User guide for Solvent DP4+

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

sarotti@iquir-conicet.gov.ar

Instructive, general recommendations and case study

Workflow and general recommendations

Step 1: Although the script can process any number of isomers, minimizing the number of candidates offers several advantages: it lowers the overall computational cost and decreases the likelihood that data calculated for an incorrect isomer will align better with experimental values than the correct candidate.

Step 2: A conformational search should accurately describe the conformational landscape of the system under study. Poor computational practices can adversely affect the overall results. While systematic sampling is ideal, it becomes impractical for highly flexible molecules. In such cases, stochastic searches with a sufficiently large number of steps are recommended. To ensure no potentially relevant conformations are overlooked, all conformations within a safe energy window from the global minimum should be retained. For general recommendations, see: *Nat. Prod. Rep.*, **2022**, *39*, 58.

Step 3: All conformations found in Step 2 must be fully optimized at the B3LYP/6-31G* level.

Step 4: After removing duplicates, all structures should be subjected to NMR calculations. For general applications, we recommend using PCM/mPW1PW91/6-31+G**, which is the default level in the script. If alternative methods are chosen, the statistical parameters required for DP4+ analysis must be determined and inserted (*vide infra*).

Step 5: The output files should be organized in a dedicated folder. Additionally, an Excel file containing the experimental data and corresponding labels is required. NMR data must be properly assigned, ensuring that each chemical shift is accurately correlated to the corresponding nuclei. Using unassigned or misassigned data can lead to significant errors. Chemical shifts for equivalent nuclei undergoing rapid interconversion should be averaged (e.g., methyl groups or certain methylene groups). Treating each individual proton within a group as separate entities is incorrect (e.g., assigning three distinct chemical shifts to the same methyl group). A common issue arises with diastereotopic methylene protons, which are often arbitrarily assigned. Unless their discrimination as pro-*R* and pro-*S* is supported by additional NMR data (e.g., NOE or J-coupling), the best approach is to treat these signals as interchangeable (*vide infra*).

Step 6: Run the script `dp4plus-solv.py` to perform the Solvent DP4+ calculations. The script opens a pop-up window prompting to select the folder containing the Gaussian output files (in *.log or *.out format) and the Excel input file (Figure 1). The script automatically extracts isotropic shielding values and energies from each output file, organizing them by isomer. Finally, the chemical shifts are averaged and correlated with the experimental data recorded in the user´s selected solvent using DP4+. Additionally, the script simulates solvent-induced changes in experimental chemical shifts (Δδi) using the statistical parameters corresponding to the selected solvent in the main window. The resulting probabilities (averaged across the selected iterations) are also provided, allowing for the analysis of the analyte's sensitivity to solvent changes.

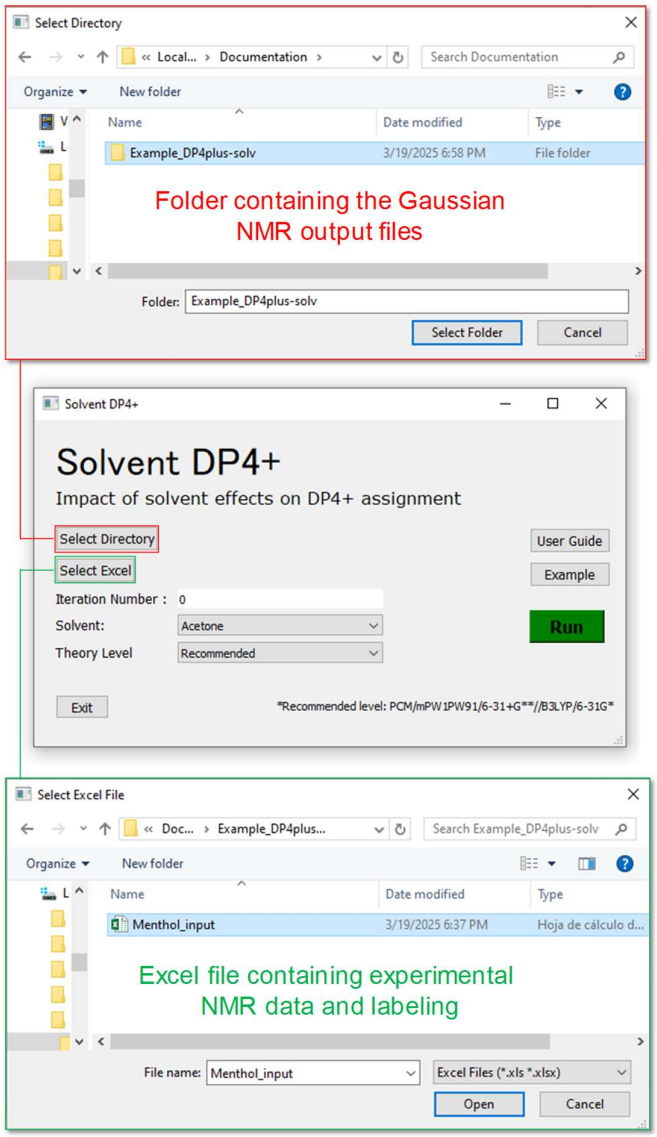

Figure 1.

Installation Requirements

Solvent DP4+ needs python 3.8 or later to work. The module can be installed by console using:

```
pip3 install DP4plus-solv
```

Once installed the python module, the program can be executed by console using:

```
DP4plus-solv
```

or generate a DP4plus-solv.py shortcut on the desktop, which allows direct execution of the program without the use of a console, using:

```
DP4plus-solv-exe
```

To verify the correct operation of the software, it is recommended to run the provided example, which can be downloaded by clicking the `Example` button (Figure 2). This will create a folder named `Example_DP4plus_solvent` on the desktop, containing all the files needed to run the example.

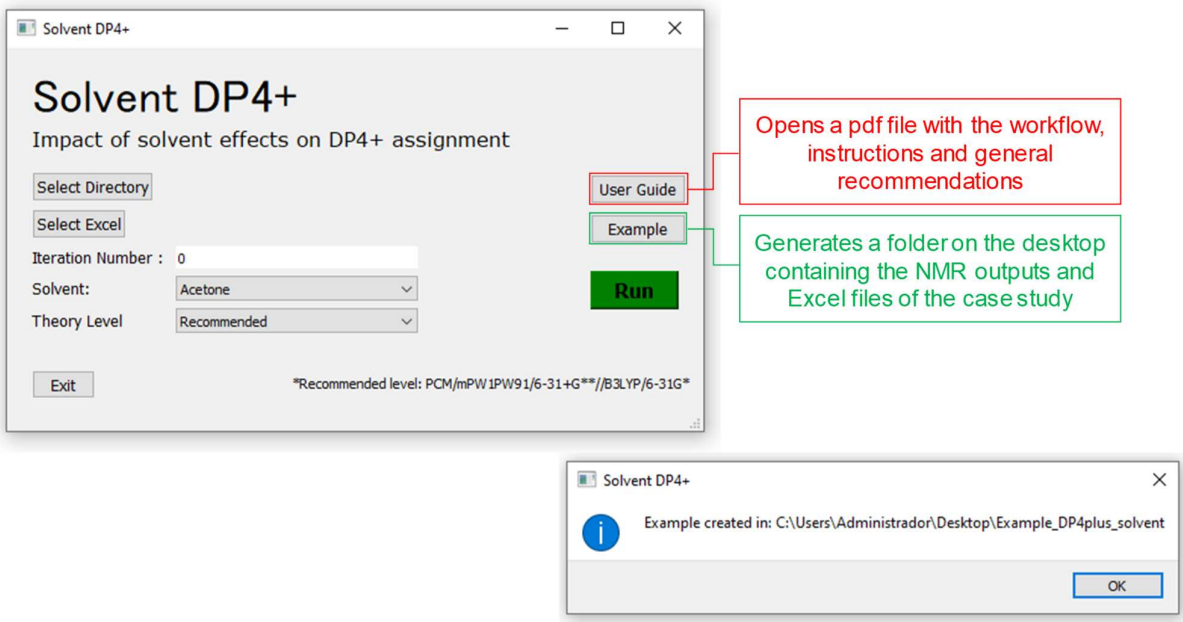

Figure 2.

Instructions for using DP4plus-solv.py

Terms of use. To run DP4plus-solv, the required information must be stored in a folder containing the following files:

- Gaussian output files for the NMR calculations (including all conformers for all isomers).
- An Excel file with the experimental data and the labels for each nucleus corresponding to each experimental value.

Technical requirements.

1) The output files: must be named following the next convention: ***n_*_m*.log*** or ***.out***, where ***n*** represent the isomer number (ranging from *1* to *N*, where *N* is the number of candidate structures), and ***m*** indicate the conformer number. For example:

1_NewNatProd_c01.log (Conformer 1 for isomer 1 of *NewNatProd*)

1_NewNatProd_c02.log (Conformer 2 for isomer 1 of *NewNatProd*)

2_NewNatProd_c01.log (Conformer 1 for isomer 2 of *NewNatProd*)

The script handles outputs from Gaussian 09 and 16.

2) The input Excel file: The experimental data and the labels of the candidate structures must be stored in an Excel file following the next rules. The Excel file must contain a single sheet named "shifts".

"shifts" sheet (Figure 3): the first column *"nuclei"* contains the identity of the atom 'c or C' for ^{13}C and 'h or H' for hydrogen atoms. The second column *"sp2"* contains the hybridization of the atoms: *0* for sp^3 C or H attached to; *1* for sp^2/sp C or H attached to). The third column *"exp_data"* contains the experimental

chemical shifts. The column *"exchange"* identifies interchangeable signals (for example, diastereotopic hydrogens). Any character can be used to indicate a pair of interchangeable signals, which will cause that the experimental and calculated values to be swapped upside-down. When dealing with more than one pair of interchangeable signals, different characters should be used to differentiate them. For example, it can be used the letter "a to indicate one pair, and the letter "b" to indicate the other pair. The following columns are intended to include the Gaussian labels of the nuclei associated to the corresponding chemical shifts. If two or more values are added in the same row, the isotropic shielding values will be averaged (as in the case of methyl groups or equivalent methylene groups). If the isomers under study have different labeling schemes (as in the case of constitutional isomers), three columns for each isomer should be provided as indicated below.

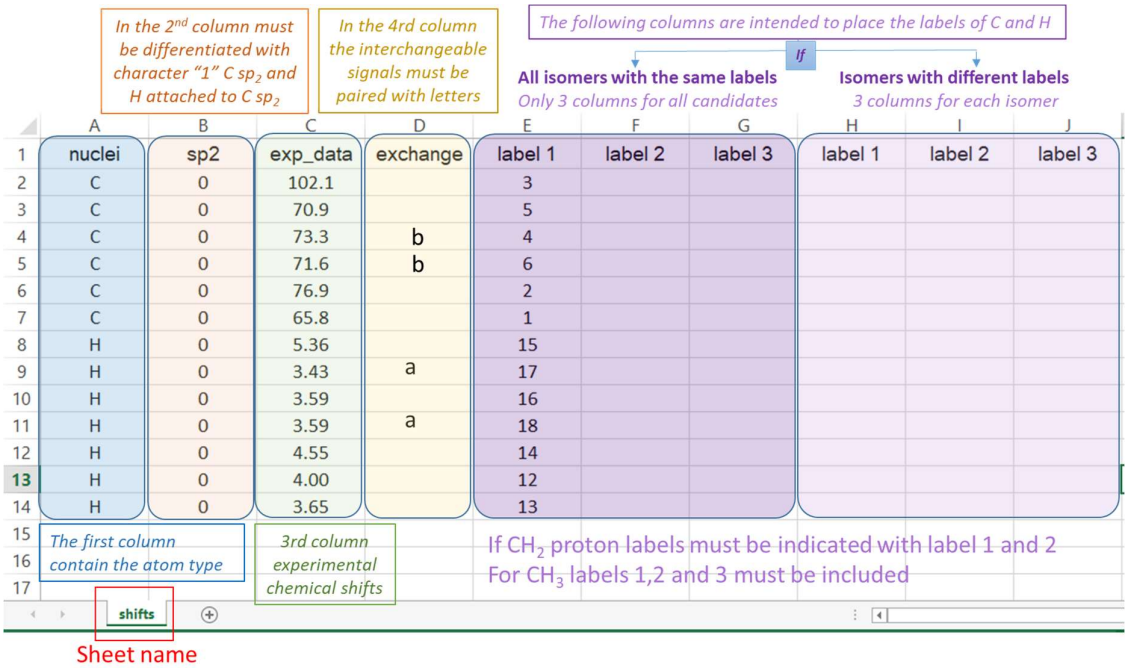

Figure 3. Instructions for proper Excel file completion.

3) The main window: In addition to the options for selecting the Directory and Excel input file, the main window of the script requires specifying the number of iterations and the solvent. Regarding the first point, although the number is variable, at least 1000 iterations are recommended to obtain statistically significant results. As for the solvent, it should match the one used for the experimental NMR measurements. It is important to note that $CDCl_3$ is not included among the options; if the experimental values are obtained in this solvent, the traditional DP4+ method should be used (see: *Nat. Prod. Rep.*, **2022**, *39*, 58). By default, the script uses the DP4+ statistical parameters obtained at the PCM/mPW1PW91/6-31+G**//B3LYP/6-31G* level, which is recommended for general applications. If the analysis is to be performed at a different level of theory, the "Other level" option should be selected (Figure 4). When the script is executed (via the Run button), a new window will open where the statistical parameters corresponding to that solvent must be completed (see: *J. Org. Chem.* **2021**, *86*, 8544 and *J. Nat. Prod.*, **2023**, *86*, 2360).

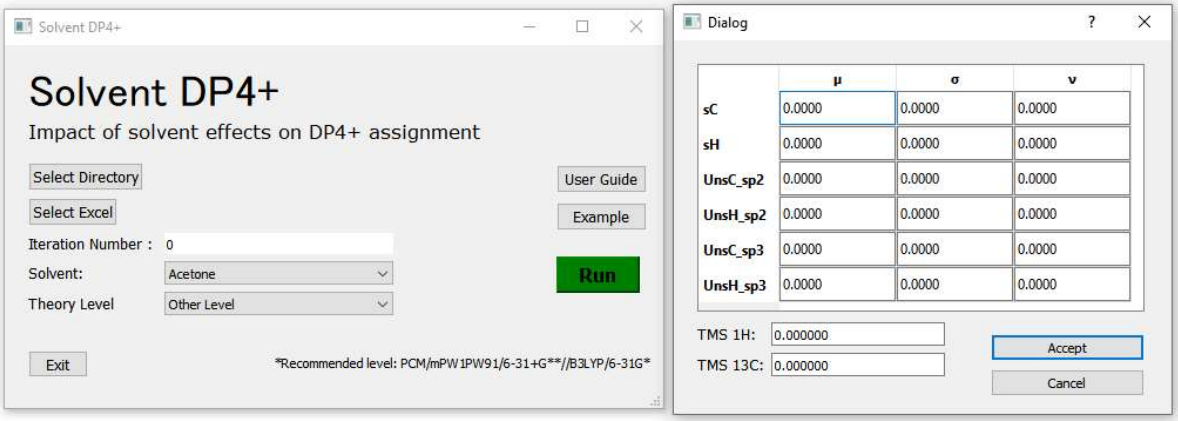

Figure 4.

4) The output Excel file: once DP4plus-solv.py is executed, a file named '*DP4+_Results.xlsx*' is created in the directory containing the Gaussian output files. The file contains 2 sheets.

DP4+ std: Displays the probabilities obtained for the different candidates using the experimental values provided in the input Excel file, following the standard DP4+ procedure (i.e., the value obtained through a traditional DP4+ analysis).

DP4+ solv: Displays the probabilities obtained (averaged across the number of selected iterations) by simulating solvent-induced changes in experimental chemical shifts (Δδi) using the statistical parameters corresponding to the selected solvent in the main window. The difference between DP4+ solv and DP4+ std shows the analyte's sensitivity to solvent changes.

Case study: menthol

To showcase the DP4plus-solv workflow, the analysis of menthol is presented using experimental data for isomer **01** collected in C_6D_6. As shown in Figure 5, four diastereoisomeric candidates are considered.

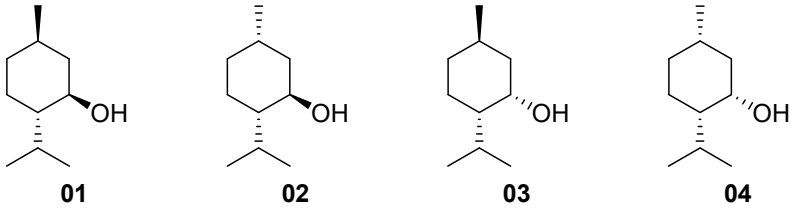

Figure 5

Following the recommended computational procedure, a total number of 53 conformers were found after the optimization at the B3LYP/6-31G* level (the standard for DP4+ calculations). Each structure was submitted to NMR and SCRF calculations at the PCM/mPW1PW91/6-31+G** level of theory, using benzene as solvent. The corresponding output files are provided in the Folder "Example". According to Gaussian numbering scheme, the labels corresponding to each nuclei are given in Figure 6.

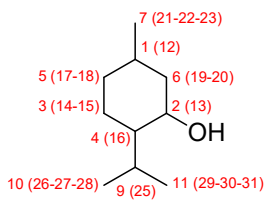

Figure 6. Carbon label followed by its corresponding proton(s) label(s) between parenthesis

With this information, the Excel input file (also provided in the "Example" folder) is completed using the chemical shift values recorded in C_6D_6 and the labels shown in Figure 6. The file should appear as shown in Figure 7.

	A	B	C	D	E	F	G
1	nuclei	sp2	exp_data	exchange	label 1	label 2	label 3
2	C	0	71.58		2		
3	C	0	50.31		4		
4	C	0	23.24		3		
5	C	0	34.68		5		
6	C	0	31.78		1		
7	C	0	45.16		6		
8	C	0	25.98		9		
9	C	0	21.15	a	10		
10	C	0	16.15	a	11		
11	C	0	22.44		7		
12	H	0	3.40		13		
13	H	0	1.09		16		
14	H	0	1.60	b	14		
15	H	0	0.97	b	15		
16	H	0	1.66	c	18		
17	H	0	0.80	c	17		
18	H	0	1.41		12		
19	H	0	1.96	d	20		
20	H	0	0.94	d	19		
21	H	0	2.16		25		
22	H	0	0.93	e	29	30	31
23	H	0	0.83	e	26	27	28
24	H	0	0.91		21	22	23

shifts

Figure 7.

With all the available information, the script is executed. Once the directory and the Excel file are selected (see previous discussion), the number of iterations and the solvent must be specified. Regarding the first point, although the number is variable, at least 1000 iterations are recommended to obtain statistically significant results. As for the solvent, it should match the one used for the experimental NMR measurements, which in this case is C_6D_6. Once the information is loaded, the program is run by pressing the Run button. The process takes a few seconds to minutes, depending on the number of iterations. When completed, an output file is generated in the selected directory.

The file displays the probability values calculated for each isomer using the experimental values obtained in the selected solvent (DP4+ std) and those obtained by applying solvent-induced changes stochastically (DP4+ solv), Figure 8. The difference between DP4+ solv and DP4+ std reflects the analyte's sensitivity to solvent changes. In this case, the sensitivity is low.

	A	B	C	D	E	F
1		Isomer 01	Isomer 02	Isomer 03	Isomer 04	
2	sDP4_H	100	0	0	0	
3	sDP4_C	99.93	0.07	0	0	
4	sDP4_Full	100	0	0	0	
5	UsDP4_H	100	0	0	0	
6	UsDP4_C	100	0	0	0	
7	UsDP4_Full	100	0	0	0	
8	DP4_H	100	0	0	0	
9	DP4_C	100	0	0	0	
10	DP4_Full	100	0	0	0	
11						
12						

DP4+ std | DP4+ solv

	A	B	C	D	E	F
1		Isomer 01	Isomer 02	Isomer 03	Isomer 04	
2	sDP4_H	98.99	0.09	0.91	0.02	
3	sDP4_C	99.87	0.12	0	0	
4	sDP4_Full	100	0	0	0	
5	UsDP4_H	91.66	3.05	3.14	2.15	
6	UsDP4_C	100	0	0	0	
7	UsDP4_Full	100	0	0	0	
8	DP4_H	98.77	0.2	1.02	0.01	
9	DP4_C	100	0	0	0	
10	DP4_Full	100	0	0	0	
11						
12						

DP4+ std | DP4+ solv

Figure 8.