	DP4+ Formalism
Sc*TMS*+**Uns** or **Sc***MSTD*+**Uns**	$P(i) = \dfrac{\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{STD}[1 - T_{u\text{-}spx}^{\nu}(|e_{u,k}^{i} - \mu_{u\text{-}spx}|/\sigma_{u\text{-}spx})]}{\sum_{j=1}^{m}\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{STD}[1 - T_{u\text{-}spx}^{\nu}(|e_{u,k}^{i} - \mu_{u\text{-}spx}|/\sigma_{u\text{-}spx})]}$
	where STD could be TMS or MSTD used as reference standard
	New Proposed Formalisms
Sc*TMS*+**Sc***MSTD*	$P(i) = \dfrac{\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{TMS}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{MSTD}}{\sum_{j=1}^{m}\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{TMS}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{MSTD}}$
Sc*TMS*+**Sc***MSTD*+**Uns**	$P(i) = \dfrac{\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{TMS}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{MSTD}[1 - T_{u-spx}^{\nu}(|e_{u,k}^{i} - \mu_{u\text{-}spx}|/\sigma_{u\text{-}spx})]}{\sum_{j=1}^{m}\prod_{k=1}^{N}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{TMS}[1 - T_s^{\nu}(|e_{s,k}^{i}|/\sigma_s)]^{MSTD}[1 - T_{u\text{-}spx}^{\nu}(|e_{u,k}^{i} - \mu_{u\text{-}spx}|/\sigma_{u\text{-}spx})]}$

[a] **Sc**: scaled, **Uns**: unscaled, *TMS*: tetramethylsilane, *MSTD*: multistandard approach.

better solvation mode. As a result, the level SMD/ωB97XD/6-31+G**//MMFF is recommended for the use of MM-DP4+.

The Mathematical Formalism of DP4-like Methods. Previous works have demonstrated that the excellent levels of correct classification achieved by DP4+ were interpreted based on constructive offsets for the errors when using both scaled and unscaled ^{1}H and ^{13}C shifts, demonstrating the advantage of incorporating unscaled data into its mathematical formalism.[16,29] Hence, in order to bolster confidence in the assignment, it was determined that all types of data should be incorporated into the probability calculation procedure.

When unscaled data was introduced, the improvement in the method's performance could be attributed to the usefulness of preserving potential systematic errors for the configurational differentiation of the compounds under analysis. However, there is a factor that has not been analyzed yet, related to enhancing the quality of the linear regression employed in the scaled δ calculation. The use of TMS as the reference standard for both sp^2 and sp^3 nuclei resulted in errors dependent on hybridization that could lead to nonideal linear regressions. Therefore, it was proposed to explore a potential enhancement of the *s*DP4+ term by using higher-quality δ_{calc} values obtained through the multistandard approach. This method employs different reference compounds depending on the hybridization of the nucleus under analysis (MeOH for sp^3 and benzene for sp^2).[30,31] Hence, it was proposed to analyze new mathematical formulations to explore which is the most appropriate to carry out a DP4-like analysis. Including calculation of the chemical shifts by the multistandard approach, we proposed using three types of data in the mathematical formalism: δ scaled using TMS as reference standard, δ scaled using the MSTD approach, and unscaled δ with TMS.

The formulations were analyzed at the original DP4 level (B3L_A_GAS) and at the optimal level for MM geometries previously establish (WB9_B_SMD). Using the complete validation set, the predictive capacity of the probability tools was analyzed using each of the formulations: **Sc***TMS,* **Sc***MSTD,* **Sc***TMS*+**Uns**, **Sc***MSTD*+**Uns**, **Sc***TMS*+**Sc***MSTD*, and **Sc***TMS*

+Sc*MSTD*+**Uns.** Each mathematical formalism is detailed in Table 1.

Using the complete validation set (**73**−**121**), the predictive capacity of the six formulations was tested. Results for both levels, represented by the percentages of examples correctly assigned, are summarized in Figure 5.

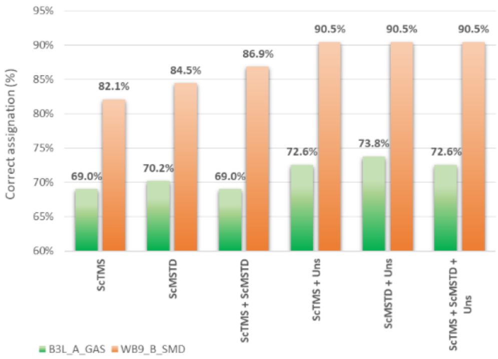


Figure 5. Percentages of examples correctly assigned by the six formulations using the original DP4 level of theory (B3L_A_GAS) and the recommended level for the new method MM-DP4+ (WB9_B_SMD).

The least accurate results (69.0% of examples correctly assigned) were obtained by using B3L_A_GAS with **Sc***TMS* data (level and formulation corresponding to the DP4 original method) and by mixing only scaled data (**Sc***TMS*+**Sc***MSTD*). For both levels of theory, the exclusive use of scaled MSTD data performs better than TMS scaled. However, the results significantly improve by adding unscaled data, such as the DP4+ method (**Sc***TMS*+**Uns**), reaching an assignation of 90.5% for WB9_B_SMD. Moreover, there is no formulation with an outstanding result because **Sc***TMS*+**Uns,** **Sc***MSTD*+**Uns,** and **Sc***TMS*+**Sc***MSTD*+**Uns** have the same outcome. The inclusion of unscaled data does not significantly impact the structural identification when using the multistandard approach. This observation could entail that the most discriminant parameter for the performance improvement was the unscaled probability. Hence, **Sc***TMS*+**Uns** was considered the better and simpler mathematical formalism because it requires only TMS calculations.

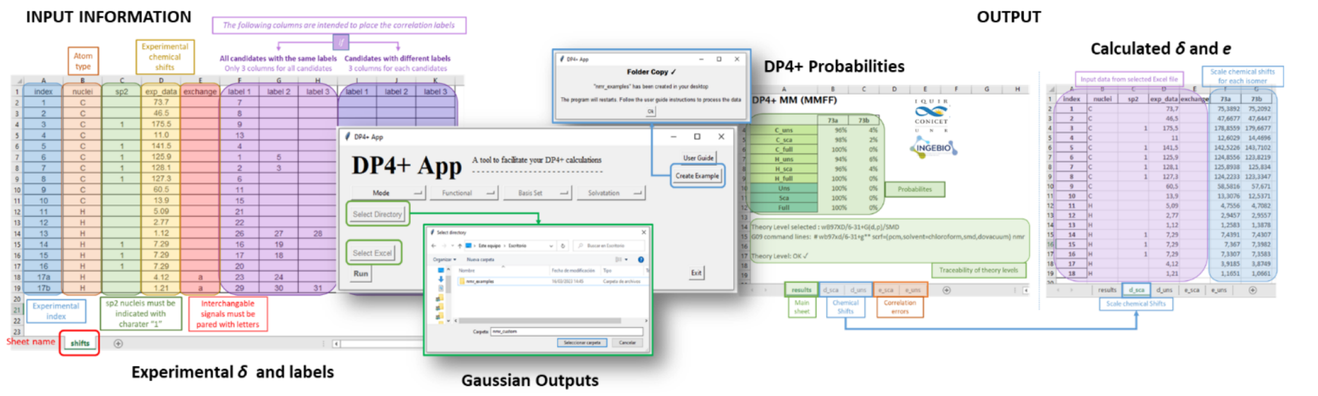
DP4+ Calculation Automation. The publication of tools featuring intuitive interfaces has greatly facilitated the adoption of *in silico* methodologies for structure elucidation purposes. Since the first DP4 applet, other informatics tools for structure elucidation have been published, each with varying degrees of ease and challenges in their utilization.[11,12,14,21,32−39] Among them, DP4+ stands out not only for its exceptional performance but also for its widespread popularity owing to its seamless user experience. The isotropic shielding constants and the experimental chemical shifts can be effortlessly loaded into an Excel spreadsheet, thereby streamlining the entire process. While these tools are undeniably innovative and user-friendly, it is worth noting that certain user operations are still required.

To simplify this task further, the use of advanced programming software for automation is an ever-growing field. As a result, a Python package named DP4+App has been developed to facilitate the implementation of DP4+-like calculations. This applet is launched under an open-source MIT license, enabling users to calculate the DP4+ probability at the original 24 levels and MM-DP4+ at the new 36 levels and even customize the method if desired. By using Gaussian output files and spreadsheet integration with experimental data and labels, this software enables users to obtain probabilities directly. This update offers the advantage of liberating users from the extraction and conditioning of NMR calculation results, saving processing time, eliminating transcription errors, and providing faster and more reliable results.

DP4+App arises as a new tool that combines the power of a graphical user interface (GUI) with its executable applet structure (Figure 6). Comprised of two modules, it offers a comprehensive set of functionalities for seamless DP4+-like correlation calculations. The main module enables DP4+, MM-DP4+, and Custom-DP4+ probability calculations with precision and reliability. Conversely, the second module is designed for training customized theory levels, thereby enhancing the software's flexibility and adaptability.

Furthermore, to improve the user experience and ensure accurate interpretation of results, DP4+App incorporates additional chemical and computational criteria. These criteria serve as warning signs, alerting users about anomalous values, misidentified experimental data, and mismatched theory levels. By flagging potential misuse and errors, these incorporated features assist users in effectively analyzing their outputs and maintaining the integrity of their research.

Results are presented in a comprehensive spreadsheet format. In addition to providing probabilities of the isomers, the spreadsheet offers complete traceability of the process, including essential information such as shielding constants (*σ*),


INPUT INFORMATION
The following columns are intended to place the correlation labels
Atom type
Experimental chemical shifts
All candidates with the same labels
Only 3 columns for all candidates
Candidates with different labels
3 columns for each candidates
Experimental index
sp2 nucleis must be indicated with charater "1"
Interchangable signals must be pared with letters
Sheet name
shifts
Experimental δ and labels
DP4+ App
Folder Copy ✓
"nmr_examples" has been created in your desktop
The program will restarts. Follow the user guide instructions to process the data
Ok
DP4+ App
A tool to facilitate your DP4+ calculations
Mode
Functional
Basis Set
Solvatation
User Guide
Create Example
Select Directory
Select Excel
Run
Exit
Select directory
Este equipo › Escritorio
Buscar en Escritorio
Organizar
Nueva carpeta
Nombre
Fecha de modificación
Tipo
nmr_examples
Carpeta de archivos
Carpeta: nmr_custom
Seleccionar carpeta
Cancelar
Gaussian Outputs
OUTPUT
DP4+ Probabilities
DP4+ MM (MMFF)
IQUIR
CONICET
UNR
INGEBIO
Probabilites
Traceability of theory levels
results
d_sca
d_uns
e_sca
e_uns
Main sheet
Chemical Shifts
Correlation errors
Calculated δ and e
Input data from selected Excel file
Scale chemical shifts for each isomer
Scale chemical Shifts

index	nuclei	sp2	exp_data	exchange	label 1	label 2	label 3	label 1	label 2	label 3
1	C		73.7		7					
2	C		46.5		8					
3	C	1	175.5		9					
4	C		11.0		13					
5	C	1	141.5		4					
6	C	1	125.9		1	5				
7	C	1	128.1		2	3				
8	C	1	127.3		6					
9	C		60.5		11					
10	C		13.9		15					
11	H		5.09		21					
12	H		2.77		22					
13	H		1.12		26	27	28			
14	H	1	7.29		16	19				
15	H	1	7.29		17	18				
16	H	1	7.29		20					
17a	H		4.12	a	23	24				
17b	H		1.21	a	29	30	31			

	73a	73b
C_uns	96%	4%
C_sca	98%	2%
C_full	100%	0%
H_uns	94%	6%
H_sca	96%	4%
H_full	100%	0%
Uns	100%	0%
Sca	100%	0%
Full	100%	0%

Theory Level selected : wB97XD/6-31+G(d,p)/SMD
G09 command lines: #wb97xd/6-31+g** scrf=(pcm,solvent=chloroform,smd,dovacuum) nmr
Theory Level: OK ✓

index	nuclei	sp2	exp_data	exchange	73a	73b
1	C		73,7		75,3892	75,2092
2	C		46,5		47,6677	47,6447
3	C	1	175,5		178,8559	179,6677
4	C		11		12,6029	14,4696
5	C	1	141,5		142,5226	143,7102
6	C	1	125,9		124,8556	123,8219
7	C	1	128,1		125,8938	125,834
8	C	1	127,3		124,2233	123,3347
9	C		60,5		58,5816	57,671
10	C		13,9		13,3076	12,5371
11	H		5,09		4,7556	4,7082
12	H		2,77		2,9457	2,9557
13	H		1,12		1,2583	1,3878
14	H	1	7,29		7,4391	7,4307
15	H	1	7,29		7,367	7,3982
16	H	1	7,29		7,3307	7,3583
17	H		4,12		3,9185	3,8749
18	H		1,21		1,1651	1,0661

Figure 6. DP4+App GUI, INPUT, and OUTPUT files.

chemical shifts (δ), and errors (e) associated with each candidate. Both scaled and unscaled versions of the chemical shifts and errors are included, ensuring a thorough understanding of the data and facilitating further analysis and comparison.

To obtain further information regarding the functionality of the DP4+App, it is advisible to refer to the User Guide.

General Recommendations. As the capabilities of powerful tools expand, users are empowered to undertake larger and more complex calculations. However, it is essential to recognize that enhanced capabilities also bring a corresponding need for diligent and responsible usage. While DP4+App has caution filters in place, it remains important to consider general recommendations when conducting a DP4+-like analysis. These recommendations serve as guiding principles to ensure the accuracy and reliability of results.

(a) *Preserving the resources:* It is important to take into account factors such as the number of isomer candidates, the conformational landscape, and computational resources to get accurate results within a reasonable time frame. When dealing with flexible molecules, a thorough conformational sampling ought to be done using a safe energy cutoff (5 kcal/mol). Keeping the candidates set to a minimum offers several advantages; it reduces both the overall computational cost and the risk of calculated data for an incorrect isomer yielding a better fit than the correct candidate.

(b) *Choosing the most suitable method:* For small sets with limited conformational flexibility, it is recommended to invest in highly accurate DP4+ calculations (PCM/mPW1PW91/6-31+G**//B3LYP/6-31G*). However, for a large number of candidates or a complex conformational space, starting with screening options using MM-DP4+ at SMD/ωB97XD/6-31+G**//MMFF could be advantageous. The most reliable correlation results are obtained using DP4+, so analysis should be concluded with this calculation using the top candidates identified during screening. It is also important to note that highly flexible polyhydroxy compounds are better tackled with specialized software like MESSI.[14]

(c) *Choosing theory levels*: The App admits the use of 60 validated levels of theory. Whenever possible, it is recommended to use NMR theory levels PCM/mPW1PW91/6-31+G** for DP4+ and SMD/ωB97XD/6-31+G** for MM-DP4+ because they have demonstrated better performance. However, if the desired theory level is not parametrized, the app allows generating a new level by following the instructions provided in the Custom-DP4+ method. The choice of appropriate level of theory is crucial because Bayesian methods construct the candidate probability by multiplying the individual probabilities associated with each individual error. The individual probabilities are *t*-distribution dependent, and which parameters describe it [μ, σ, ν] depend on the level of theory. The use of improper distribution might impact the DP4+ values leading to potentially wrong assignments.

(d) *Training method*: If a customizable method is chosen it is important to consider that the sample size affects the accuracy and reliability of estimations. With a larger sample size, uncertainty is reduced, and more precise estimates of the parameters of the Student's *t*-distribution [μ, σ, ν] are obtained. The values of probabilities do not strongly depend on freedom degrees (ν), but if a reduced number of compounds are used the recommendation is to use a mean value as was previously described.[17]

(e) *Checking all conformations*: When DFT geometries are used for the NMR shift calculations, it is crucial to ensure that all conformations are optimized at the desired level and remove any duplicates. It is a suitable habit to perform a frequency analysis on the most stable structures to verify the nature of the stationary points obtained.

(f) *Validation of the results*: It is always recommended to validate the results with the experimental NMR information available (such as homo- and heteronuclear coupling values and/or interatomic distances obtained through NOE/ROE experiments).

CONCLUSION

The main limitation of the use of DP4+ in optimal conditions is the high computational cost, associated with performing geometry optimizations at quantum levels. Thirty-six new levels were trained for the calculation of DP4+ over MM geometries. To know their performance, these levels were validated in real cases of configurational assignments, determining that the best level to perform MM-DP4+ analysis is SMD/ωB97XD/6-31+G**//MMFF, with 90.5% of accuracy and an average CPU time savings of 70%.

We studied the influence on the predictive capacity of each of the factors involved in the process of DP4+ probability calculation, to better understand the uncertainty when the customizable method is used. The suggestions are M062X is the worst functional to perform NMR shift calculations and ωB97XD exhibits the best performance. The basis set does not show clear tendencies, but it is important to point out that no accuracy is gained with a triple-ζ basis, and its computational cost does not justify its use. With respect to the media, we recommend the use of PCM or SMD as solvation modes.

To reduce the analysis time necessary for a DP4+ assignment, an interactive and easy-to-use App was developed. It allows automation of the entire calculation process, requiring only the outputs of the NMR Gaussian calculation and a spreadsheet containing the experimental data and labels. DP4+App performs probability calculations over B3LYP/6-31G* geometries in combination with 24 NMR levels and 36 levels of NMR for MMFF geometries. In these 60 cases, the certainty of the method has been tested. If another level is desired, the App enables the calculations of the customizable method.

The DP4+App code was launched by MIT license, and its installer is available at https://github.com/Sarotti-Lab/DP4plus-App (including instructions and tutorials) and from the Python Package Index (https://pypi.org/project/dp4plus-app/).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jnatprod.3c00566.

Computational methods; case study of the workflow for the use of DP4+ and MM-DP4+ (PDF)

AUTHOR INFORMATION

Corresponding Authors

Ariel M. Sarotti − *Instituto de Química Rosario (CONICET), Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Rosario 2000, Argentina;* orcid.org/0000-0002-8151-0306; Email: sarotti@iquir-conicet.gov.ar

María M. Zanardi − *Instituto de Investigaciones en Ingeniería Ambiental, Química y Biotecnología Aplicada (INGEBIO), Facultad de Química e Ingeniería del Rosario, Pontificia Universidad Católica Argentina, Rosario 2000, Argentina;* orcid.org/0000-0002-7145-5358; Email: zanardi@inv.rosario-conicet.gov.ar

Authors

Bruno A. Franco − *Instituto de Investigaciones en Ingeniería Ambiental, Química y Biotecnología Aplicada (INGEBIO), Facultad de Química e Ingeniería del Rosario, Pontificia Universidad Católica Argentina, Rosario 2000, Argentina*

Ezequiel R. Luciano − *Instituto de Investigaciones en Ingeniería Ambiental, Química y Biotecnología Aplicada (INGEBIO), Facultad de Química e Ingeniería del Rosario, Pontificia Universidad Católica Argentina, Rosario 2000, Argentina*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jnatprod.3c00566

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was funded by ANPCyT (PICT-2016-0116, PICT-2017-1524, and PICT-2019-4052) and CONICET (PIP 11220200102205CO and PIBAA 28720210101005CO). B.A.F. and E.R.L. thank CONICET for the award of a fellowship.

REFERENCES

(1) Yoo, H. D.; Nam, S. J.; Chin, Y. W.; Kim, M. S. *Arch. Pharm. Res.* **2016**, *39* (2), 143−153.

(2) Nicolaou, K. C.; Snyder, S. A. *Angew. Chemie - Int. Ed.* **2005**, *44* (7), 1012−1044.

(3) Chhetri, B. K.; Lavoie, S.; Sweeney-Jones, A. M.; Kubanek, J. *Nat. Prod. Rep.* **2018**, *35* (6), 514−531.

(4) Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. *Chem. Rev.* **2012**, *112* (3), 1839−1862.

(5) Grimblat, N.; Sarotti, A. M. *Chem. - A Eur. J.* **2016**, *22* (35), 12246−12261.

(6) Lauro, G.; Das, P.; Riccio, R.; Reddy, D. S.; Bifulco, G. *J. Org. Chem.* **2020**, *85*, 3297.

(7) Costa, F. L. P.; De Albuquerque, A. C. F.; Fiorot, R. G.; Lião, L. M.; Martorano, L. H.; Mota, G. V. S.; Valverde, A. L.; Carneiro, J. W. M.; Dos Santos Junior, F. M. *Org. Chem. Front.* **2021**, *8* (9), 2019−2058.

(8) Marcarino, M. O.; Zanardi, M. M.; Cicetti, S.; Sarotti, A. M. *Acc. Chem. Res.* **2020**, *53* (9), 1922−1932.

(9) Smith, S. G.; Goodman, J. M. *J. Am. Chem. Soc.* **2010**, *132* (37), 12946−12959.

(10) Grimblat, N.; Gavín, J. A.; Hernández Daranas, A.; Sarotti, A. M. *Org. Lett.* **2019**, *21* (11), 4003−4007.

(11) Tsai, Y. H.; Amichetti, M.; Zanardi, M. M.; Grimson, R.; Daranas, A. H.; Sarotti, A. M. *Org. Lett.* **2022**, *24* (41), 7487−7491.

(12) Grimblat, N.; Zanardi, M. M.; Sarotti, A. M. *J. Org. Chem.* **2015**, *80* (24), 12526−12534.

(13) Xin, D.; Jones, P. J.; Gonnella, N. C. *J. Org. Chem.* **2018**, *83* (9), 5035−5043.

(14) Marcarino, M. O.; Passaglia, L.; Zanardi, M. M.; Sarotti, A. M. *Chem. - A Eur. J.* **2023**, *29*, No. e202300420.

(15) Marcarino, M. O.; Cicetti, S.; Zanardi, M. M.; Sarotti, A. M. *Nat. Prod. Rep.* **2022**, *39* (1), 58−76.

(16) Zanardi, M. M.; Suárez, A. G.; Sarotti, A. M. *J. Org. Chem.* **2017**, *82* (4), 1873−1879.

(17) Zanardi, M. M.; Sarotti, A. M. *J. Org. Chem.* **2021**, *86* (12), 8544−8548.

(18) Lam, C. C.; Goodman, J. M. *CONFPASS: Fast DFT Re-Optimizations of Structures from Conformation Searches* **2023**, *63*, 4364.

(19) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22* (14), 7169−7192.

(20) Ermanis, K.; Parkes, K. E. B.; Agback, T.; Goodman, J. M. *Org. Biomol. Chem.* **2017**, *15* (42), 8998−9007.

(21) Zanardi, M. M.; Marcarino, M. O.; Sarotti, A. M. *Org. Lett.* **2020**, *22* (1), 52−56.

(22) Marcarino, M. O.; Zanardi, M. M.; Sarotti, A. M. *Org. Lett.* **2020**, *22* (9), 3561−3565.

(23) Zanardi, M. M.; Sortino, M. A.; Sarotti, A. M. *Carbohydr. Res.* **2019**, *474*, 72−79.

(24) Amovilli, C.; Barone, V.; Cammi, R.; Cancès, E.; Cossi, M.; Mennucci, B.; Pomelli, C. S.; Tomasi, J. *Adv. Quantum Chem.* **1998**, *32* (C), 227−261.

(25) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55* (1), 117−129.

(26) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113* (18), 6378−6396.

(27) Kutateladze, A. G.; Reddy, D. S. *J. Org. Chem.* **2017**, *82* (7), 3368−3381.

(28) Safi, Z. S.; Wazzan, N. *Sci. Rep.* **2022**, *12* (1), 17798.

(29) Zanardi, M. M.; Biglione, F. A.; Sortino, M. A.; Sarotti, A. M. *J. Org. Chem.* **2018**, *83* (19), 11839−11849.

(30) Sarotti, A. M.; Pellegrinet, S. C. *J. Org. Chem.* **2009**, *74* (19), 7254−7260.

(31) Sarotti, A. M.; Pellegrinet, S. C. *J. Org. Chem.* **2012**, *77* (14), 6059−6065.

(32) Zanardi, M. M.; Sarotti, A. M. *J. Org. Chem.* **2015**, *80* (19), 9371−9378.

(33) Grimblat, N.; Gavín, J. A.; Hernández Daranas, A.; Sarotti, A. M. *Org. Lett.* **2019**, *21* (11), 4003.

(34) Ermanis, K.; Parkes, K. E. B.; Agback, T.; Goodman, J. M. *Org. Biomol. Chem.* **2016**, *14* (16), 3943−3949.

(35) Troche-Pesqueira, E.; Anklin, C.; Gil, R. R.; Navarro-Vázquez, A. *Angew. Chemie - Int. Ed.* **2017**, *56* (13), 3660−3664.

(36) Yesiltepe, Y.; Nuñez, J. R.; Colby, S. M.; Thomas, D. G.; Borkum, M. I.; Reardon, P. N.; Washton, N. M.; Metz, T. O.; Teeguarden, J. G.; Govind, N.; et al. *J. Cheminform.* **2018**, *10* (1), 52−68.

(37) Shenderovich, I. G. *J. Chem. Phys.* **2018**, *148* (12), No. 124313.

(38) Howarth, A.; Ermanis, K.; Goodman, J. M. *Chem. Sci.* **2020**, *11* (17), 4351−4359.

(39) Willoughby, P. H.; Jansma, M. J.; Hoye, T. R. *Nat. Protoc.* **2014**, *9* (3), 643−660.