DP4+ App

Instructive, general recommendations and case study

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.

Probability calculations: DP4+, MM-DP4+ and Custom-DP4+

Prepare your files

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

The Excel file (.xlsx) should include the information on the "shifts" sheet. This sheet will be the only one read by the program and must adhere to the structure defined in Figure 1 (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 during the process.

Figure 1. Excel sheet (experimental information and correlation labels)

The Gaussian files should be the results of "nmr" calculations obtained from the Gaussian software (.log or .out files). Follow the labeling convention as follows: *n_m_*_nmr.log*, where:

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

To select these files, use the "Select..." buttons, which will prompt popup windows for file selection. It is highly recommended to utilize the provided example ("Create Example" button) as a template to build your working directory effectively.

By adhering to these guidelines and organizing the files in a suitable manner, a seamless and precise execution of the probability calculations can be achieved using the DP4+ App.
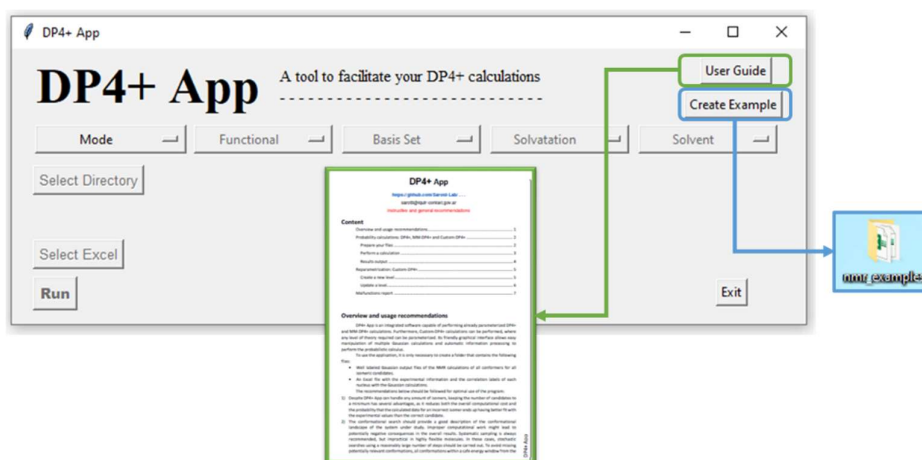

Figure 2. User Guide and Example buttons

Perform a calculation

With the DP4+ App, it is possible to determine the correlation probability using 60 pre-parameterized theory levels. These levels encompass a combination of different functionals, basis sets, and solvation modes. Out of the total, 24 levels were derived from geometries optimized using quantum mechanics at the B3LYP/6-31G* level (QM mode), while the remaining 36 levels were obtained through molecular mechanics using the MMFF force field (MM mode).

QM mode theory levels combinations

Functional	Basis Set		Solvatation
B3L	6-31G(d)	6-31G(d,p)	GAS
mPW1PW91	6-31+G(d,p)	6-311G(d)	PCM (CH$_3$Cl)
	6-311G(d,p)	6-311+G(d,p)	

MM mode theory levels combinations

Functional	Basis Set	Solvatation
B3L	6-31G(d,p)	GAS
M06-2x	6-31+G(d,p)	PCM (CH$_3$Cl)
mPW1PW91	6-311+G(d,p)	SMD (CH$_3$Cl)
wB97XD		

For those theory levels using solvatation mode (PCM or SMD), there are 10 solvents available: *CHCl3, CH2Cl2, CCl4, H2O, MeOH, MeCN, DMSO, THF, Pyridine, Acetone, Benzene*. If a different solvent is needed, the *'Other'* option can be used for entering H^1 and C^{13} TMS tensors (a popup window will appear after pressing the "*Run*" button).
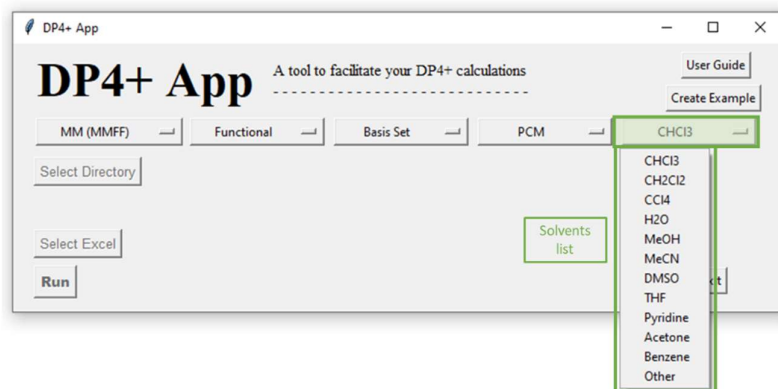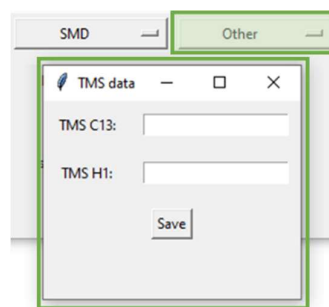
Figure 3. Solvent drop-down list

Figure 4. Pop-up window for "Other" option

While it is possible to perform calculations at any selected level, for optimal utilization of the DP4+ App, the program ensures the alignment between the command lines of the Gaussian files and the selected theory level. In the event of a mismatch, a warning will appear, but it will not impede the execution of the calculation (refer to Figure 5).
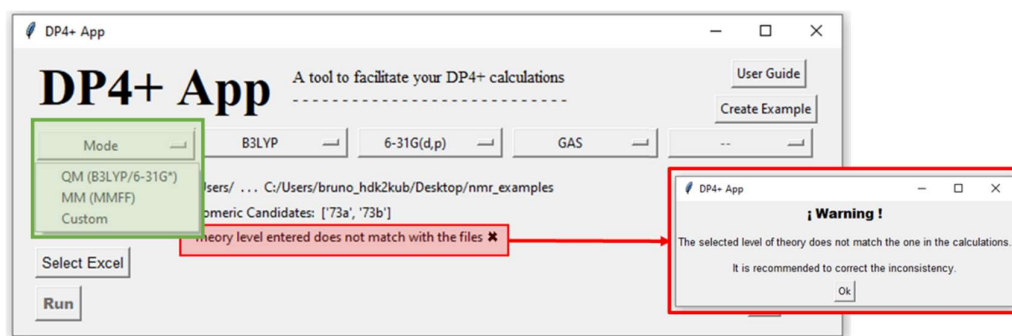

Figure 5. Modes selection and example of miss matching theory level and command lines

For performing calculations at a theory level that is not mentioned, it is necessary to parameterize it first by following the instructions provided in the Custom mode (covered in the next section). In the Custom mode, the automatic validation of command line is not performed. Instead, manual verification can be conducted on the final results sheet.

By adhering to these guidelines, calculations using the DP4+ App can be performed effectively. This approach ensures compatibility between theory levels and provides the opportunity to utilize the Custom mode for parameterizing new levels if necessary.

Results output

After the completion of the calculation, a pop-up notification will confirm the successful execution, and the results will be presented in an Excel file located within the selected working folder. The name of the output file will correspond to the calculation mode utilized.

The Excel file will consist of:

- *Probability Results*: This sheet, labeled as "**results**," provides the probabilities of the candidates 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*: Similarly, two sheets are allocated for presenting the correlation errors. ("**e_sca**" and "**e_uns**").

- *Parameters*: Reports distribution parameters and reference standard for correlation calculation.
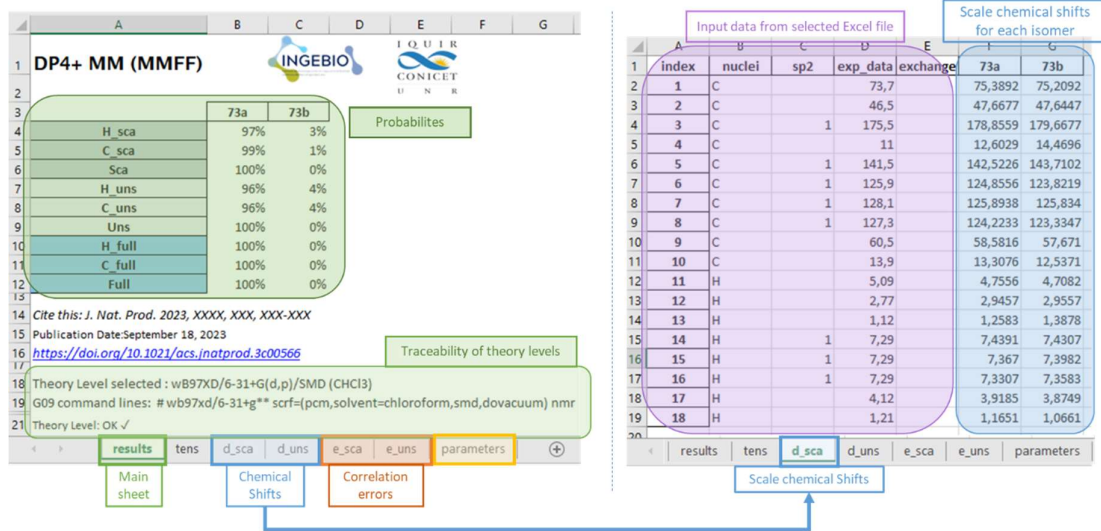

Figure 6. Output Excel file

In cases where the selected theory level does not align with the Gaussian calculation command line, a warning will be displayed, indicating any inconsistencies encountered (refer to Figure 7).

In the Custom mode, it will be explicitly stated that the theory level cannot be verified automatically, and the responsibility falls upon the user to perform the necessary verification.

By reviewing the comprehensive results output in the provided Excel file, users can analyze the probability results, chemical shifts, and correlation errors, and ensure the appropriate utilization of the selected theory level and Custom mode if applicable.
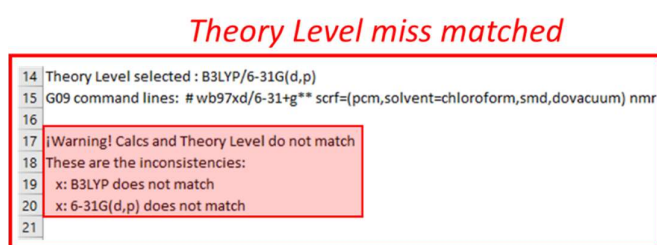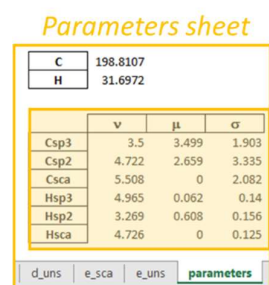

Figure 7. Examples of theory level miss selection

Figure 8. *"Parameters" sheet view*

Reparametrization: Custom-DP4+

Create a new level

In the Custom mode of the main window, the option **+ new** redirects to the reparameterization module. Within this module, the parameterization mode can be selected (**Input Parameter** or **Load files**), and a name can be assigned to the custom level. It is important to use a name consisting only of numbers and lowercase letters, while avoiding the use of special characters.
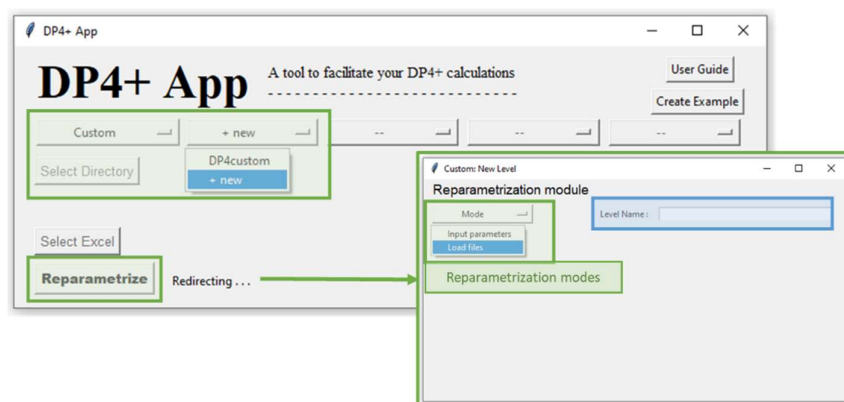
Figure 9. Selection to create a new custom level and redirecting window.

If a theory level has already been parameterized, the distribution parameters and tensors of the TMS (Figure 7.a) can be manually entered using the keyboard.

However, for calculating distribution statistics, it is necessary to load a working directory and the corresponding Excel file like in the main window. The working folder should contain the NMR calculations for all labeled conformations in the format 'n_m_nmr.log', and the TMS calculation specified as 'TMS_nmr.log'. The Excel spreadsheet should have the same number of sheets as the parametrization molecules, with each sheet named after the ID of the respective compounds. The application will automatically determine the parameters, and you can view them when using the custom level in a calculation.
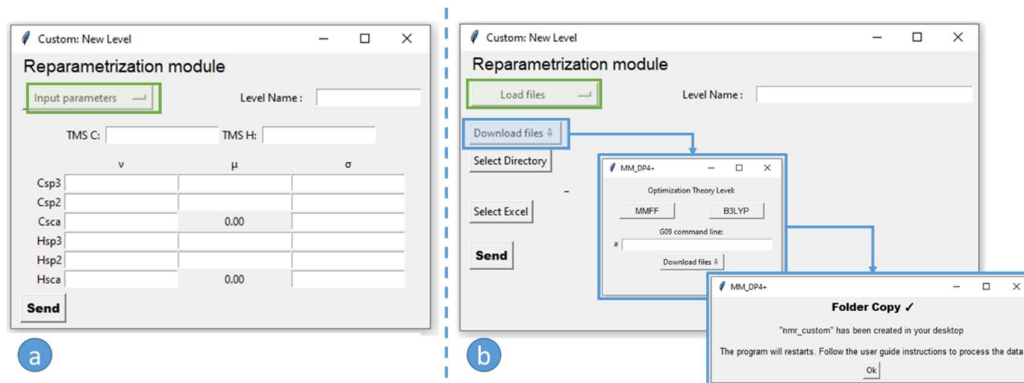

Figure 7. a) Input parameters mode: allow to entry the values of an already parametrized level. b) Load mode: capable of automatically determine the distribution parameters using the training set calculus. Also, offers templates of the training set in QM y MM optimization.

Based on the findings presented in the article *J. Org. Chem.* **2021**, *86*, 12, 8544–8548, it is recommended to use the set of 8 molecules listed in Table 1 for parametrizing your theory level. These molecules have been thoroughly tested and proven to accurately estimate the distribution parameters. To facilitate the preparation process, the input files for Gaussian calculations in both MM (using the MMFF force field) and QM (using the B3LYP theory level) optimization can be downloaded (Figure 7.b). In the input window, the command line can be specified as required by Gaussian, along with the 'nmr' calculation instruction.

The Excel file containing the experimental data and the assigned correlation labels will be provided along with the Gaussian inputs. Each molecule's data set will be placed in a sheet named after the associated ID, following the same structure as the correlation "shifts" sheet shown in Figure 1.
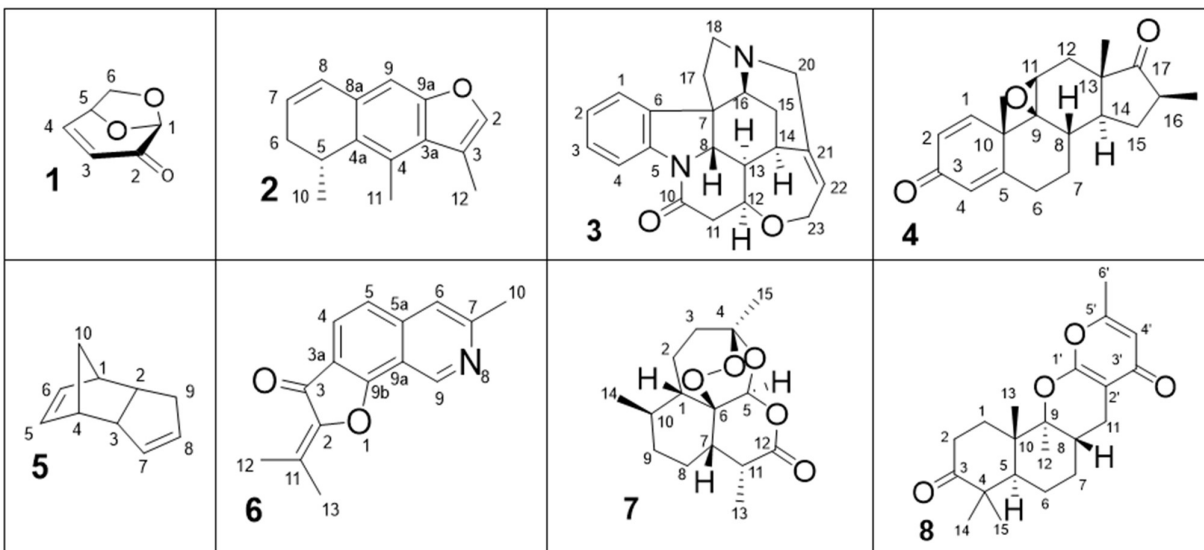
Table 1. Training set molecules with experimental labels.

As mentioned by Sarotti (2021)[1], it is crucial to consider the number of sampling points used for parameterization analysis to avoid potential inaccuracies in estimating the degrees of freedom. If a data set consists of fewer than 150 elements, a pop-up window will notify you about the potential inaccuracies in fitting a t-Student distribution. This notification specifically applies to the recommended set mentioned in Table 1.

There are two options: either utilizing the average values from the original publication, which have been proven to yield accurate results for DP4+ type calculations, or retaining the actual values. If the estimated values are chosen, it is strongly recommended to verify that the degrees of freedom (v) in the calculations are less than 30.

By adhering to these guidelines for reparameterization in the Custom-DP4+ mode, a new level can be created, allowing for the effective determination of distribution parameters for accurate correlation calculations.
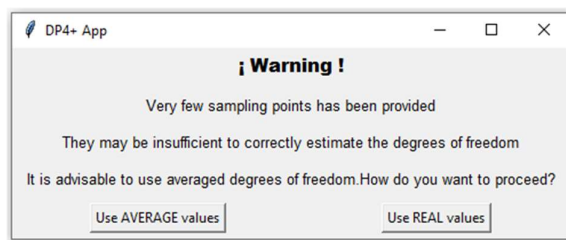

Figure 8. Warning popup for insufficient parametrization set

Update a level

To update a custom level, the steps outlined in the previous section can be followed, and the name of the desired custom level can be overwritten. If a level with the same name already exists, a popup will notify and prompt a decision. If both parametrizations are to be retained, the name can be modified accordingly by going back (refer to Figure 8). It is important to note that the update can be generated in any mode, and it is not necessary to use the same mode as before.

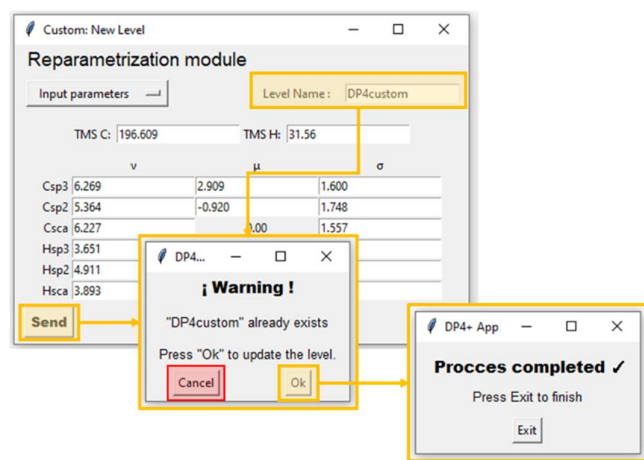[1] *J. Org. Chem.* **2021**, *86, 12*, 8544–8548

Figure 8. Example of update/overwriting a custom level

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.

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:

- $\sigma_H > 6$ppm and $\sigma_C > 120$ ppm, not marked as *sp2*
- $\sigma_H > 14$ppm, identified as ^{13}C
- $e_{sca-H} > 0.7$ and $e_{sca-C} > 10$, 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. During the parameterization process, a popup will appear to confirm whether to proceed, and the input spreadsheet will show the highlighted values.

Figure 9. Example of deviant values for parametrization method

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 "Removed Files." A popup will then offer the option to proceed with the calculation without those files or initiate a recalculation.
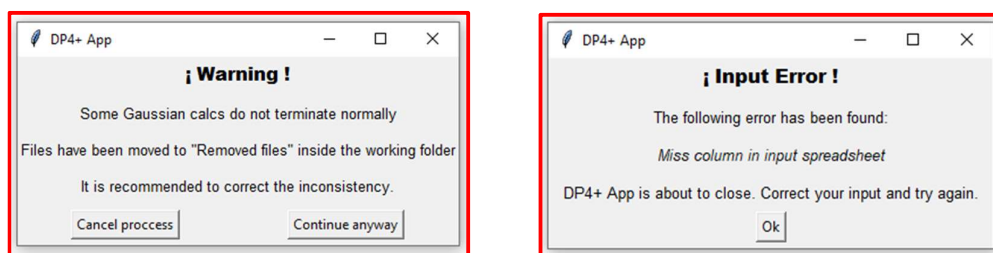

Figure 10. Examples of warning and input error

Data spreadsheet

The Excel spreadsheet provided for DP4+ App must follow a specific format, as illustrated in Figure 1. 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'.
- Label index out of range: Alerts if a label does not match any nuclei in the Gaussian calculation matrix.
- 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, it is essential for the user to rectify any inconsistencies before proceeding further.

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