SUPPORTING INFORMATION

A Database for Rapid Dereplication of Known Natural Products Using Data from MS and Fast NMR Experiments

Carlos L. Zani[†]* and Anthony R. Carroll*[‡]

[†]Centro de Pesquisa René Rachou – Fiocruz. Av. Augusto de Lima, 1715, Belo Horizonte, 30190-002, MG, Brazil.

[‡]Griffith School of Environment, Griffith University, Gold Coast Campus, QLD 4222, Australia

*zani@cpqrr.fiocruz.br, +55 31 3349 7791 (tel) +55 31 32953115 (fax)

Contents

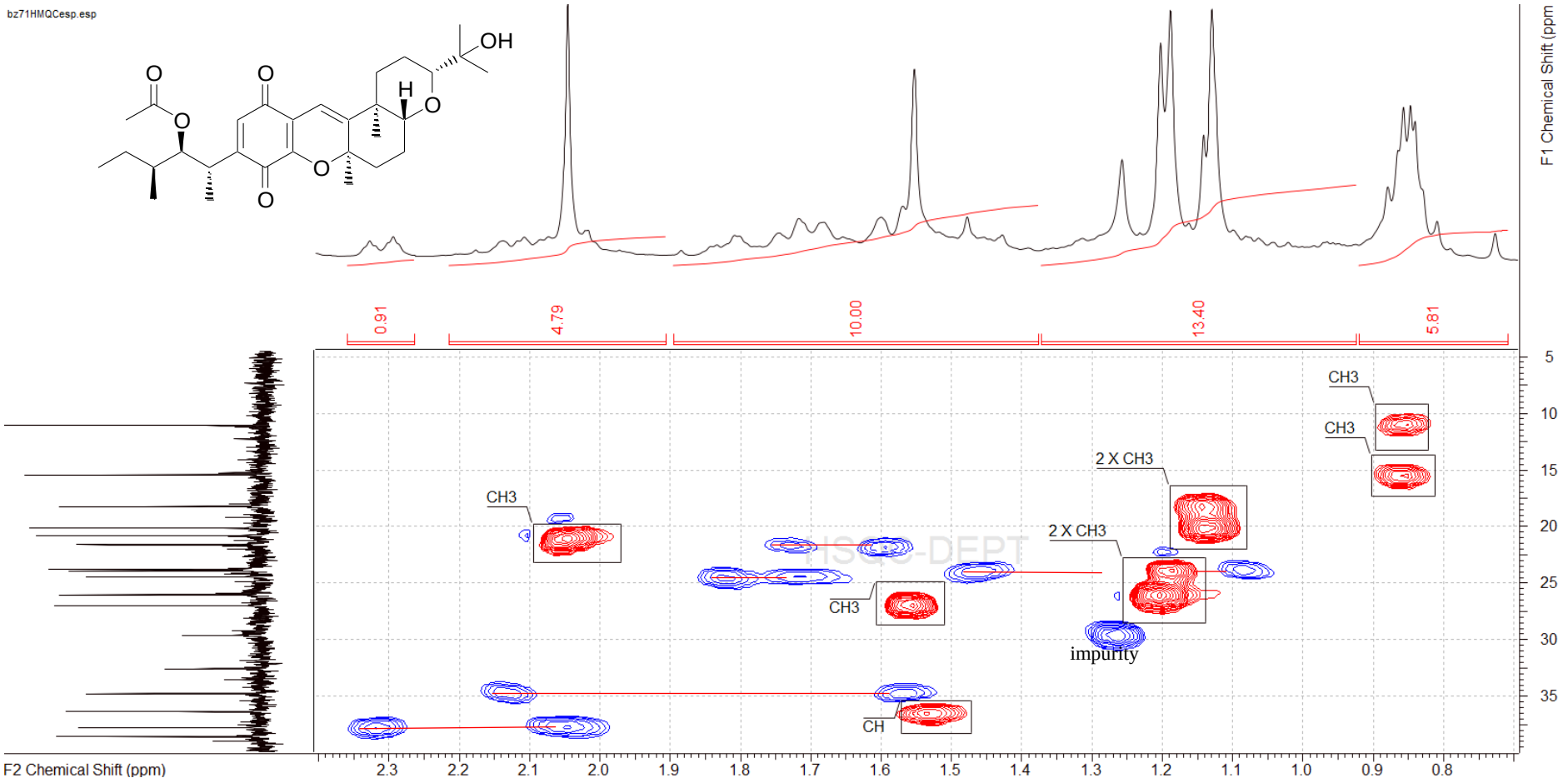

Figure S1. High field region of the HSQC contour map of compound **1** (400 MHz, $CDCl_3$).

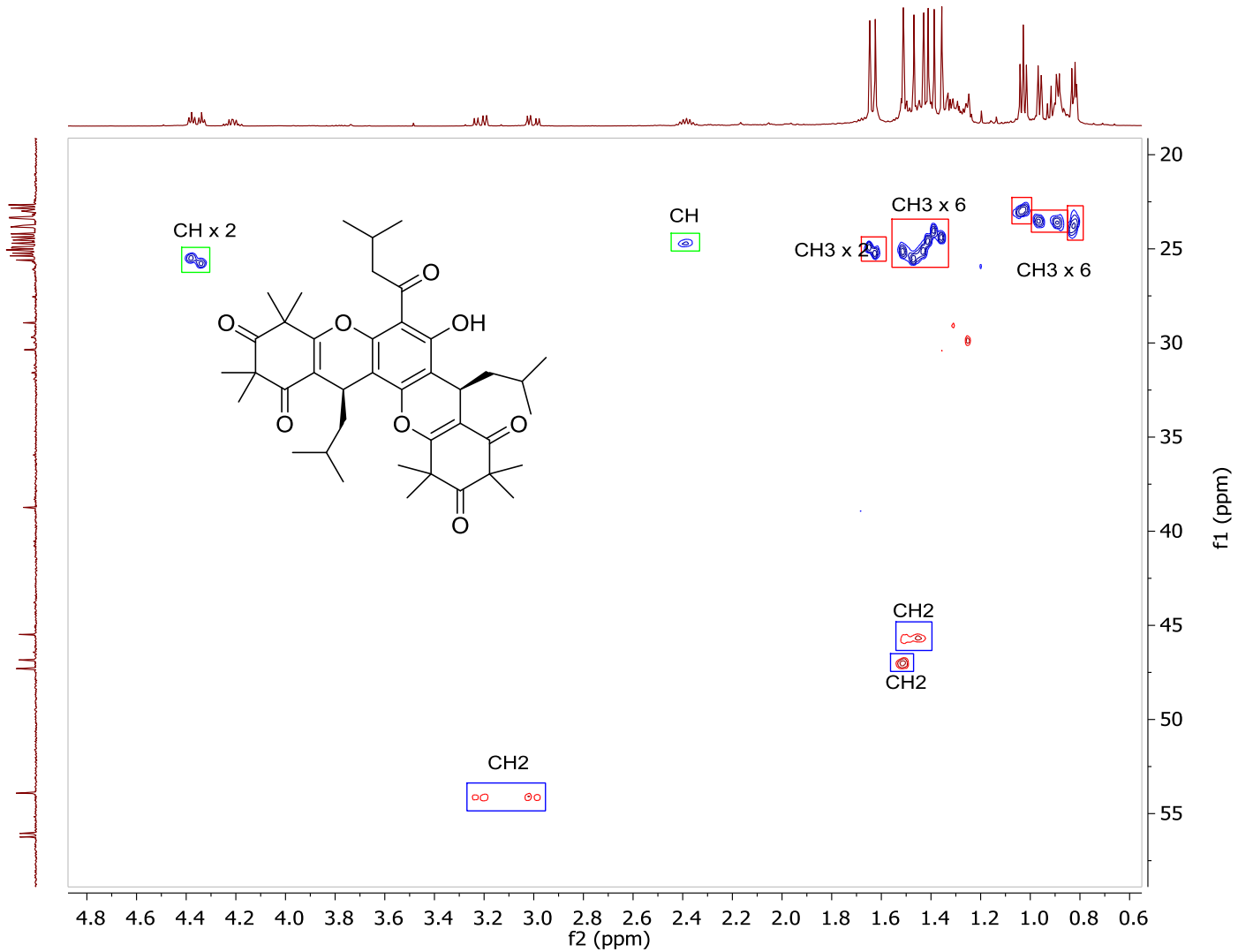

Figure S2. Edited HSQC spectrum of **2** (rhodomyrtosone C) (500 MHz, $CDCl_3$).

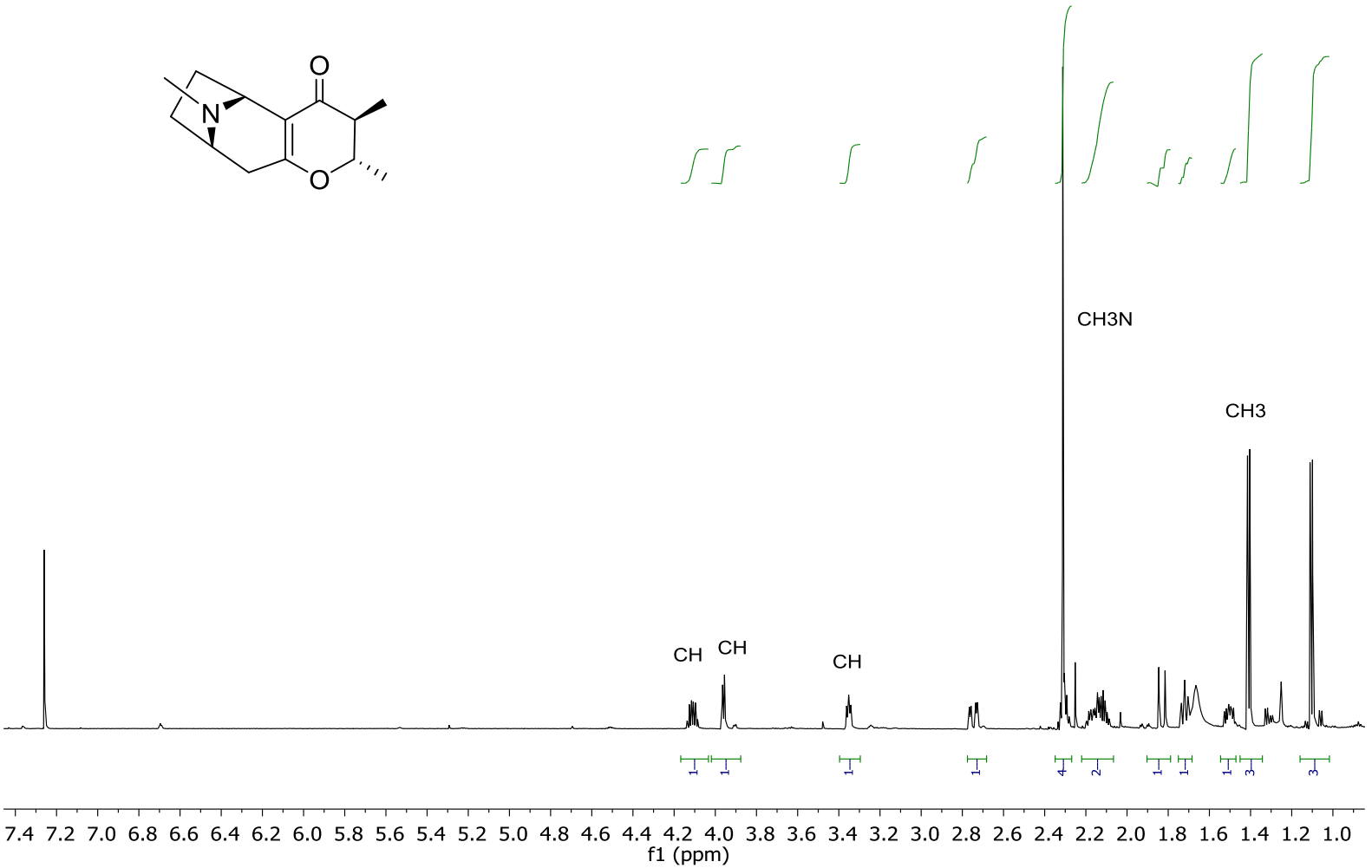

Figure S3. ^{1}H NMR spectrum of **3** (2,3-dihydrodarlingine) (600 MHz, $CDCl_3$).

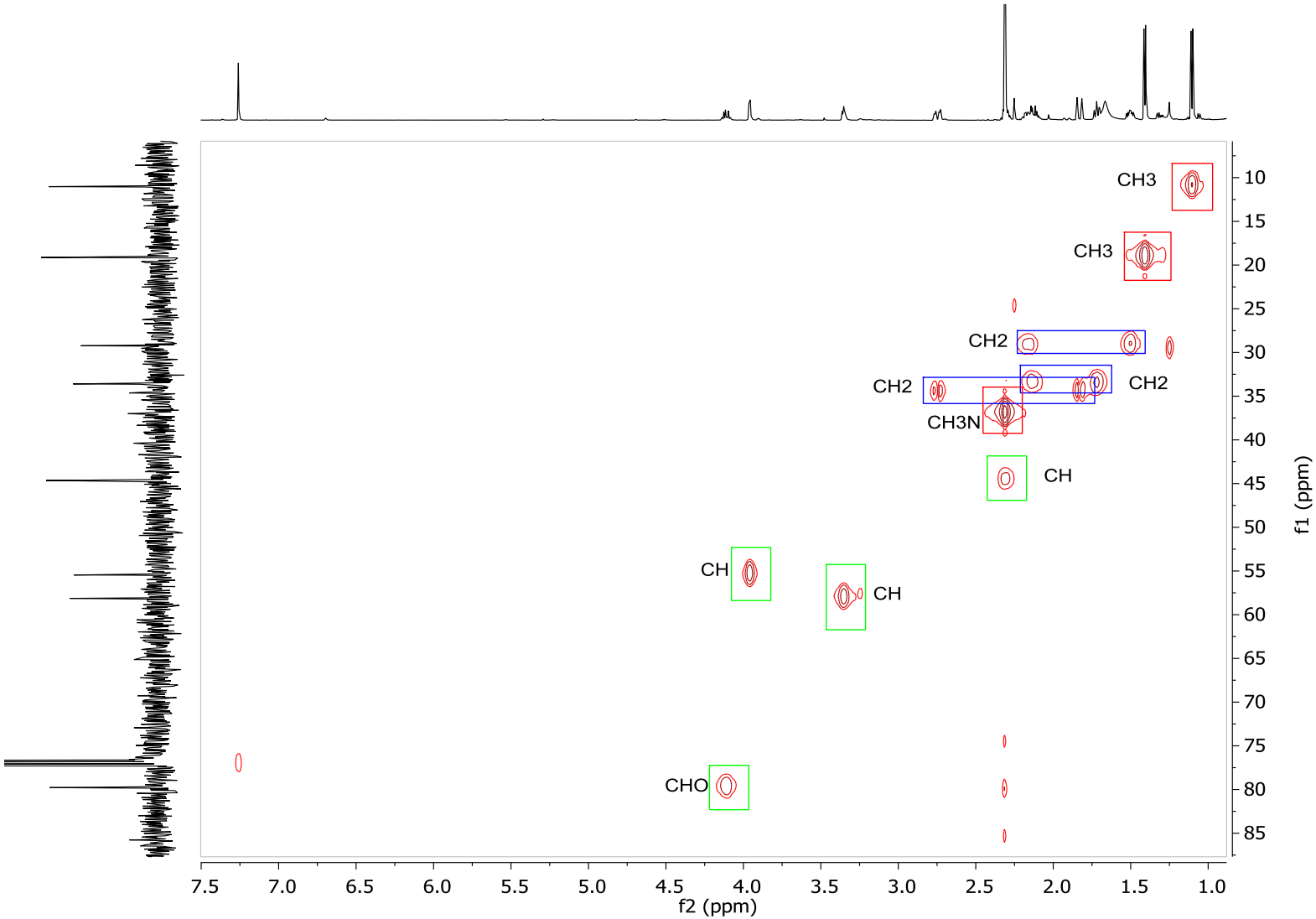

Figure S4. HSQC NMR spectrum of **3** (2,3-dihydrodarlingine) (600 MHz, $CDCl_3$).

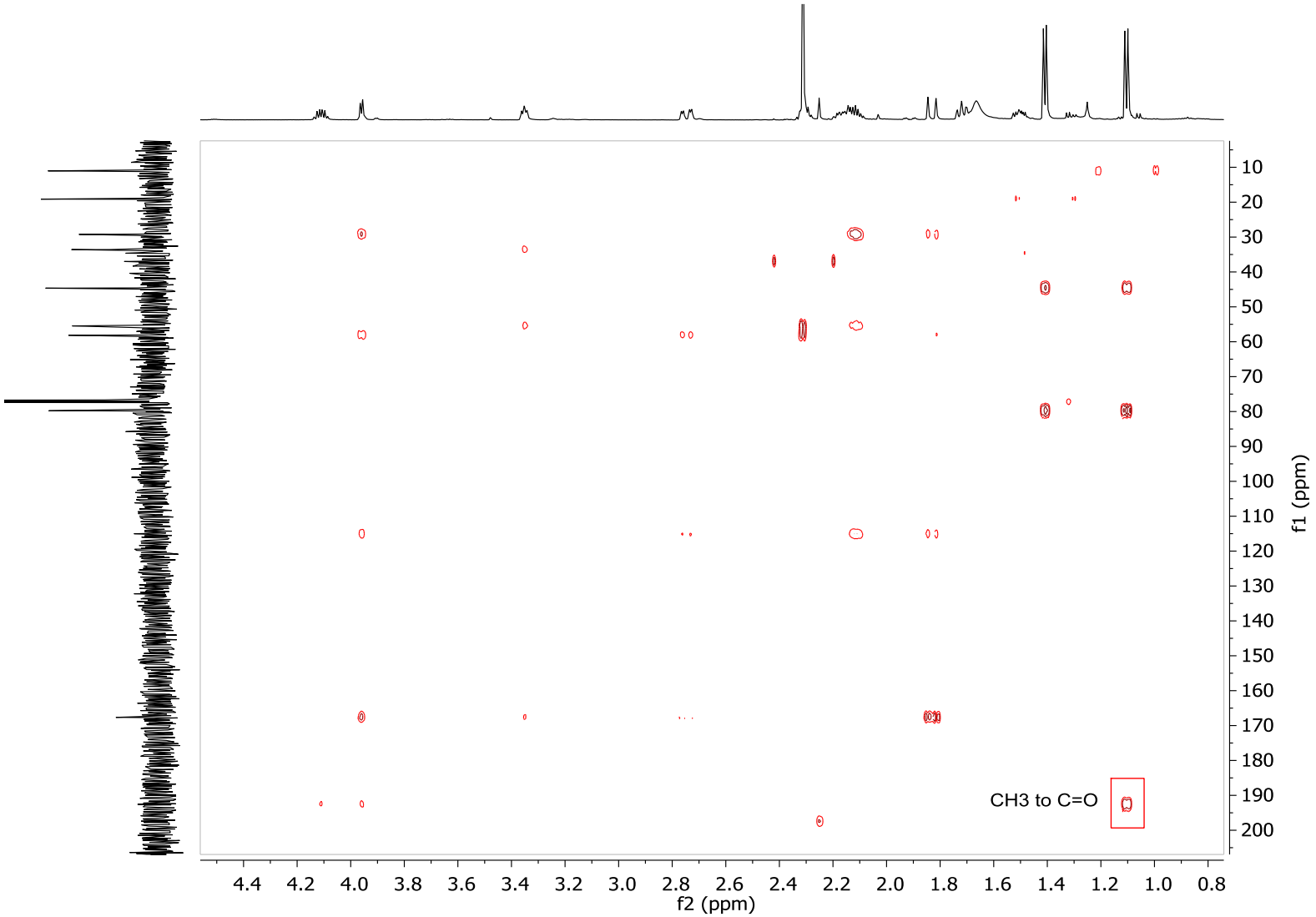

Figure S5. HMBC NMR spectrum of **3** (2,3-dihydrodarlingine) (600 MHz, $CDCl_3$).

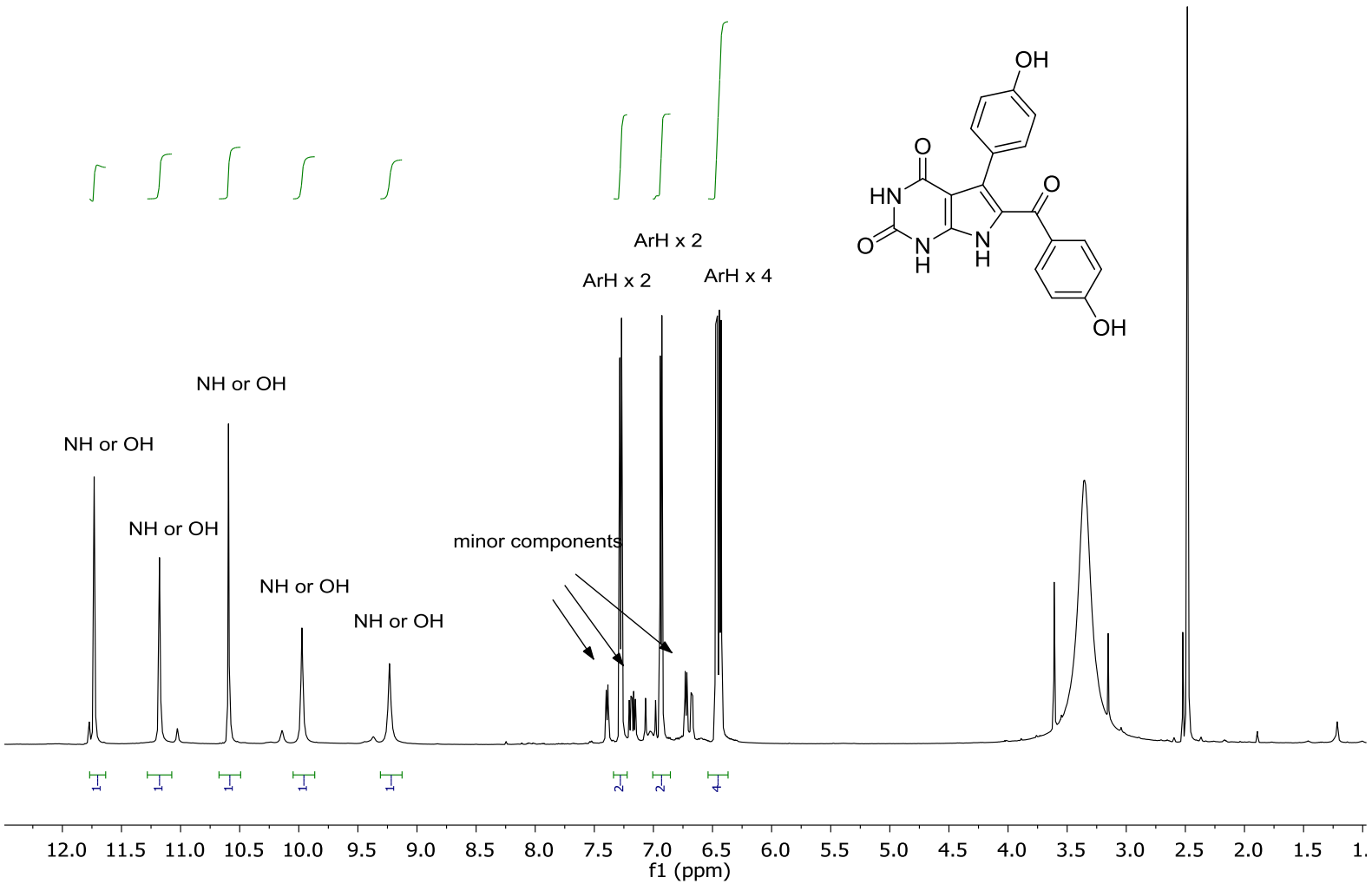

Figure S6. ^{1}H NMR spectrum of **4** (rigidin) (600 MHz, DMSO-*d*6)

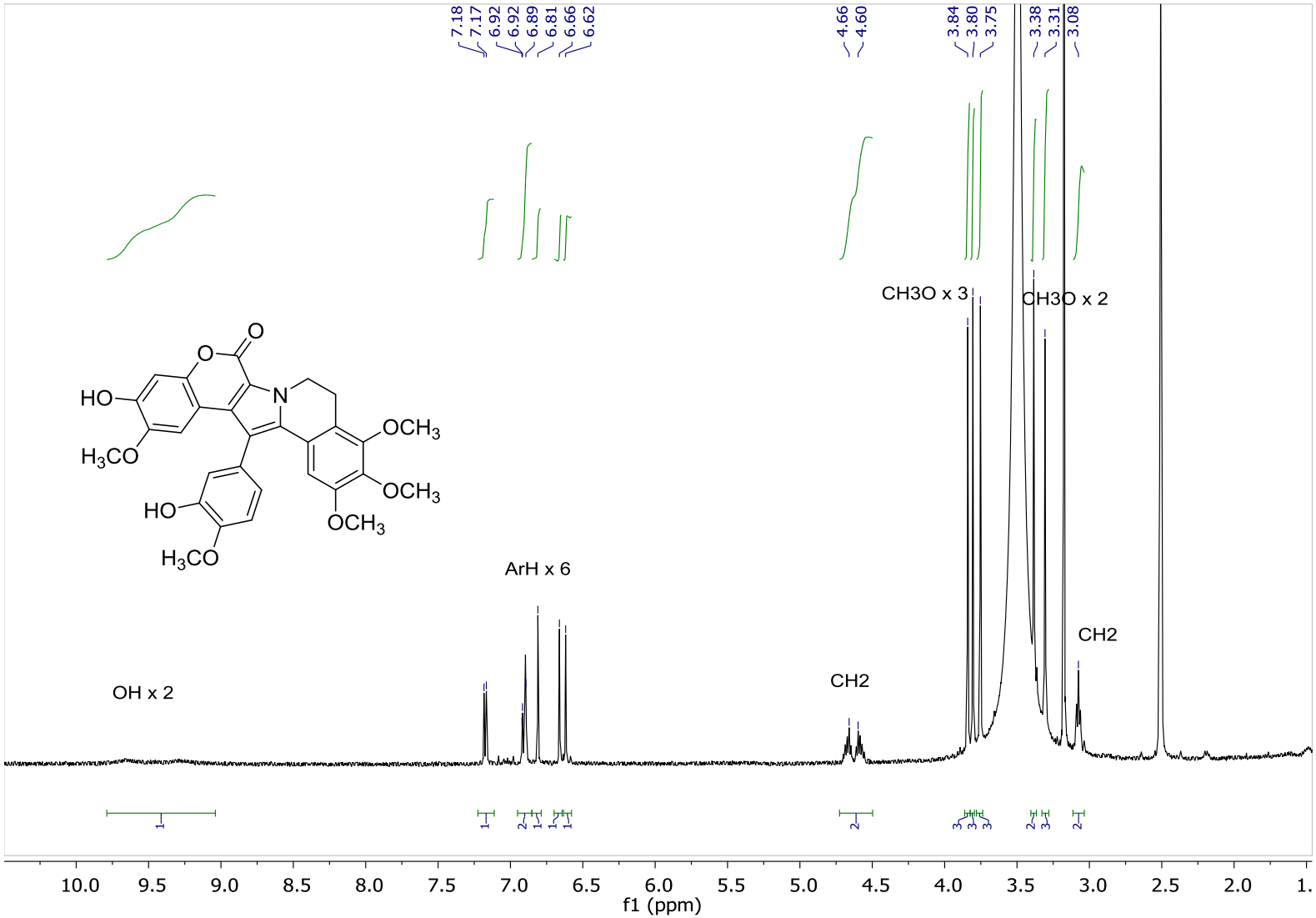

Figure S7. ^{1}H NMR spectrum of **5** (lamellarin T) (600 MHz, DMSO-d6)

Quick Tips to Use the Database in DataWarrior

1) When you open the DataWarrior file you downloaded, close all graphics windows and leave only the Table and Structures windows open.

2) Delete all search criteria on the right pane, if any.

3) In the EDIT menu select NEW FILTER. Use text filters and the appropriate options in the dropdown box (starts with, contains, equals). To search ranges, use the SLIDER filter format when selecting the new filter and then in the right pane select the min and max values by double-clicking on the numbers at the ends of the slider.

4) If you could deduce a partial structure from the NMR data, include it in the search: In the EDIT menu select NEW FILTER and then Structure [Structure List, SSS]. On the filters pane, right click in the white part of the filter and choose ADD and draw the substructure. Note that when you hover over an atom a blue dot appears and if click Q, you can define various atom properties. If you draw two substructures in the same filter it will work as OR Boolean. To create an AND query, you need set two Structure [Structure List, SSS] filters. To do this, create a new filter and select SHOW DUPLICATE FILTERS check box. Now, both SS must be present in the same structure. Note, however, that using this procedure you are not counting number of the SS, just their presence.

5) The yin-yang bouton in each search means a NOT Boolean and can be used to omit structures that comply to the query.

6) For other functionalities read DataWarrior help.

KNIME workflow and node setings for substructure counting

Free resources employed:
UNDP files (SDF and csv format) available at
http://pkuxxj.pku.edu.cn/UNPD/download.php were downloaded.
KNIME and all chemistry nodes were downloaded from
https://www.knime.org/downloads/overview
SMARTS strings were build according to Daylight instructions
(http://daylight.com/dayhtml_tutorials/languages/smarts/index.html) and checked with
SMARTS viewer (http://smartsview.zbh.uni-hamburg.de/smartsview/view).
DataWarrior was obtained at
http://www.openmolecules.org/datawarrior/download.html.

DEREP-NP generation:

1) Before running the workflow (Figure S5), the UNDP files were merged using KNIME "SDF Reader" and "SDF writer". From the merged file, only the fields corresponding to 2D and 3D structures, molecule name, CAS number, molecular formula and molecular weight were written to a new SDF file (DEREP-NP).
2) The SMARTS queries, corresponding to the structural features to be counted in each structure were created and manually
transferred (copy and paste) to the "Table Creator" node in a KNIME workflow (Figure S6).
3) The SMARTS statements present as strings in each row in the previous node are then translated and replaced by the Molecule Type Cast node into SMARTS structure type queries (Figure S7).
4) The SDF file prepared in step 1 is read together with the output of the Molecule Type Cast node are processed by the Substructure Match Counter node (Figure S8) which includes new columns corresponding to each feature to be counted.
5) The output of the later node is written to a new SDF file that contains the previous data and new columns with the features counting for each molecule.
6) This SDF file can be read by DataWarrior or any other chemical-aware software.

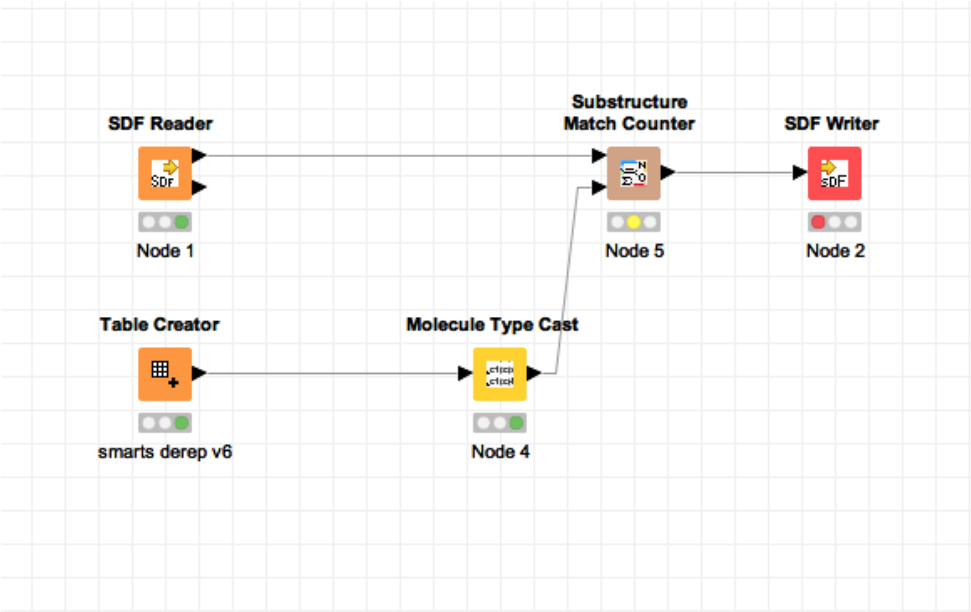

Figure S8. Knime workflow indicating the nodes used

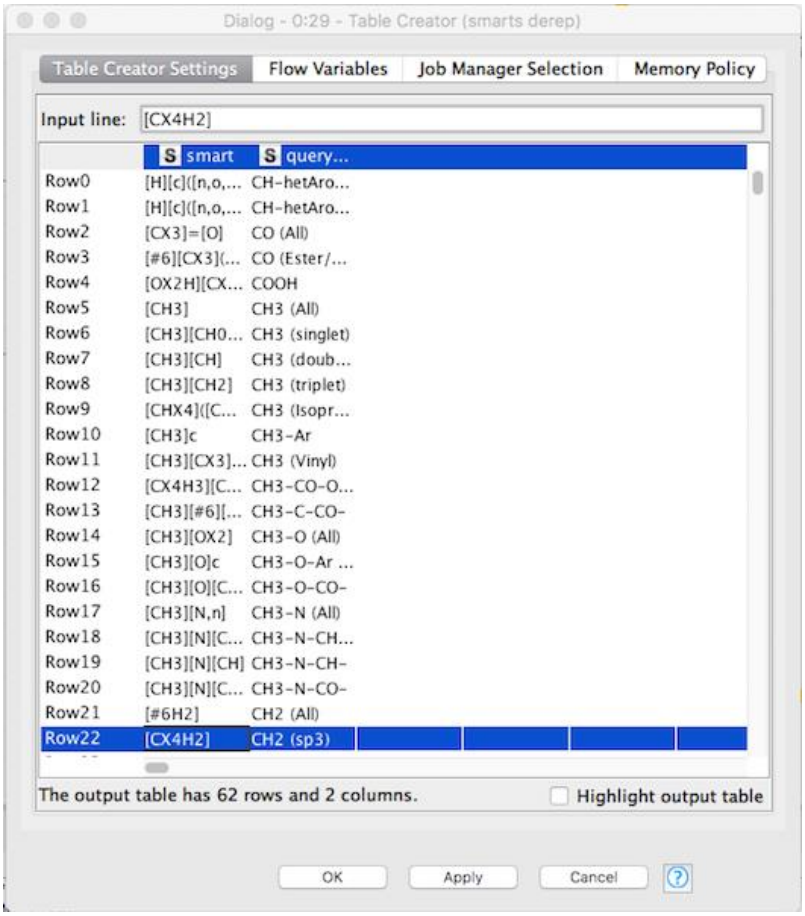

Figure S9. Dialog box of the node "Table Creator" with the pasted SMART strings and query name

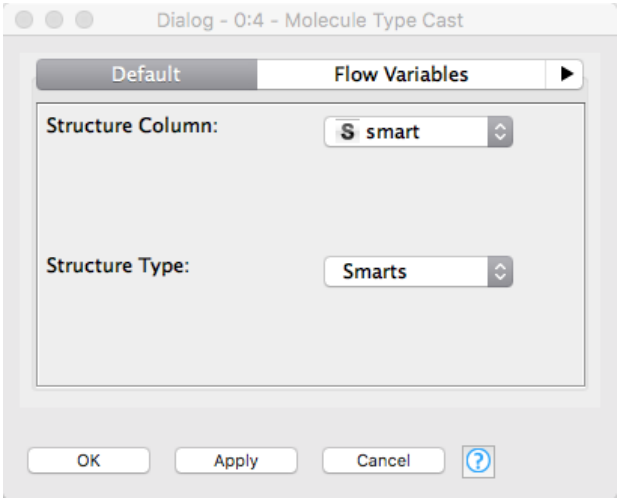

Figure S10. Setings of the node "Molecule Type Cast"

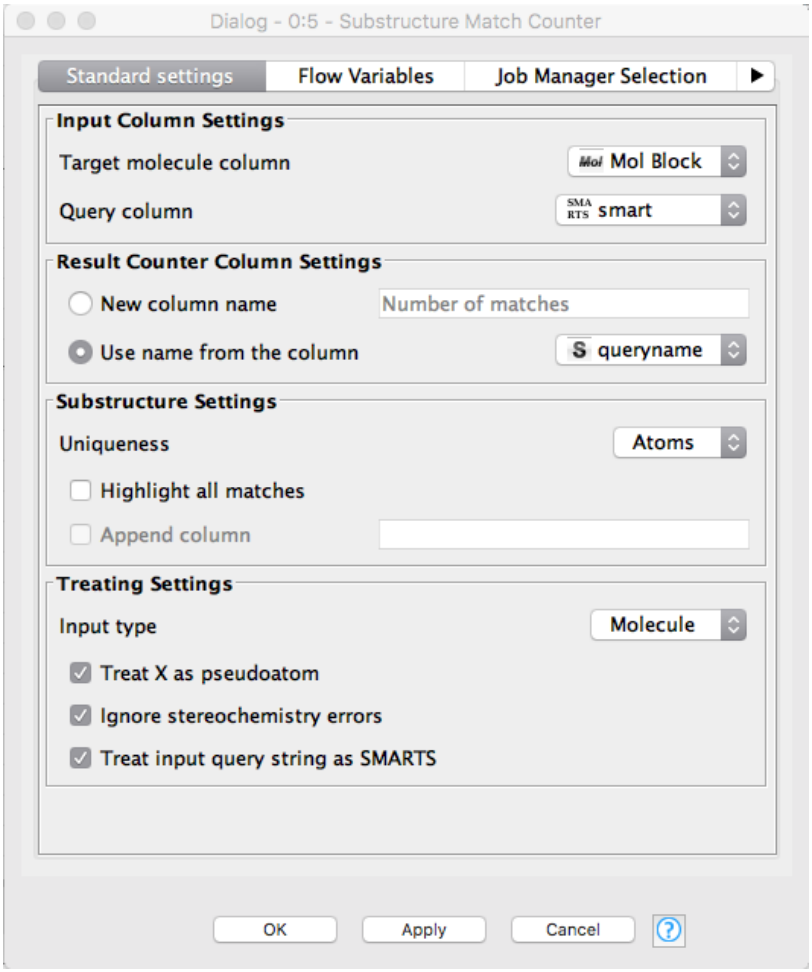

Figure S11. Dialog box of the "Substructure Mach Counter" node with the settings used