# simplePREDICT

## Introduction

The key aims behind the simpleNMR suite of utilities are:

* to identify CH(n) groups (where n = 0, 1, 2, or 3) in a set of NMR experiments on a particular sample
* identify the interactions (correlations) between those CH(n) groups using COSY and HMBC experiments
* map that information onto a postulated structural isomer in a way that makes it relatively easy to answer the question “Is this NMR data consistent with this structural isomer and can a satisfactory assignment of the NMR data be made on that basis?”
* Facilitate the assignment and generate a report in an easy way.

simplePREDICT is the main tool in the simpleNMR suite of utilities. The tool takes the peak-picked data sets, and uses predicted 13C chemical shifts for the molecule to map the NMR data onto the molecular structure using information from chemical shifts, CH(n) groups (from HSQC), and correlations (from HMBC and COSY) in an attempt to interpret the data in terms of the molecular structure provided by the user.

## Step Through Guide

### NMR Data Sets

The tool simplePREDICT requires a set of NMR experiments and a molecular structure to be present in the Mnova document. Below is a list of the datasets that can be used, with an indication of which are required and which are optional.

* **HSQC** (required)
  + Multiplicity Edited version preferred with CH3 and CH1 phased to be positive and CH2 phased to be negative.
  + Peaks picked and integrated
  + CH3 set in the annotations field of the peak table (MNOVA)
  + If the HSQC experiment is not multiplicity edited then a DEPT-135 dataset is also needed.
* **Molecular structure** diagram present in the file. (required)
* **1-D Proton Pureshift** (optional)
  + Required for simplePeakPicking tool to be used
* **1-D Proton** (optional)
  + Useful for identifying CH3 resonances that can then be used to set the corresponding annotation field in the HSQC data set
* **1-D Carbon** (optional)
  + Useful for identifying all the carbon resonances in the molecule.
  + Used by the simplePeakPicking tool
  + Useful for quarternary carbon atoms that have no HMBC couplings
* **HMBC** (optional)
  + Used to check if the resonances assigned to the molecule are geometrically correct.
  + Used to find quarternary carbons if no 1-D carbon is present.
  + Used in automatic simulated annealing algorithm to refine the prediction.
* **COSY** (optional)
  + Used to identify 1H-1H correlations between CH(n) groups. Helps to check if assignments are correct.
  + Used in automatic simulated annealing algorithm to refine the prediction.
* **HSQC-CLIP-COSY** (optional)
  + Useful to increase the dispersion of COSY information and pick more correlations with increased certainty.
* **DDEPT-CH3-Only** (optional)
  + Used to identify experimentally CH3 resonances. Very useful when the 1-D proton spectrum is difficult to analyse due to peak crowding / overlap.
* **DEPT-135** (optional)
  + Used when working with historical data that do not have multiplicity-edited HSQC data, in order to differentiate between CH2 resonances and CH1/ CH3 resonances.

## The “Ideal” data set

In view of the number of “optional” datasets listed above, a reasonable question would be “so which experiments should I be using?” The only experiment that is absolutely required is the HSQC, since this is the experiment that is used to identify the CH(n) fragments that are central to the method. It is strongly recommended to use a multiplicity edited version of the HSQC experiment to facilitate identification of CH2 groups. Beyond that, the method uses correlations between the CH(n) fragments to refine and confirm their positioning on the molecular structure. So both HMBC and some form of COSY experiment are very useful, but note that neither is an absolute requirement. It may be, for example, that some structures show hardly any useful COSY correlations, so it makes no sense to absolutely require that a COSY experiment is present but, in the general case, we would expect both COSY and HMBC to be present. We might then ask what form of COSY experiment should we choose. The tool will accept both classic 1H-1H COSY spectra and the heteronuclear HSQC-Clip-COSY. Generally, the HSQC-Clip-COSY experiment will have better peak dispersion, so would probably be favoured, but it may be that in cases where there is near accidental degeneracy in the carbon spectrum, but not in the proton, the classic 1H-1H COSY gives better results. So the user can select whichever experiment seems more appropriate. Finally among the 2D data sets, if the molecule contains a significant number of methyl groups that are overlapped with other signals so that it is not immediately apparent to the user which HSQC correlations are due to methyl groups (in some steroids, for example) the tool can make use of the DDEPT-CH3-Only experiment to identify the methyl groups, but note that this is often not needed.

Turning to the 1-D experiments, it makes no sense not to acquire a standard 1-D proton spectrum as part of the dataset. It is not generally used by the simplePREDICT tool, but is a useful reference point and takes less than a minute to acquire. It is generally good policy to acquire the whole suite of NMR experiments on a particular sample at the same time if possible. In addition, both a 1-D carbon spectrum and a pure-shift proton spectrum (PSYCHE, for example) are very useful in that they greatly simplify the process of peak picking the 2-D spectra (see the documentation for simplePeakPick) and facilitate the correct identification of **all** quaternary carbon atoms. If you are collecting NMR data with the intention of reporting the characterisation of your molecule in the literature you will most likely require a 1-D carbon spectrum anyway so it makes sense to include it in this dataset.

So, the “ideal” data set might consist of the following spectra: 1-D proton, 1-D carbon, proton PSYCHE, 1H-13C HSQC (multiplicity edited), 1H-13C HMBC, 1H-13C HSQC-Clip\_COSY. But note that the tool has been successfully used with a range of other datasets.

### The simplePREDICT Dialog

After clicking on the simplePREDICT icon, the simplePREDICT dialog window will appear.

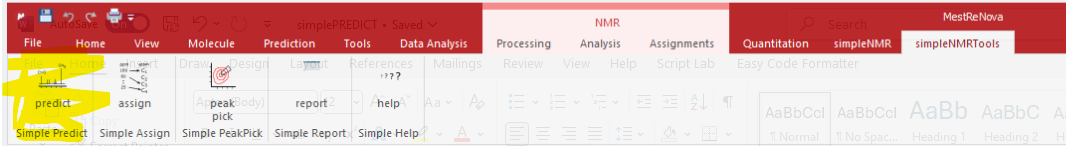


Figure 1 The simplePREDICT icon under the simpleNMRTools.

The simplePREDICT dialog controls how the tool operates. There are three main parts to the dialog.

1. NMR datasets to use in the prediction
2. Which carbon chemical shift prediction source to use
3. Optimization of results using simulated annealing on COSY and HMBC correlations

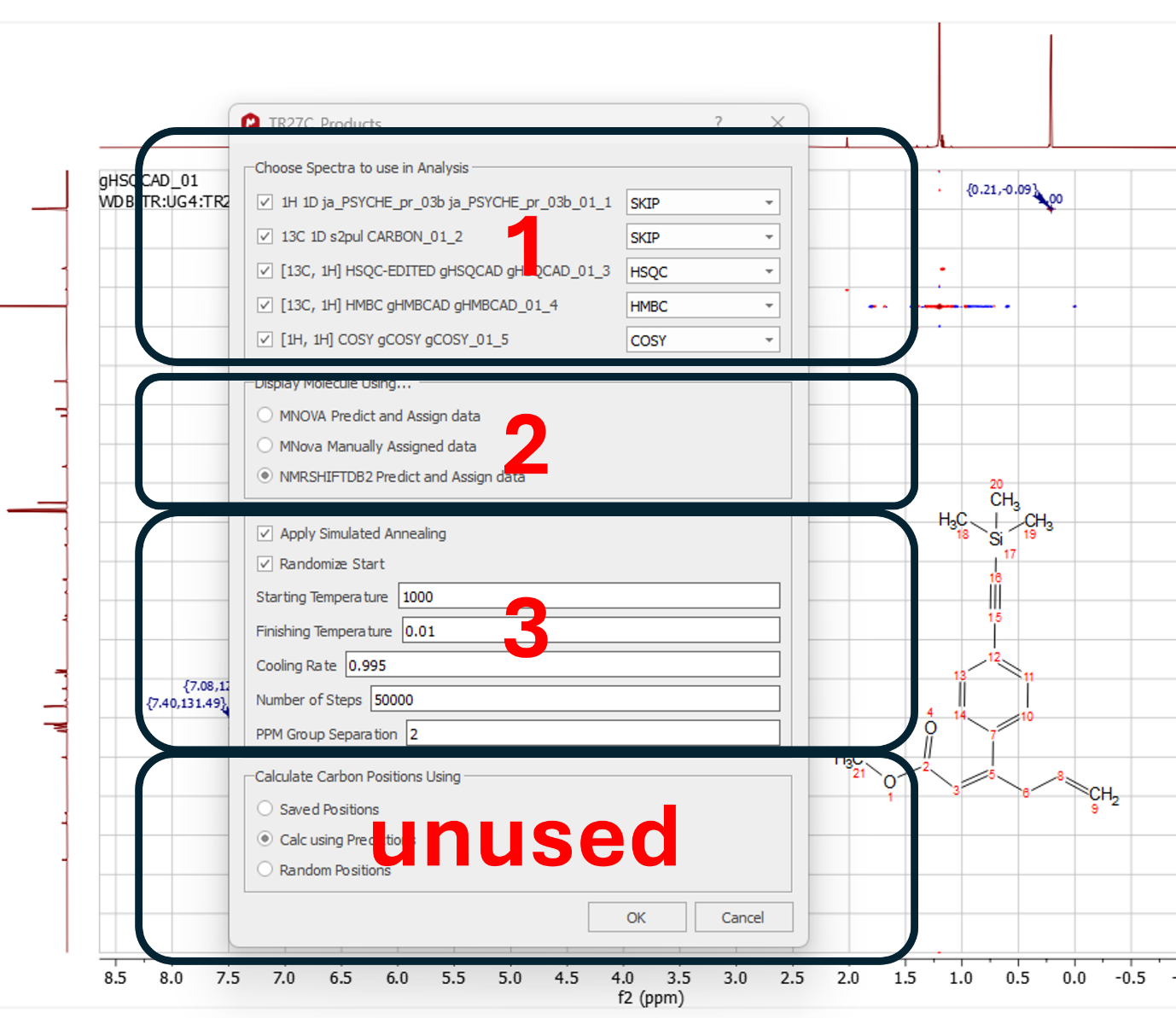


Figure 2 simplePREDICT dialog showing three distinct parts

#### 1. NMR data sets to use in the prediction

All the NMR experiments present in the MNOVA file which have been “peak-picked” will show up in this section. Initially the drop-down windows will be set to SKIP. The user is required to identify the datasets in terms of HSQC, HMBC, COSY etc manually. We have attempted in the past to automatically identify the spectra, but this has proven difficult to implement as no one procedure is able to reliably identify all of the various possible flavours of experiment from all of the possible equipment manufacturers (past and present!). Therefore, we have gone with a simple manual solution. Once the datasets have been identified, the information is stored so that the user does not have to perform the action again unless new data files have been peak picked and the simplePREDICT tool is run again.

Note that an HSQC dataset **must** be present in the list.

#### 2. Carbon Prediction software to use.

In this part of the dialog the user has three options on how the simplePREDICT tool calculates the carbon ppm values.

The first option is to use the prediction tool from MNOVA if the user has a license.

The second option can be used if the user has manually assigned the data already using the manual or semi-automatic tools from MNOVA.

The third option uses the NMRSHIFTDB hose code to predict the chemical shifts. This code is free to use, but the predictions are generally less accurate than those from the Mnova software.

#### 3. Optimization of Prediction Results using Simulated Annealing.

Originally, the simplePREDICT tool matched the predicted carbon chemical shifts to the experimental data by grouping the carbon ppm values into categories based on the number of protons attached to the carbon. Then matching the calculated chemical shifts to the experimental chemical shifts in each category. The user then had to use the graphical user display to check whether the HMBC and COSY correlations looked reasonable and re-arrange the carbon atoms over the molecular structure until the graph representation of the interactions made sense.

We have attempted to automate this step by implementing a simple simulated annealing optimization algorithm to minimize the HMBC and COSY correlations over the graph representation of the molecule that the user thinks they have made.

The default parameters for the simulated annealing algorithm are usually good enough to find a good optimum, but on occasion they may have to be changed if the molecule is large and the HMBC and / or COSY correlations are sparse.

Typically, if the default parameters prove to be inadequate, the user can try some combination of increasing the “Starting Temperature”, reducing the “Finishing Temperature”, and increasing the “Cooling Rate” (note that the parameter given as “Cooling Rate” is actually the fractional decrease in temperature per step and is therefore the inverse of a rate – the higher the value of this parameter, the slower the rate of cooling), but be aware that any of these actions is liable to increase the execution time.

## Errors and Problems

The error reporting with simplePREDICT is not very informative at present and we are working to improve this. The tool catches a number of simple errors and reports them via MNOVA warning dialogs or via html output if the tool has reached that stage.

The simple errors include the following:

* Absence of a molecule diagram in the MNOVA data.
* Missing HSQC dataset.

The simplePREDICT program attempts to match up the number of carbon groups in the molecule (CH3, CH2, CH1, C) with those found in the experimental NMR data via the HSQC information, proton integrals (if used) and CH3 only NMR data if present.

On many occasions the molecule will have NMR symmetry present and therefore the steps taken to decide if the number of carbons groups match up with the experimental information is quite complex. Unfortunately, when things don’t match up the error messages reported are quite cryptic and not very helpful to the novice (or, indeed, the experienced user!).

Typically, the error message will be something like len(CH0) > len(CH0) 6>5. This means there are more experimental quaternary carbons present than expected in the molecule structure provided.

In such cases, the user then has to resort to looking at the HSQC and 1-D carbon experimental data to see if a peak has been picked erroneously or is missing.

These types of errors may occur if the carbon chemical shift separation is very small for a couple of carbon resonances. This type of error is difficult to overcome as the user does not have access to this adjustable tolerance.

A second reason for errors based on mismatched number of carbons in a certain group is when the HSQC data has only been peak picked and not also integrated and there are clear doublets in the proton dimension of the HSQC, corresponding to, for example, a CH2 group. Typically, these peaks should have a negative intensity, but if the position of the peak is picked in the centre of the doublet the intensity maybe 0 or even positive. If the integral is measured in addition to peak picking this usually integrates to be negative and so the peak will be correctly recognised as a CH2 group and the number of carbons in all the other groups will be counted correctly.