

DP4+ App

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

sarotti@iquir-conicet.gov.ar

Instructive and general recommendations for Custom-STX150

Content

Overview and usage recommendations.....	1
Custom-STX150	2
Step 1: Download	2
Step 2: Train.....	3
Results output	3
DP4+ calculation with Custom-STX150	5

Overview and usage recommendations

The DP4+ App is a comprehensive software designed to perform parameterized DP4+, MM-DP4+ and xTB-DP4+ calculations seamlessly. Additionally, the application includes two additional modules: Custom-DP4+, which allows parameterization of any desired level of theory using user-defined molecules and data, and Custom-STX150, which is based on a predefined set of 150 previously calculated molecules. 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+, MM-DP4+ and xTB-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-STX150

The **Custom-STX150** module offers a streamlined approach to DP4+ calculations using custom distributions trained on the **STX150 dataset**, a curated set of 150 previously calculated molecules. Unlike the Custom-DP4+ method, which requires the user to provide their own training data, Custom-STX150 relies on this predefined dataset, which can be downloaded directly from within the application. Within the Custom-STX150 section, there are two tabs:

- **Download:** Allows the user to generate Gaussian input files (.gjc) for the full STX150 dataset, which will serve as the training basis for the Custom-STX150 parameterization.
- **Train:** Enables training a theory level using the Gaussian output files obtained from the STX150 dataset.

The characteristics of each section will be described below.

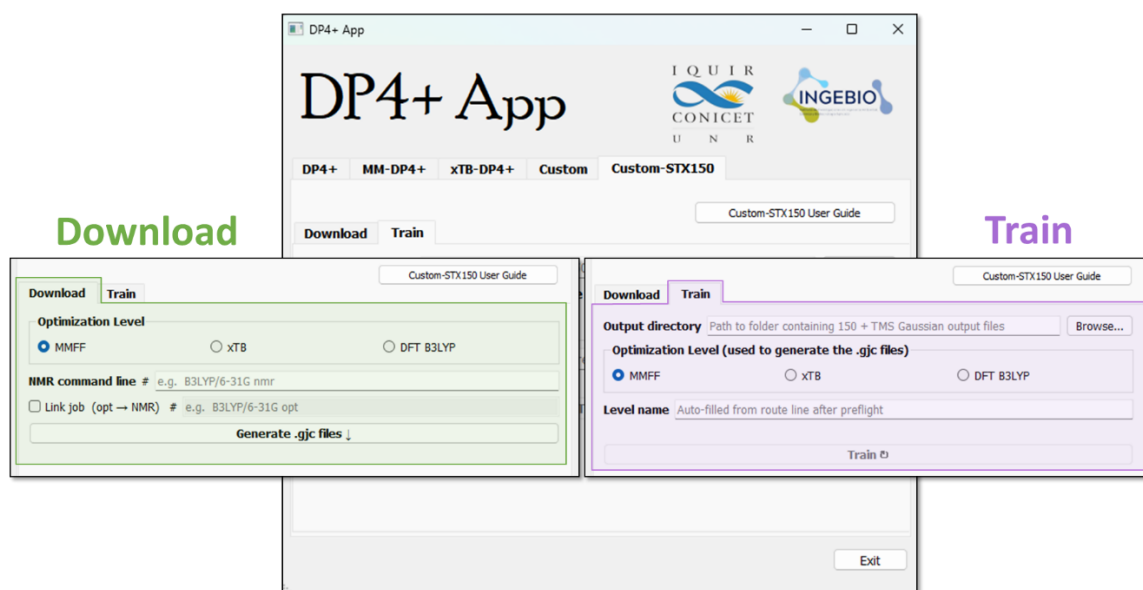


Figure 1. DP4+ Custom-STX150 module overview

Step 1: Download

The **Download tab** allows the user to generate Gaussian input files (.gjc) for the full STX150 dataset. To do so, the user must first select the desired **Optimization level** (MMFF, xTB or DFT/B3LYP), which determines the pre-optimized geometries to be used, and enter the corresponding Gaussian command line in the **NMR command line** field (e.g., `# wB97XD/6-31+G* scrf=(pcm,solvent=chloroform,smd,dovacuum) nmr`).

Optionally, the **Link job (opt → NMR)** checkbox may be enabled to reoptimize the geometries at a different level of theory prior to the NMR calculation. When selected, an additional field becomes available to

specify the desired optimization command line (e.g., *# M062X/6-31+G(d,p) opt*), resulting in a linked Gaussian job that performs both steps sequentially in a single input file.

Once all options are set, clicking the **Generate .gjc** files button will produce the input files for all 150 molecules of the STX150 dataset, along with the **TMS** reference calculation. The files are saved in a folder automatically created on the desktop, named following the convention `index_OptLevel.gjc` (e.g., `6_MMFF.gjc`, `7_MMFF.gjc`), where the index corresponds to the molecule number within the STX150 dataset and the suffix to the selected optimization level.

Step 2: Train

The **Train tab** parameterizes the desired level of theory once the Gaussian calculations for the STX150 dataset have been completed. The user must specify the folder containing the output files (including NMR results for all 150 molecules and the TMS reference calculations) either by typing the path directly in the **Output directory** field or using the **Browse...** button. The Optimization level must then be selected to match the choice made during the Download step. The Level name field is populated automatically based on the selected options. Clicking **Train** button will initiate the parameterization and display the resulting training information.

Upon selecting the optimization level, a preflight check is automatically performed to verify consistency between the selected level and the provided output files. The result is displayed as a color-coded message:

- **Green:** The selected optimization level matches the output files. The parameterization can proceed normally.
- **Yellow:** A limited number of geometry warnings were detected but are non-critical. The process will proceed with a notice (e.g., *"3 geometry warnings acknowledged – proceeding"*).
- **Red:** A significant number of molecules present geometric inconsistencies ($\text{RMSD} > 0.5 \text{ \AA}$), suggesting a mismatch between the selected optimization level and the actual calculations. The user should verify the correct level and run the preflight check again before proceeding.

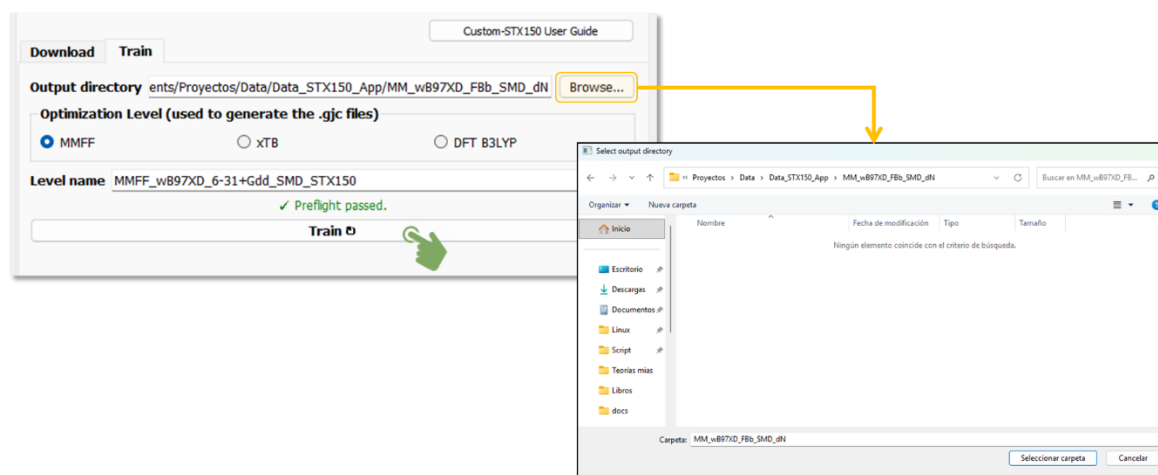


Figure 2. Train tab usage diagram

Results output

After the completion of the training, a PDF report is automatically generated and displayed in a pop-up window. The file is saved in the same folder as the molecular output files, and is named automatically based on the selected options and the current date (e.g., *MMFF_wB97XD_6-31+Gdd_SMD_STX150_report_2026-05-07.pdf*). The report is organized into the following sections:

- ❖ **Metadata:** Displays the level name (generated automatically from the selected options), the date of the training, the optimization level, the geometry match percentage obtained in the preflight check, and the full NMR command line as entered by the user in the Download tab.

- ❖ **Parametrization — TMS tensors:** Reports the TMS reference tensors for ^{13}C and ^1H derived from the Gaussian output files calculated at the user-selected level of theory. These values are used as the chemical shift reference standard for subsequent DP4+ calculations.
- ❖ **T-Student parameters:** A table listing the distribution parameters (ν , μ , σ) for each error distribution series (Csp3, Csp2, Cspa, Hsp3, Hsp2, Hspa), fitted to the user-selected level of theory.
- ❖ **Statistics:** A summary table reporting the CMAE, Max Error, and Within threshold percentage for both ^{13}C and ^1H , providing an overall assessment of the parameterization quality.
- ❖ **Spidergram:** A radar chart comparing the statistical performance of the trained level against a fixed reference level across all metrics.

An example of a completed training report is shown below.

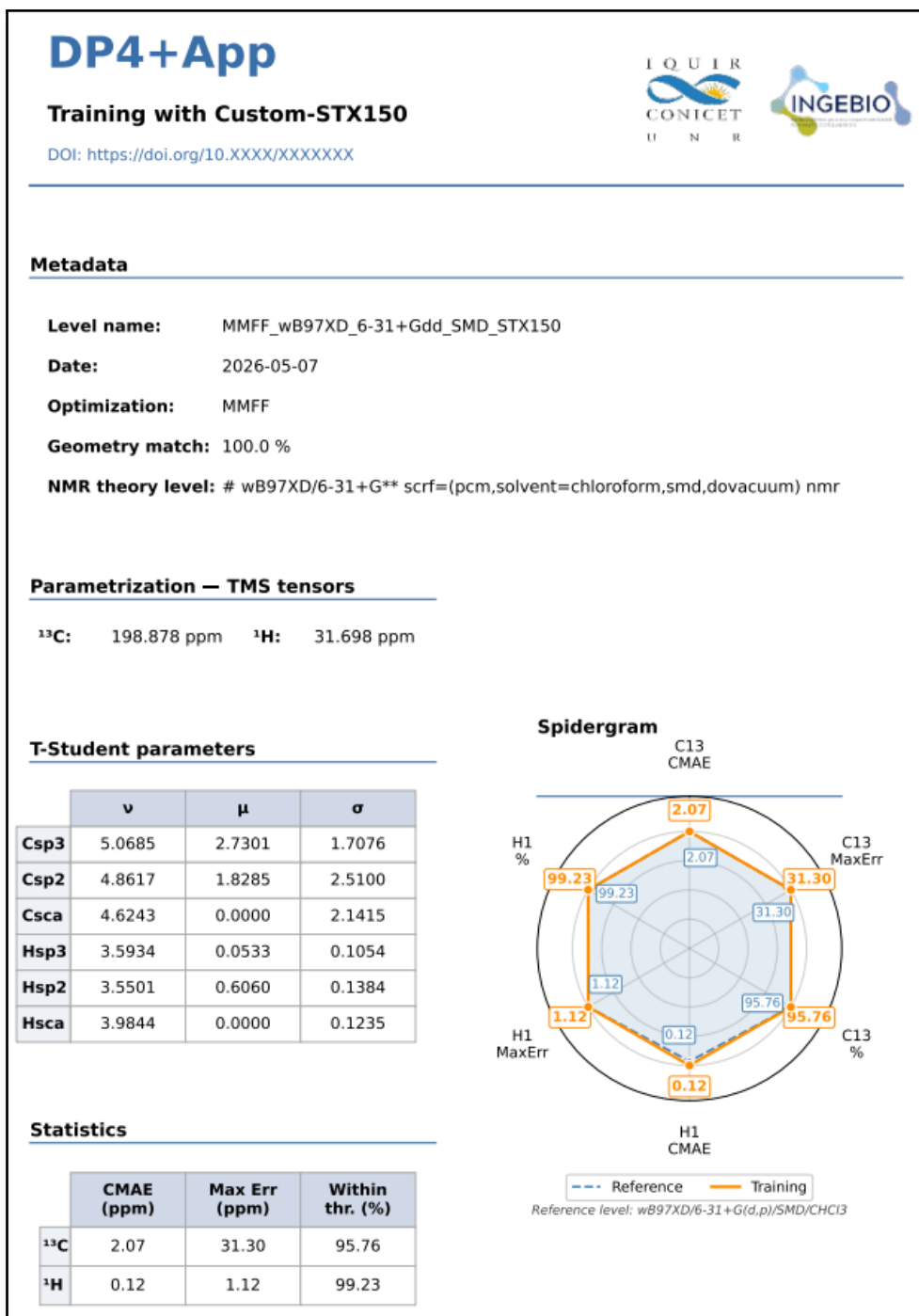


Figure 3. Representative PDF report generated upon completion of the Custom-STX150 training.

DP4+ calculation with Custom-STX150

Upon successful completion of the training, a confirmation pop-up will appear indicating that the new level has been saved to the Custom parameter bank. Clicking **OK** will automatically redirect the user to the **Custom Calc** tab, where the trained level will be readily available in the **Custom Level** dropdown menu.

To proceed with the DP4+ calculation, the working folder containing the Gaussian NMR output files for all conformers of each candidate structure must be specified, along with the correlation spreadsheet containing the experimental NMR data and the corresponding correlation labels. These are selected using the **NMR** and **Correlation** buttons respectively, each opening a navigation dialogue for folder and file selection. Once both inputs are set, clicking **Run** will initiate the DP4+ calculation using the newly trained level.

For detailed instructions on file preparation and calculation setup, refer to the Custom-DP4+ User Guide.

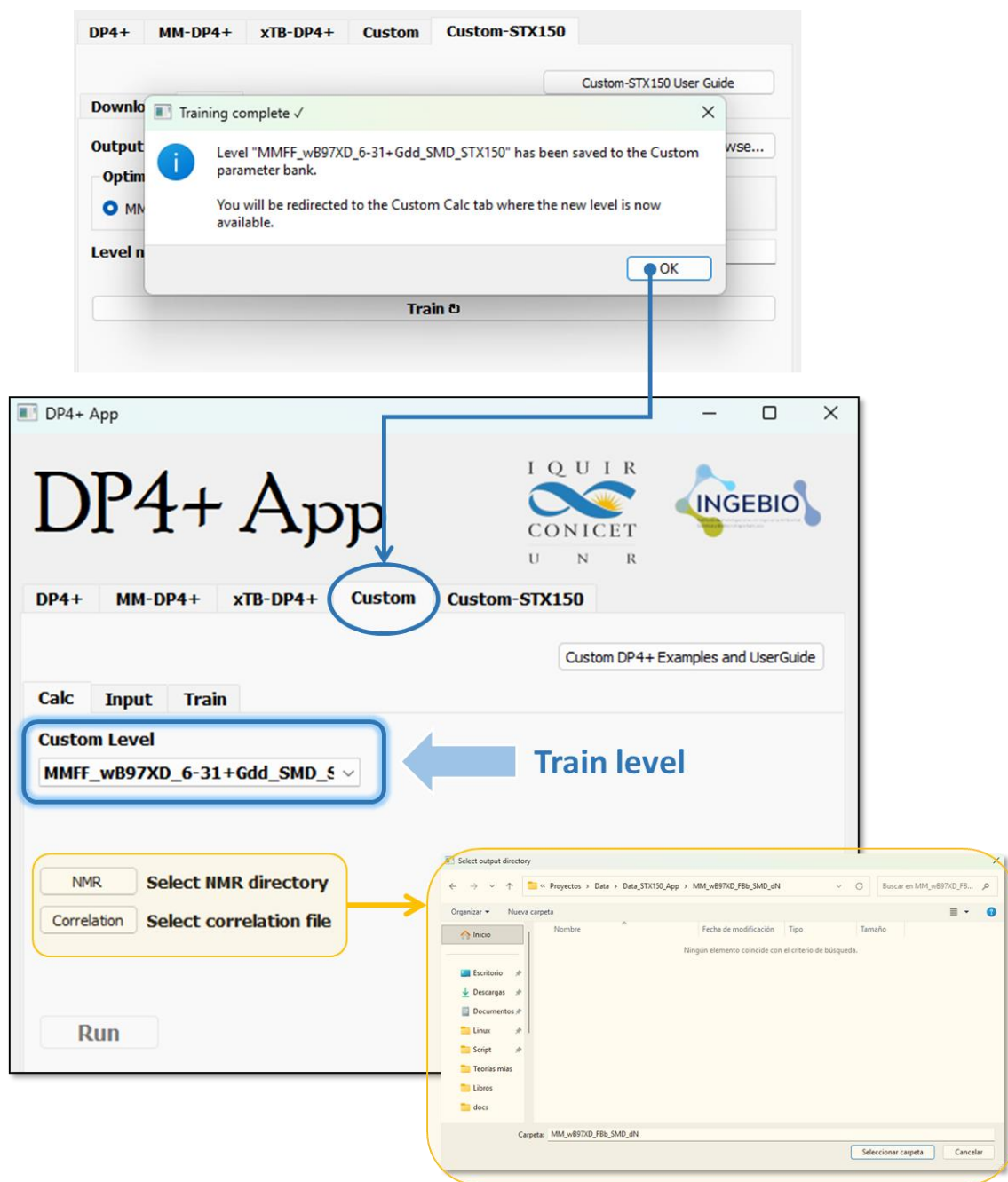


Figure 4. The application redirects to the Custom Calc tab with the trained level preloaded.