DP4+ App

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

sarotti@iquir-conicet.gov.ar

Instructive, general recommendations and case study

Content

Overview and usage recommendations

DP4+ App is an integrated software capable of performing already parameterized DP4+ and MM-DP4+ calculations. Furthermore, Custom-DP4+ calculations can be performed, where any level of theory required can be parameterized. Its friendly graphical interface allows easy manipulation of multiple Gaussian calculations and automatic information processing to perform the probabilistic calculus.

To use the application, it is only necessary to create a folder that contains the following files:

- Well-labeled Gaussian output files, from NMR calculations, of all conformers for all isomeric candidates.
- An Excel file with the experimental information and the correlation labels of each nucleus with the Gaussian calculations.

The recommendations below should be followed for the optimal use of the program:

1) Despite DP4+ App can handle any amount of isomers, keeping the number of candidates to a minimum has several advantages, as it reduces both the overall computational cost and the probability that the calculated data for an incorrect isomer ends up having a better fit with the experimental values than the correct candidate.
2) The conformational search should provide a good description of the conformational landscape of the system under study. Improper computational work might lead to potentially negative consequences in the overall results. Systematic sampling is always recommended, but impractical in highly flexible molecules.

In those cases, stochastic searches using a reasonably large number of steps should be carried out. All conformations within a safe energy window from the corresponding global minimum should be kept to avoid missing potentially relevant conformations. We recommend a 10 kcal/mol cutoff value for this application using the MMFF force field.

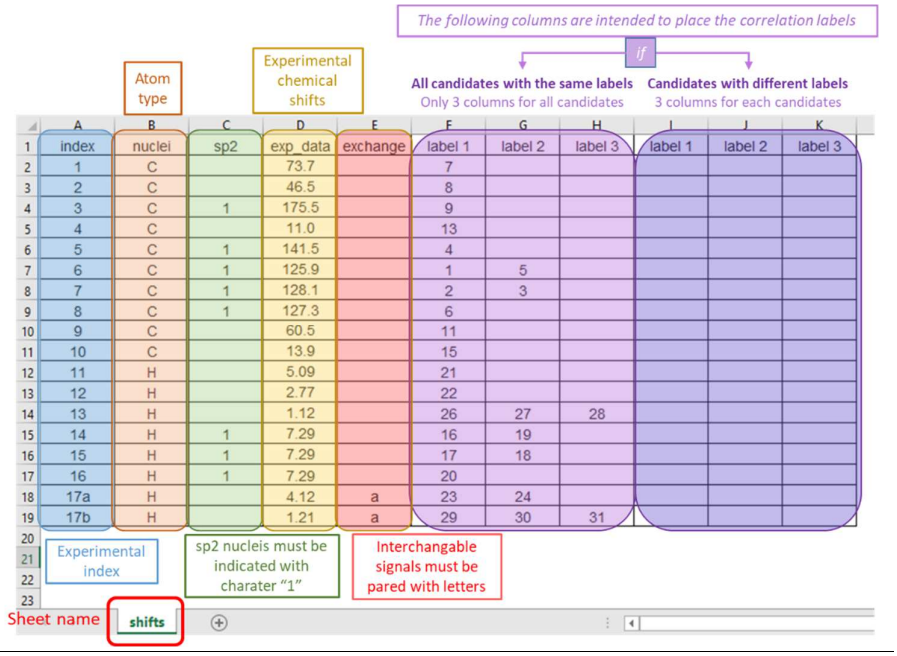
3) It is important to respect the suggested theory levels since DP4+ and MM-DP4+ were optimized for those levels. If the desired theory level is not parameterized, you can generate your own level following the instructions of the Custom-DP4+ method.
4) Using unassigned or misassigned NMR data can lead to erroneous results. The chemical shifts of equivalent nuclei that show fast interconversion should be averaged (such as the case of methyl groups, or some methylene groups). Treating the signal of each proton independently is wrong (for example, computing three different chemical shifts for the same methyl group). Another problem arises when dealing with diastereotopic methylene protons, which are often arbitrarily correlated. Unless the discrimination of both signals as pro-R and pro-S is made using additional NMR information (such as NOE or J coupling), the most convenient way to tackle this issue is to treat them as interchangeable signals. Follow the instructions to learn how to deal with these issues.

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

Prepare your files

To run a correlation calculus, DP4+ App needs the selection of a working directory and an Excel file. The program has a series of controls to ensure correct data entry.

The Excel file (.xlsx) must contain the information in the "*shifts*" sheet. This will be the only sheet read by the program and must have the structure defined in Figure 1 (see *Warnings and Input control)*. The column headers must be the same. For isomers with the same labels, only three columns need to be used. 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 requirement, since it will be selected individually.



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

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

The Gaussian files must be "*nmr*" calculations results from Gaussian software (.log or .out). The labeling must be as follows: *n_m_*_nmr.log*, where *n* is the isomer id, *m* is the conformer number, and * is a user annotation.

The selection will be made by pressing the "**Select . . .**" buttons via popup windows. We strongly recommend using the given example ("**Create Example**" button) as a template to build your working directory.

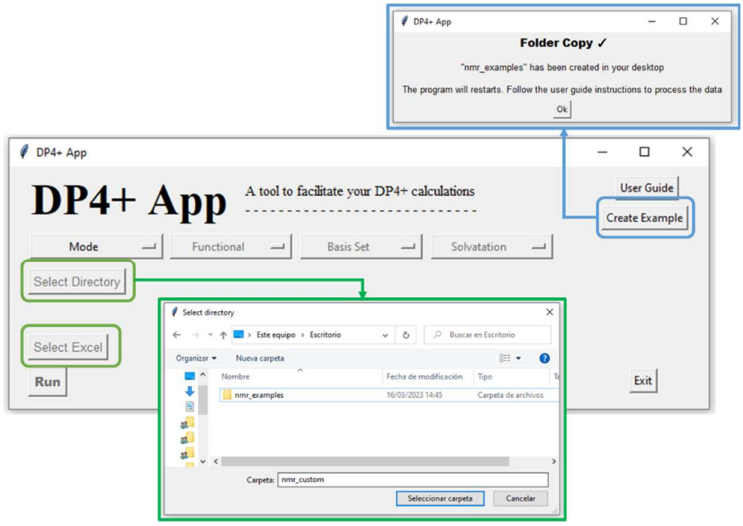


Figure 2. Entry buttons and example button

Perform a calculation

With DP4+ App it is possible to determine the probability of correlation at 60 already parameterized theory levels. They arise from the combination of various functionals, basis sets, and solvation modes. Of the total, 24 levels were parameterized from geometries optimized with quantum mechanics at the B3LYP/6-31G* (*QM mode*) and the remaining 36, through molecular mechanics with 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_3Cl)
	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_3Cl)
mPW1PW91	6-311+G(d,p)	SMD (CH_3Cl)
wB97XD		

Although it is allowed to carry out calculations at any selected level, for a better use of DP4+ App, the program controls the coincidence between the command lines of the Gaussian files with the selected theory level. In case the levels are not matched a warning will pop up, but it will not prevent the calculus execution (figure 3).

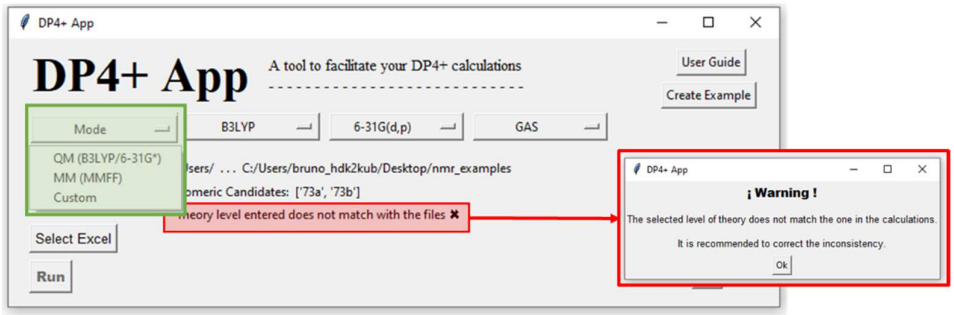


Figure 3. Modes selection and example of miss matching theory level and command lines

To perform calculations with a different theory level than those mentioned you must first parameterize it following the instructions of the *Custom mode* (next section). For this mode, the matching of the command lines is not checked automatically, but it is possible to do it yourself on the final results sheet.

Results output

A pop-up will indicate the calculation has finished correctly and the results will be presented in an Excel file inside the selected working folder. The name of the output will correspond with the calculation mode used.

There will be five sheets, one with the probability results, two with the chemical shifts and two with the correlation errors. In the main sheet ("**results**"), you will find the probabilities of the candidates classified by their nuclei, scaling and the full version. In addition, the selected theory level, the command line of the Gaussian calculations and the automatic coincidence check are printed.

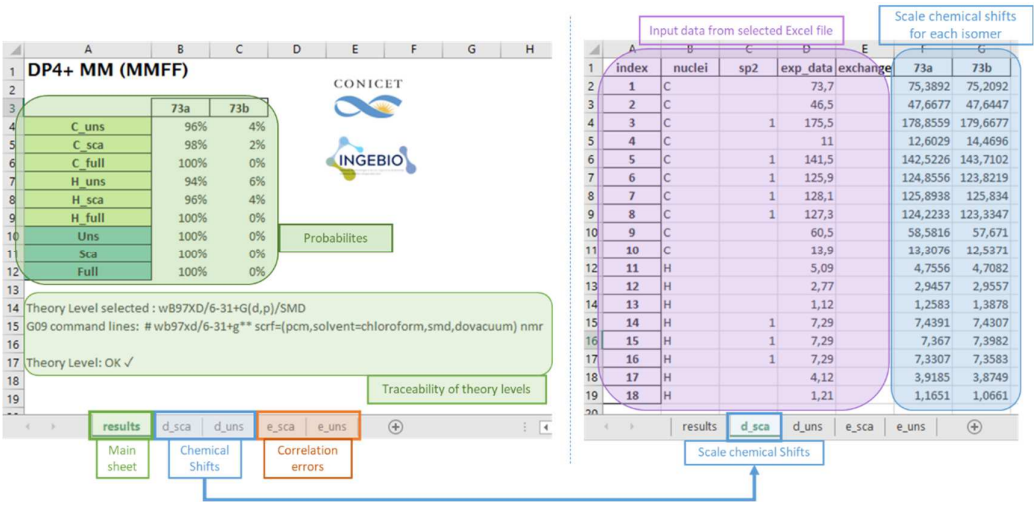


Figure 4. Output Excel file

In cases where the selected theory level does not coincide with the Gaussian calculation command line, it will warn about the misuse of the tool and the inconsistencies found (figure 5).

In Custom mode, it will be indicated that the theory level cannot be verified, and it becomes the user's responsibility to perform this verification. Additionally, below the results, the database of the Custom level used will be provided, including the standard tensors, distribution parameters, date, and parametrization method.

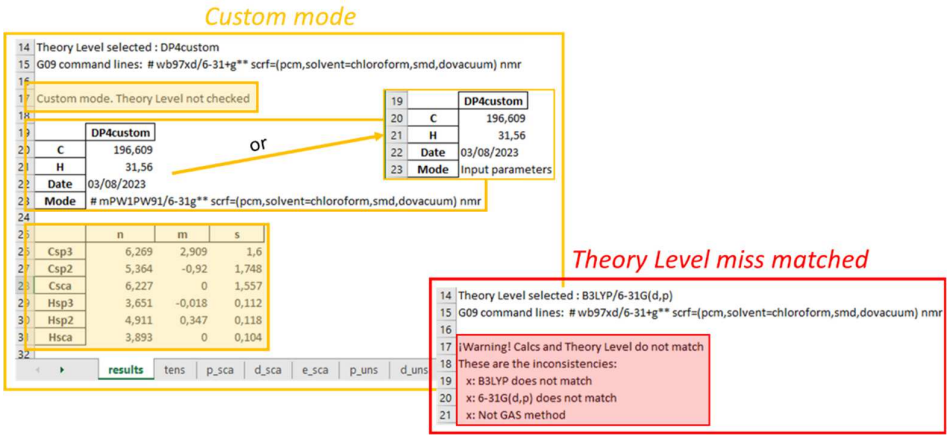


Figure 5. Examples of Custom mode traceability and theory level miss selection

Reparametrization: Custom-DP4+

Create a new level

Within the *Custom mode* of the main window, there is the "**+ new**" option that will redirect you to the reparameterization module. There you have to select the parameterization mode and assign a name to your custom level. The name can only have numbers and lowercase letters (do not use special characters either).

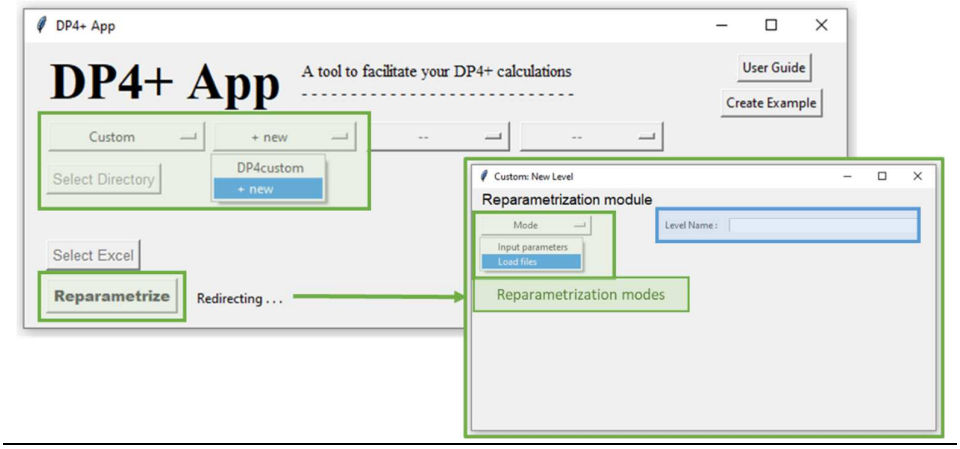


Figure 6. Selection to create a new custom level and redirecting window.

In the case you have already parameterized a theory level, you can enter the distribution parameters and tensors of the TMS by keyboard (Figure 7.a).

However, if you need to calculate distribution statistics, you will need to load a working directory and the corresponding Excel file in the main window. The working folder should contain the NMR calculations for all labeled conformations in the format *'n_m_nmr.log'*, and TMS 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.

Based on the findings presented in J. Org. Chem. 2021, 86, 12, 8544–8548, it is recommended to use a set of 8 molecules in Table 1 for parametrizing your theory level, as they have been tested and proven to estimate the distribution parameters accurately. To facilitate their preparation, you can download the input files for Gaussian calculations in both MM (force field MMFF) and QM (theory level B3LYP) optimization.

(Figure 7.b). In the window, you can type the command line that should specify the theory level as specified by Gaussian, along with the 'nmr' calculation instruction.

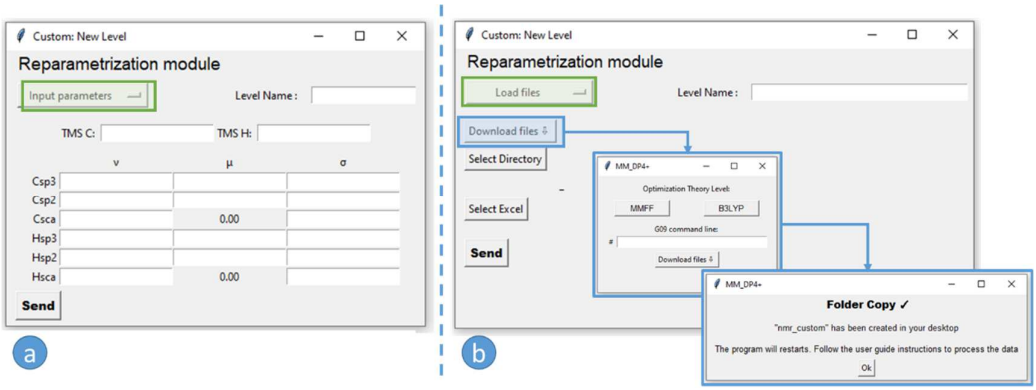


Figure 7. a) Input parameters mode: allow to entry the values of an already parametrized level. b) Load mode: capable of automatically determine the distribution parameters using the training set calculus. Also, offers templates of the training set in QM y MM optimization.

The Excel file with the experimental data and the correlation labels already assigned will be provided with the Gaussian inputs. In it, the data set of each molecule is placed in a sheet with the associated ID as mentioned before. The information follows the same structure as the correlation "**shifts**" sheet in figure 1.

Table 1. Training set molecules with experimental labels.

As mentioned by Sarotti (2021), it is important 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 about the potential inaccuracies in fitting a t-Student distribution. This notification specifically applies to the recommended set mentioned in Table 1.

You have two options: you can either utilize the average values from the original publication, which have been proven to yield accurate results for DP4+ type calculations or retain the actual values. If you choose to use the estimated values, it is strongly recommended to verify that the degrees of freedom (v) in your calculations are less than 30.

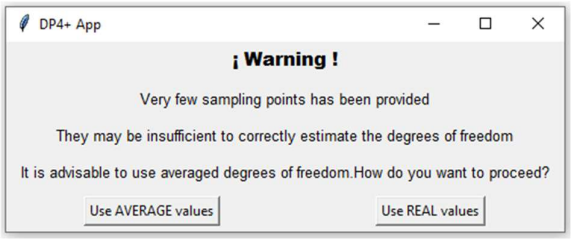


Figure 8. Warning popup for insufficient parametrization set

Update a level

To update a level simply follow the steps in the previous section and overwrite the name of the desire custom level. A popup will warn you about the existence of that level before updating the data. You can go back and change the name if you want to keep both parametrizations (figure 8). The update can be generated in any mode, it is not necessary to use the same as the one before.

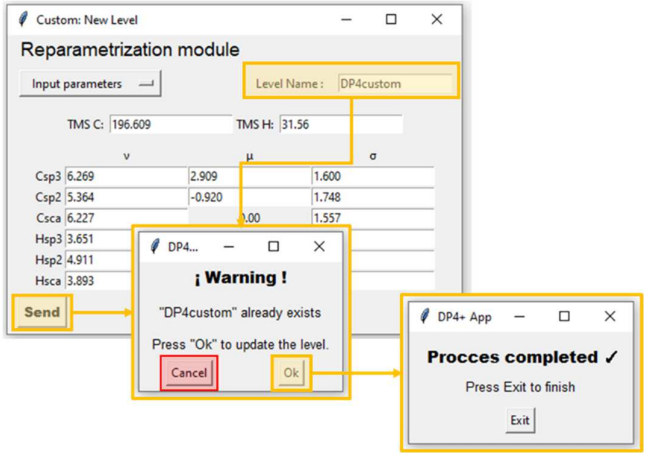


Figure 8. Example of update/overwriting a custom level

Warnings and Input control

To enhance the user's understanding of anomalous results in DP4+ type calculations, a warning system has been implemented. This system assists in interpreting and identifying any unusual outcomes. Additionally, DP4+ App includes multiple checkpoints to validate the accuracy of data entry. If any discrepancies or inconsistencies are detected that do not meet the program's requirements, it will promptly notify the user.

Questionable values

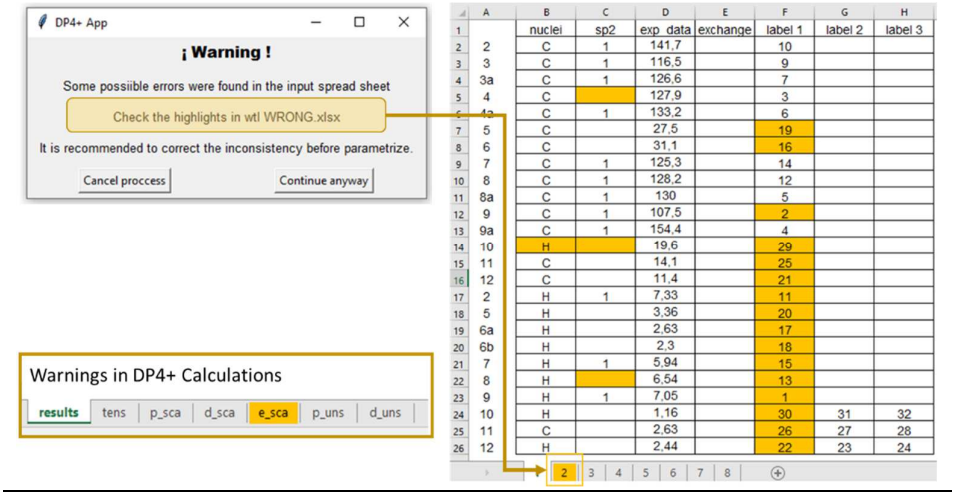
Although DP4+ App can perform calculations with any given input (which follows the requested format), there are some chemical criteria that must be taken into account for a result to be valid.

Those calculations that do not meet the following requirements will be warned about said deviations:

- σ_H > 6ppm and σ_C > 120ppm, 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

The calculations can proceed as usual, and any warnings will be easily identifiable as highlighted cells. In the case of DP4+ type calculations, the warnings will be printed on the *e_sca* sheet within the results. During

the parameterization process, a popup will prompt you to confirm whether you wish to proceed, and the highlights will be displayed in the input spreadsheet.



	nuclei	sp2	exp_data	exchange	label 1	label 2	label 3
2	C	1	141,7		10		
3	C	1	116,5		9		
3a	C	1	126,6		7		
4	C		127,9		3		
4a	C	1	133,2		6		
5	C		27,5		19		
6	C		31,1		16		
7	C	1	125,3		14		
8	C	1	128,2		12		
8a	C	1	130		5		
9	C	1	107,5		2		
9a	C	1	154,4		4		
10	H		19,6		29		
11	C		14,1		25		
12	C		11,4		21		
2	H	1	7,33		11		
5	H		3,36		20		
6a	H		2,63		17		
6b	H		2,3		18		
7	H	1	5,94		15		
8	H		6,54		13		
9	H	1	7,05		1		
10	H		1,16		30	31	32
11	C		2,63		26	27	28
12	H		2,44		22	23	24

Figure 9. Example of deviant values for parametrization method

Gaussian calculation files

In order to ensure the completeness of information provided by the Gaussian outputs, it is important to verify that the last line of each file indicates 'Normal Termination'. Any files where this string cannot be found will be automatically separated into a folder labeled 'Removed Files'. A popup will then allow you to choose whether you want to proceed with the calculation without those files or cancel to initiate a recalculation.

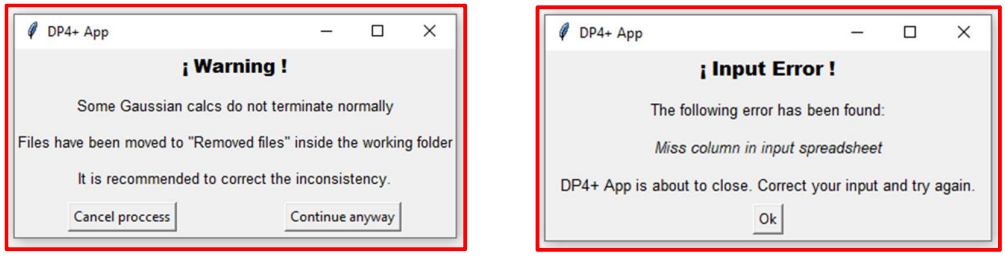


Figure 10. Examples of warning and input error

Data spreadsheet

The provided spreadsheet must adhere to the format shown in Figure 1, as mentioned earlier. The following checkpoints are in place for this file:

- Column not found: If a header is missing or not found.
- Data not found: If there is missing data in the 'nuclei', 'exp_data', or 'labels' columns.
- Incorrect data:
  - 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 'fos sp2' column, the data must be 'X', 'x', or '1'.
- Label index out of range: In cases where the label does not match any nuclei in the Gaussian calculation matrix.

- A different number of candidate isomers and set of labels: If there is a mismatch between the number of candidate isomers and the set of labels used.
- Mismatched diastereotopic labels: When the diastereotopic labels are not paired correctly.

If any of these situations occur, the program will be unable to perform the calculation. In such cases, you will need to correct the inconsistency before proceeding.

Malfunctions report

If you find a faulty operation of DP4+ App, please report your situation in detail to the following emails:

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