DP4+ App

https://github.com/Sarotti-Lab/DP4plus-App

sarotti@iquir-conicet.gov.ar

Instructive and general recommendations for Custom-DP4+

Content

Overview and usage recommendations

The DP4+ App is a comprehensive software designed to perform parameterized DP4+ and MM-DP4+ calculations seamlessly. Additionally, it offers the capability to conduct Custom-DP4+ calculations, allowing users to parameterize any required level of theory. With its friendly graphical interface, users can easily manage multiple Gaussian calculations and automate information processing for probabilistic calculations.

To get started with the application, simply create a folder and ensure that it contains the following files:

- Well-labeled Gaussian output files: These files should include NMR calculations for all conformers of each isomeric candidate. Make sure to label them appropriately for easy identification.
- Excel file with experimental information: This file should contain the necessary experimental data along with the correlation labels for each nucleus corresponding to the Gaussian calculations.

By providing these files, the DP4+ App can efficiently process the information and perform the desired calculations.

To ensure optimal use of the program, it is recommended to follow the guidelines below:

- Minimize the number of candidates: While the DP4+ App can handle any number of isomers, keeping the candidate count 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 with experimental values compared to the correct candidate.
- Conduct a thorough conformational search: It is essential to obtain an accurate depiction of the conformational landscape of the system under study. Care should be taken to avoid improper computational work that could potentially affect the overall results. Systematic sampling is always

recommended, but in the case of highly flexible molecules, stochastic searches with a reasonably large number of steps should be carried out. All conformations within a safe energy window from the corresponding global minimum should be retained to avoid missing potentially significant conformations. For this application, it is advised to use a 5 to 10 kcal/mol cutoff value, employing the MMFF force field.

- Adhere to the suggested theory levels: It is important to use the recommended theory levels since DP4+ and MM-DP4+ were optimized for these levels. If the desired theory level is not parameterized, there is the option to parametrize the desired level by following the instructions provided in the Custom-DP4+ method.
- Ensure correct assignment of NMR data: The use of unassigned or misassigned NMR data can lead to erroneous results. When dealing with equivalent nuclei that undergo fast interconversion (e.g., methyl or some equivalent methylene groups), it is necessary to average the chemical shifts. Treating each proton signal independently, such as computing different chemical shifts for the same methyl group, is incorrect. Additionally, diastereotopic methylene protons often pose challenges with arbitrary correlation. Unless additional NMR information, such as NOE or J coupling, is available to discriminate between the pro-*R* and pro-*S* signals, the most suitable approach is to treat them as interchangeable signals. Detailed instructions are provided to assist you in addressing these issues effectively.

Custom-DP4+

The **Custom** module offers enhanced flexibility in DP4+ calculations by enabling the creation of custom distributions for preference theory levels. This can be achieved in two ways: by loading pre-defined distribution parameters or by training the method using experimental data and NMR calculations. Within the **Custom** section, there are three tabs:

- **Calc**: Facilitates DP4+ calculations using an already parameterized level
- **Input**: Allows loading a pre-parameterized theory level based on its distribution parameters and TMS standard tensors.
- **Train**: Enables training a theory level using experimental data and Gaussian GIAO-NMR calculations

The characteristics of each section will be described below.

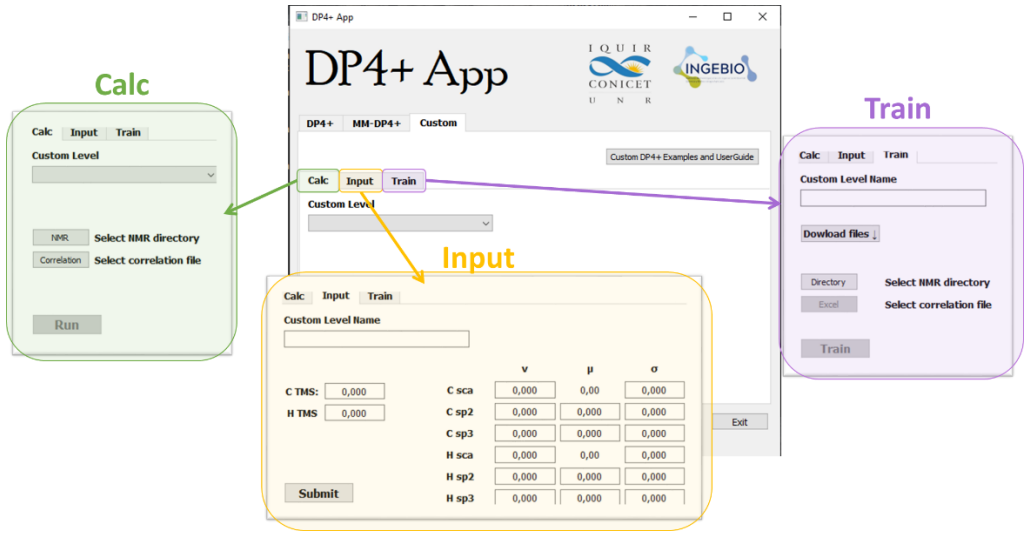

Figure 1. DP4+ Custom module overview

Calculation with Custom-DP4+

After loading or training a new theory level, it becomes available for calculations in the **Calc** tab. Select the desired level from the dropdown list and then specify the working folder and the correlation spreadsheet. Utilize the designated buttons situated beneath the theory-level selection panel. Each button will instantiate a navigation dialogue, facilitating the selection of the desired directory and file.

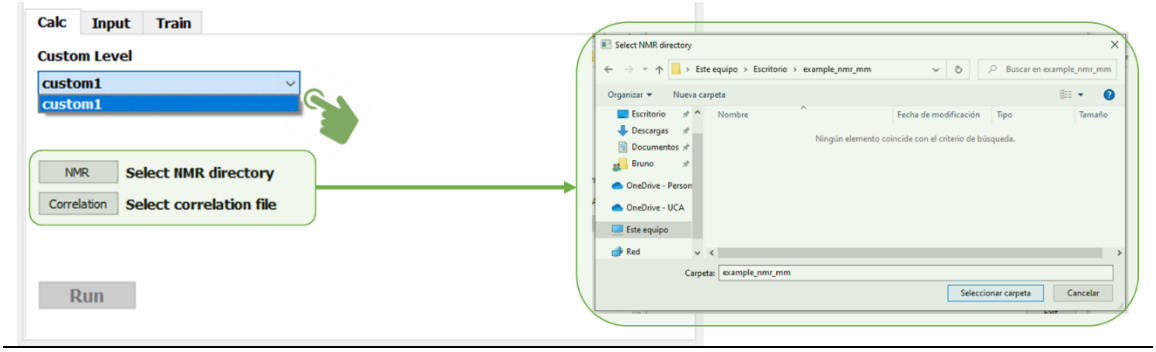

Figure 2. Calc tab usage diagram

To perform correlation calculations using the DP4+ App, it is necessary to prepare the required files. The program offers a range of controls to ensure precise data entry. The following guidelines should be followed to set up the files correctly:

The correlation file (.xlsx) should include the information in the "***shifts***" sheet. This sheet will be the only one read by the program and must adhere to the structure defined in Figure 3 (refer to Warnings and Input Control). Ensure that the column headers match accordingly. For isomers with the same labels, only three columns are required. However, if isomers use different labels, each candidate should have three labeling columns (label 1 | label 2 | label 3). The name of this document does not have any specific requirements as it will be selected individually.

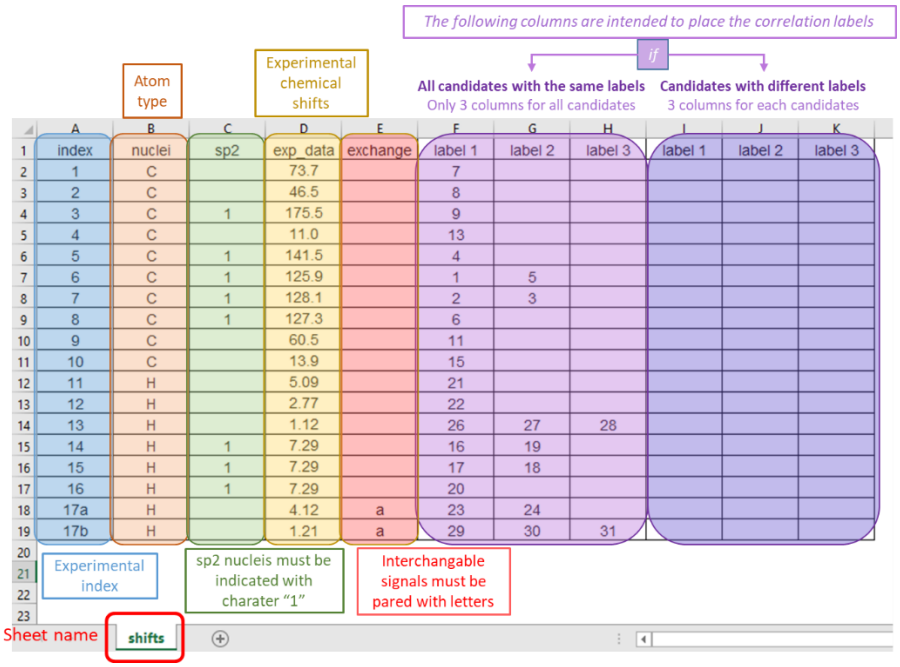

Figure 3. Correlation sheet (experimental information and correlation labels)

NMR files have to be generated from calculations using the Gaussian software (.log or .out files) with the command line containing "`#... nmr`". Label these files according to the following convention, "`n_m_*_nmr.log`", where :

- *"n"* represents the isomer ID,
- "*m*" denotes the conformer number, and
- "*" indicates a user annotation.

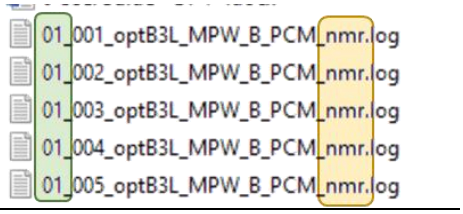

Figure 4. NMR Gaussian outputs

Results output

After the completion of the calculation, a pop-up will confirm the successful execution, and results will be presented in an Excel file located within the selected working folder, named *Custom-DP4plus_results*.

The Excel file will consist of:

- *Probability Results*: This sheet, labeled as "**results**," provides the candidate's probabilities categorized by their nuclei, scaling, and the full version. It also displays the selected theory level, the command line used for the Gaussian calculations, and the automatic coincidence check.
- *Tensors*: Contains the Boltzmann-weighted tensors sorted according to the input labels.
- *Chemical Shifts*: Two sheets are dedicated to displaying the chemical shifts obtained from the calculations ("**d_sca**" and "**d_uns**").
- *Correlation Errors*: two sheets are allocated to present the correlation errors. ("**e_sca**" and "**e_uns**").
- *Parameters*: Reports distribution parameters and reference standard for correlation calculation.

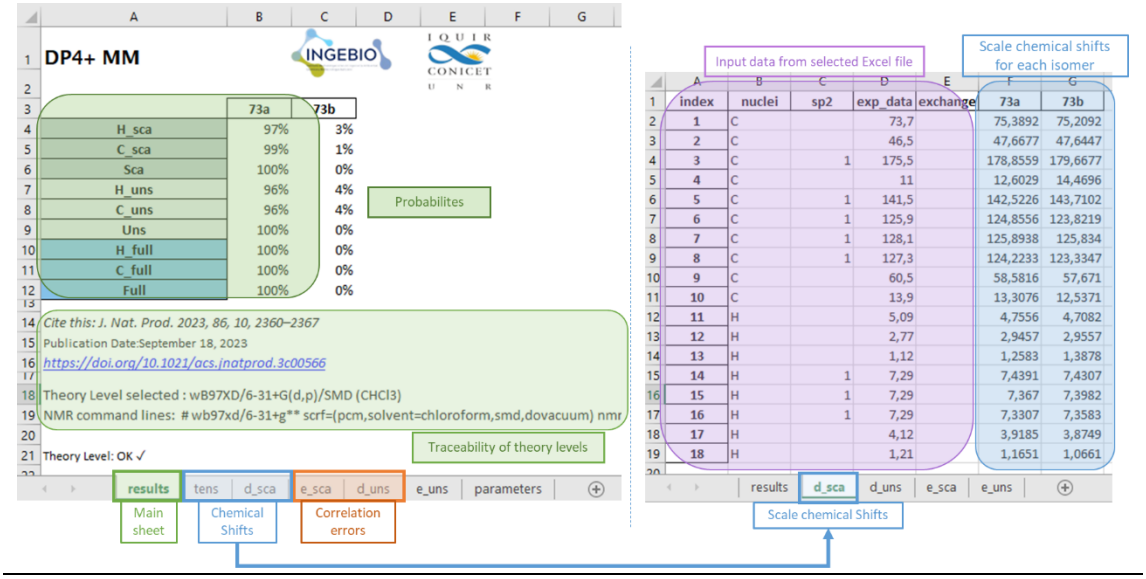

Figure 5. Output results file

Input already parametrize theory level

To load a pre-parameterized theory level, the distribution parameter values must be uploaded in the Input window. Additionally, the TMS tensors corresponding to the theory level being used need to be specified, and a valid name must be assigned.

The name must be lowercase and contain no special characters. No value should be 0, and only 3 decimal places are allowed. The status bar will provide guidance throughout the process. Once all fields have been completed, click Submit to store the level. A notification will confirm successful storage.

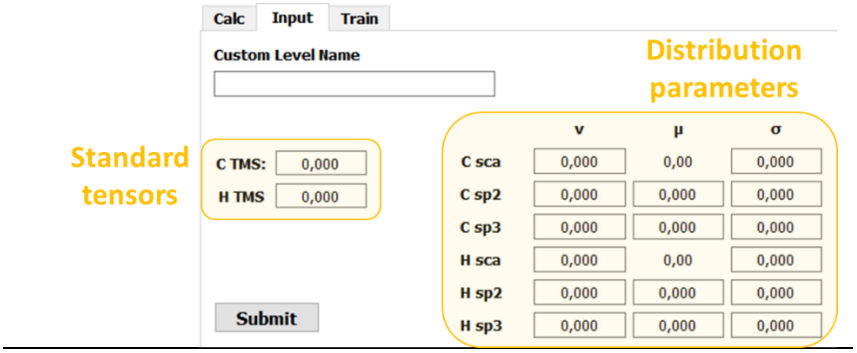

Figure 6. Input tab frames

Train a new theory level

A theory level can be parameterized from scratch using a spreadsheet with experimental data and correlation labels, along with NMR calculations of reference molecules. In the **Train** tab, the "**Download Files**" button provides templates for 8 suggested molecules, complete with their respective correlation sheets and chemical shifts.

Additionally, it is possible to choose between structures optimized with B3LYP/6-31G* or MMFF, to train DP4+ or MM-DP4+ respectively. The interactive window simplifies the command line insertion into Gaussian input files. Files should be labeled in the same way as calculations in the **Calc** tab, and for this case, TMS should be included. Calculations must be performed using Gaussian 09 or 16.

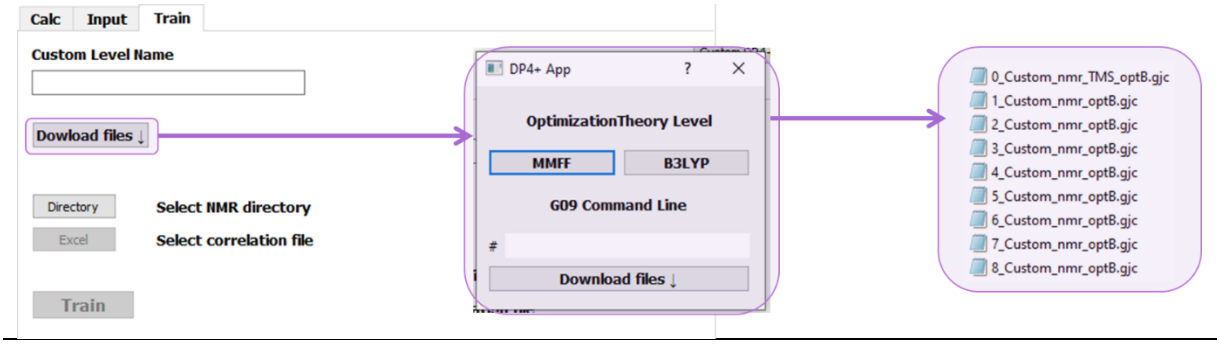

Figure 7. Template molecule download scheme

The spreadsheet contains experimental datasets for each structure, along with their assigned correlation labels. The correlation information should follow the same structure as shown in Figure 5. However, in this case, each sheet will be named after its corresponding molecule, meaning "**shifts**" sheet will not be used. To better understand this procedure, we recommend exploring the example included with this manual.

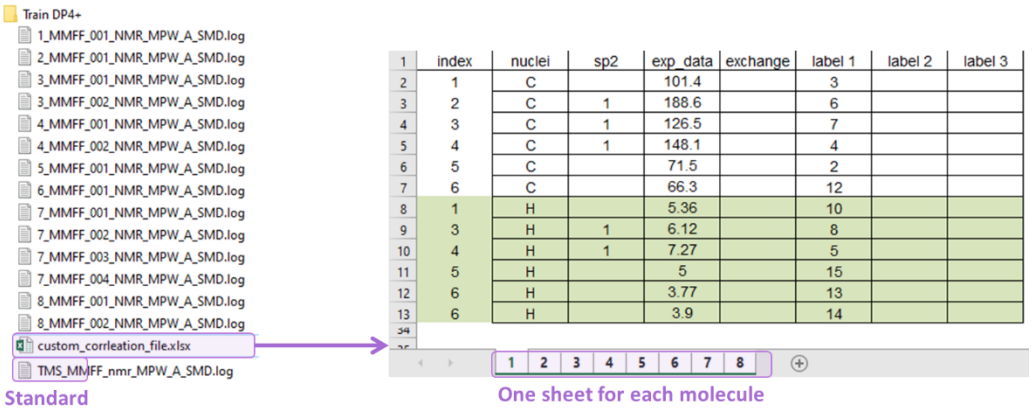

Figure 8. Spreadsheet structure with correlation and experimental data

Training can be initiated once NMR calculations are complete. To begin, assign a valid name in the frame and specify the working directory and the provided correlation file. The program will verify the accuracy of the input data prior to commencing the parameterization process.

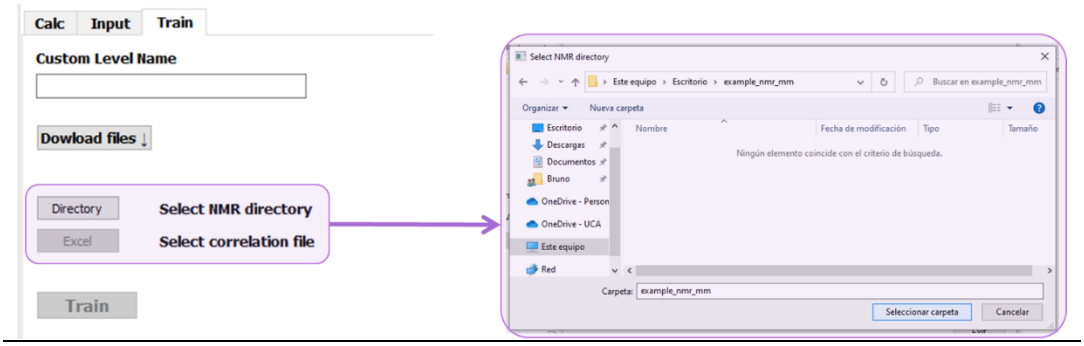

Figure 9. Training usage diagram

During training, sampling points will be counted across the six DP4+ distributions. It's well established that with limited data, the algorithm's t-distributions exhibit high degrees of freedom (v > 20), resembling normal distributions. Our previous work[1] indicates that utilizing average degrees of freedom from validated theory levels is preferable. Therefore, when parameterizing with a small molecular dataset, there is an option to select either averaged or actual degrees of freedom.

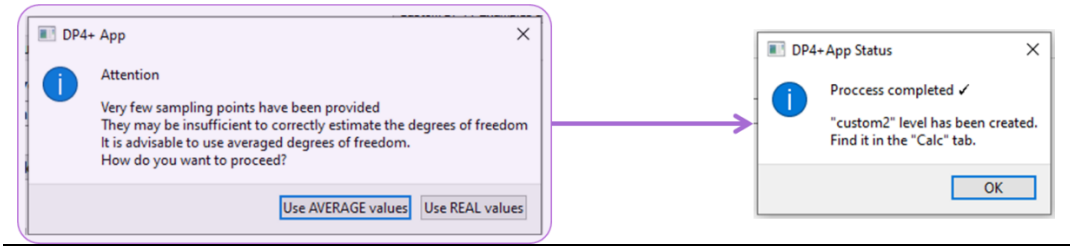

Figure 10. Degree of freedom decision window

Warnings and Input control

To enhance the user's understanding of anomalous results in DP4+ type calculations, DP4+App has implemented a comprehensive warning system. This system assists users in interpreting and identifying any unusual outcomes that may arise during the calculation process. Additionally, the application includes multiple checkpoints to validate the accuracy of data entry, ensuring reliable and consistent results.

Gaussian calculation files

DP4+App ensures the completeness of information from Gaussian calculation files by verifying the presence of the "*Normal Termination*" indicator in each file. If any file lacks this indicator, it will be automatically moved to a folder labeled "*fail files*" within the working folder.

[1] **Zanardi, M. M., & Sarotti, A. M. (2021).** *Sensitivity analysis of DP4+ with the probability distribution terms: Development of a universal and customizable method*. The Journal of Organic Chemistry, 86(12), 8544-8548.

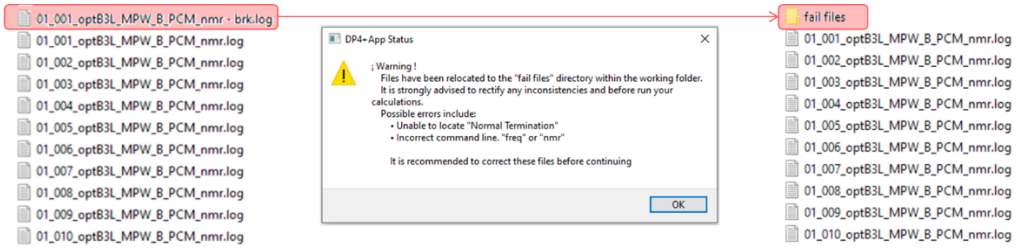

Figure 11. Example of a broken file (identified with brk) that is separated in fail folder

Data spreadsheet

The Excel spreadsheet provided for DP4+ App must follow a specific format, as illustrated in Figure 5. To ensure the accuracy of data entry, the program performs several checks on the spreadsheet:

- Column not found: Detects missing or incorrect column headers.
- Data not found: Identifies missing data in the *'nuclei'*, *'exp_data'*, or *'labels'* columns.
- Incorrect data: Checks for valid data types in specific columns
  - For the 'nuclei' column, the data must be either 'C' or 'H'.
  - For the 'exp_data' column, the data must be a numerical value.
  - For the 'labels' column, the data must be an integer number.
  - For the 'sp2' column, the data must be 'X', 'x', or '1'.
- Mismatched diastereotopic labels: Notifies when diastereotopic labels are not paired correctly.
- Different number of candidate isomers and set of labels: Detects inconsistencies between the number of candidate isomers and the set of labels used.

If any of these situations occur, the program will be unable to proceed with the calculation. Therefore, the user needs to rectify any inconsistencies before proceeding further.

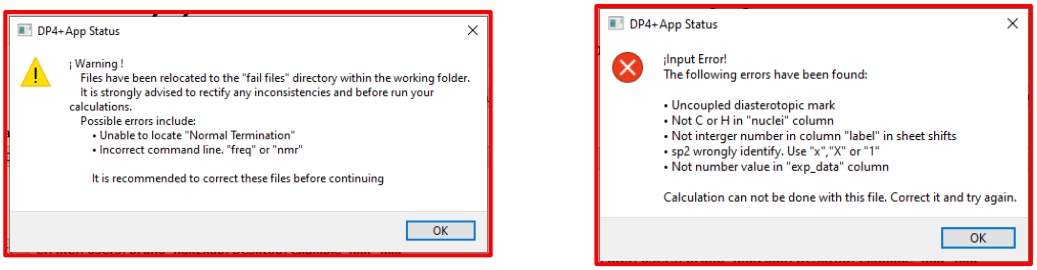

Figure 11. Examples of warning and input error

Calculation aborted

The software includes control checks that are started after the initiation of each calculation. Failure of these checks will result in the calculation being terminated and an error reported. Common causes for termination include:

- Inability to match a corresponding label (columns: label 1 | label 2 | label 3).

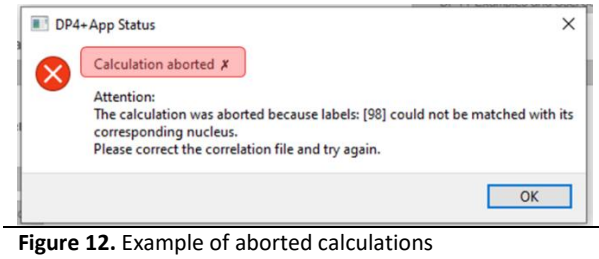

Figure 12. Example of aborted calculations

Questionable values

DP4+ App applies specific chemical criteria to validate the results of calculations. Any values that deviate from these criteria will trigger warnings to alert the user. The following conditions are considered questionable:

- σ_H > 6ppm and σ_C > 120 ppm, not marked as *sp2*
- σ_H > 14ppm, identified as ^{13}C
- $e_{sca\text{-}H}$ > 0.7 and $e_{sca\text{-}C}$ > 10, related to possible miscorrelation/missed assignment

While these warnings indicate potential issues, the calculations can still proceed. The highlighted cells will draw attention to the questionable values, allowing the user to evaluate their significance. If possible, it is recommended to experiment with different paths and interpret the results to make informed decisions.

For DP4+ type calculations, the warnings will be displayed on the **e_sca** sheet of the results. n the training scenarios, the specific points of interest or emphasis are indicated within the original correlation file, specifically on the sheets that have been marked as noteworthy.

	A	B	C	D	E	F	G	H
1	index	nuclei	sp2	exp_data	exchange	73a	73b	
2	1	C		73.7		-8.64583	-8.97892	
3	2	C		46.5		-14.5614	-14.7679	
4	3	C	1	175.5		13.15343	14.23648	
5	4	C		11		-20.9491	-18.9718	
6	5	C	1	141.5		3.750364	5.202798	
7	6	C		125.9		-1.75426	-2.99944	
8	7	C	1	128.1		-2.71407	-2.79137	
9	8	C	1	127.3		-3.90952	-4.98248	
10	9	C		60.5		-15.5238	-16.7685	
11	10	C		13.9		51.15417	50.82108	
12	11	H		5.09		-5.45829	-5.50291	
13	12	H		2.77		7.399502	7.581393	
14	13	H		1.12		-0.6099	-0.68703	
15	14	H	1	7.29		-8.33225	-8.39644	
16	15	H		7.29		-8.31415	-8.38816	
17	16	H	1	7.29		20.26961	20.40886	
18	17	H		4.12		-4.27804	-4.32062	
19	18	H		1.21		-0.67649	-0.69509	
20								

results | tens | d_sca | e_sca | d_uns | e_uns | parameters

Figure 13. Example of deviant values for parametrization method

Malfunctions report

We strive to provide a reliable and efficient user experience with DP4+ App. However, if you encounter any malfunctions or issues while using the software, we appreciate your assistance in reporting them. By providing detailed information about the problem you encountered, you can contribute to the continuous improvement of DP4+ App.

Please report any faulty operations or unexpected behavior to the following email addresses:

- brunoafranco@uca.edu.ar
- zanardi@inv.rosario-conicet.gov.ar
- sarotti@iquir-conicet.gov.ar