ReCQC User Guide

Version 1.0

contents

ReCQC 1.0 User Guide

I. Software Overview

ReCQC 1.0 is a professional natural product analysis software designed for rapid compound identification using carbon NMR and DEPT spectral data. Developed with Python, it employs advanced matching algorithms and scoring strategies to provide natural product researchers with an efficient and accurate analytical tool.

II. Installation and Launch

2.1.System Requirements

Operating System: Windows 10/11

Memory: 8GB (16GB recommended)

Storage Space : 2GB available space

2.2 Installation Methods

2.2.1 One-Click Installation (Recommended)

Download the installation package from GitHub, extract the files, and the software is ready to use.

2.2.2 Source Code Installation

Python Environment: 3.7 or later (required for source code installation). Create a Python virtual environment. Install dependencies: rdkit, tkinter, pillow, pandas. Run the gui.py file to launch the software

2.3 Operation

After installation, double-click the desktop shortcut to launch the software. The main interface consists of three primary areas:

(1) **Main Page**: Entry point for spectral search functionality, as shown in the figure.

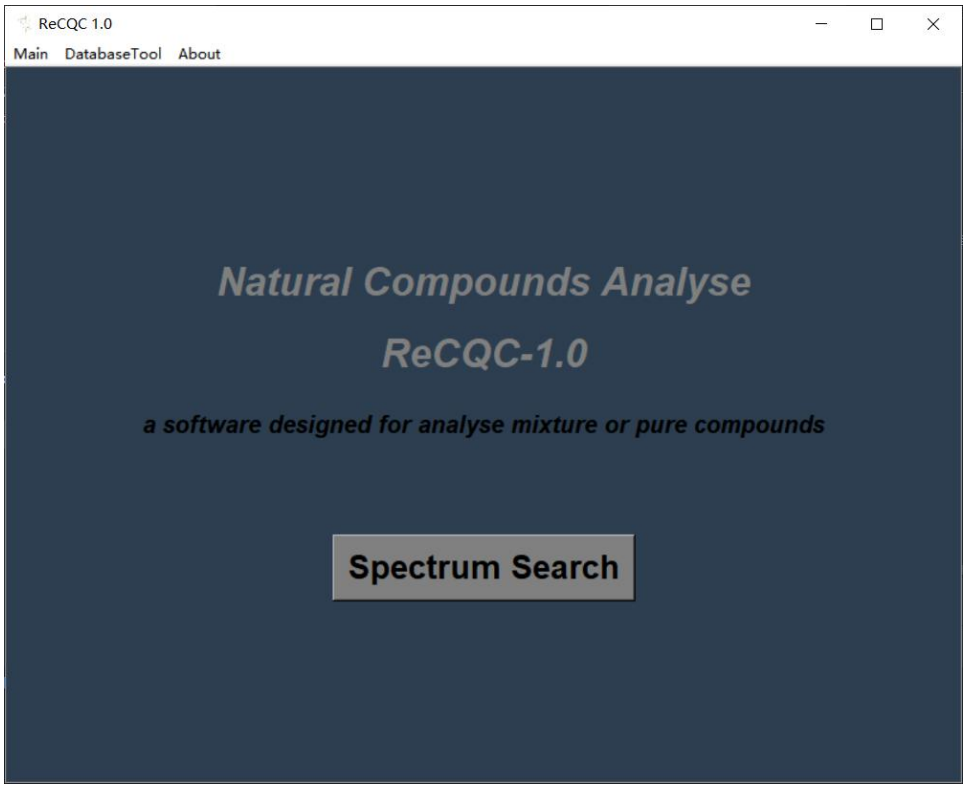


Figure 1. ReCQC Main Interface

(2) **Parameter Page**: Parameter configuration interface, as shown in the figure.

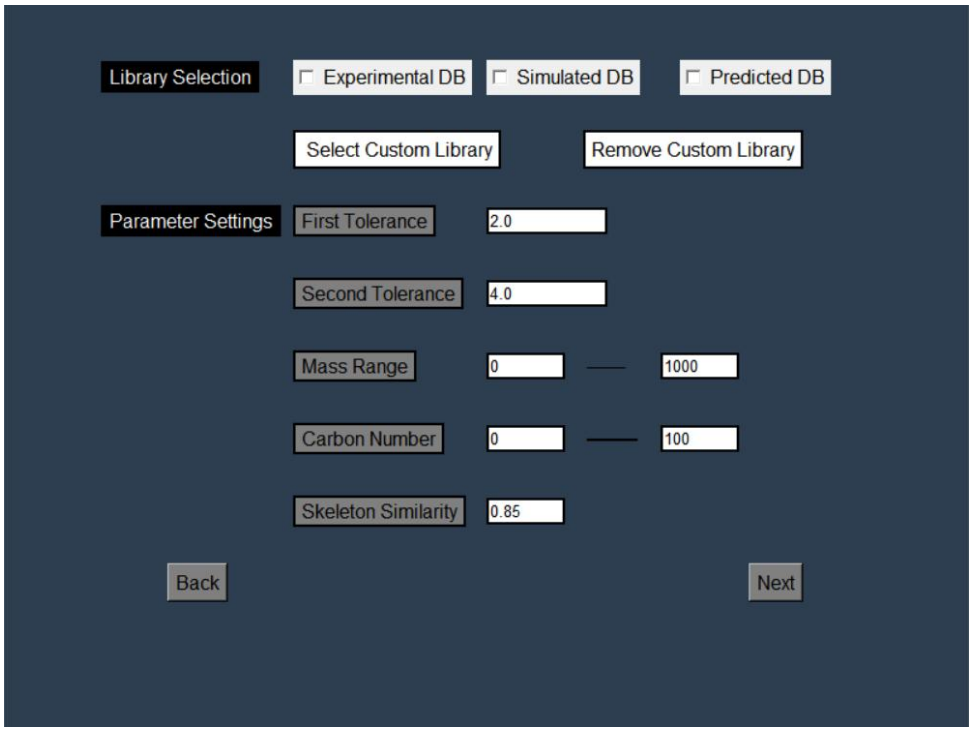


Figure 2. Parameter configuration interface

(3) **Database Tools**: Tools for database generation and management, as shown in the figure.

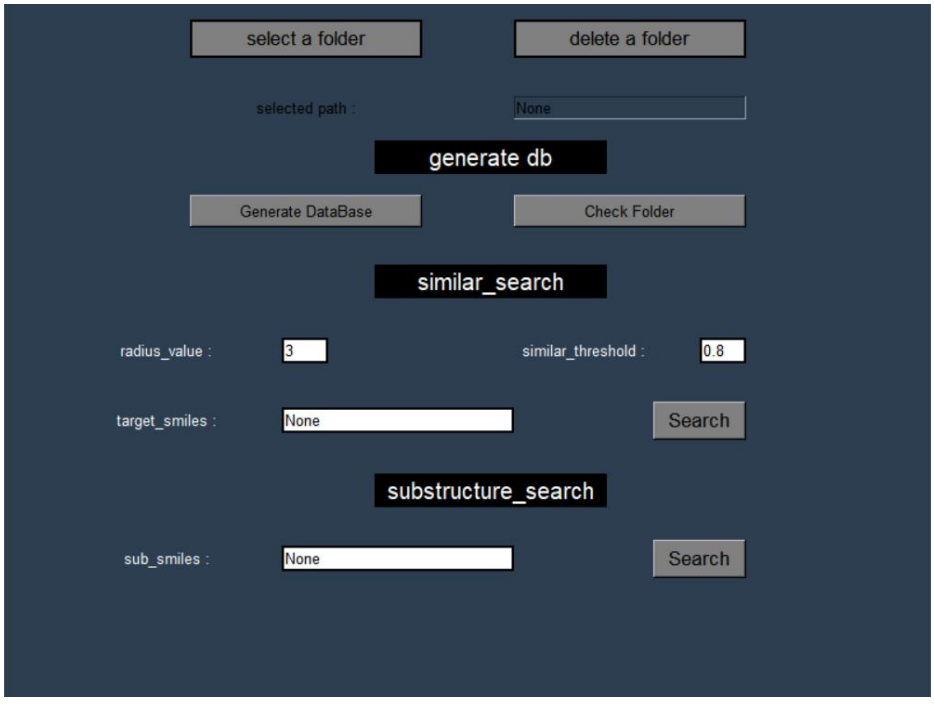


Figure 3. Plugin Interface

III. File Format Requirements

3.1. Database Generation Plugin File Format Requirements

Folder structure for database generation: Must contain paired SDF and Excel files. SDF files must contain accurate molecular structure information. Each SDF file must contain one and only one molecule.

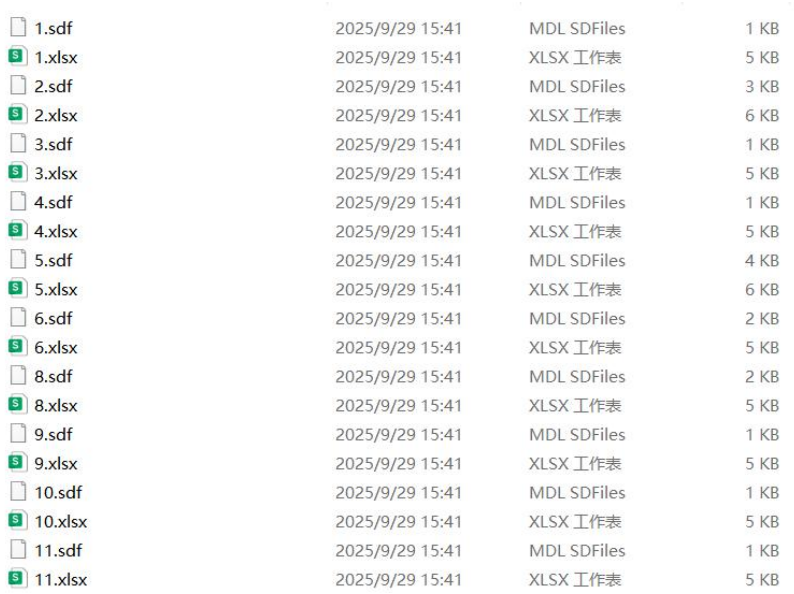
```
  RDKit          2D

 6  6  0  0  0  0  0  0  0  0999 V2000
   2.3105   -1.3282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.1576   -1.9923    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.1529   -3.3205    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.3013   -3.9892    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   3.4542   -3.3251    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   3.4588   -1.9969    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
 1  2  2  0
 1  6  1  0
 2  3  1  0
 3  4  2  0
 4  5  1  0
 5  6  2  0
M  END

$$$$
```

Figure 4. Molecular Structure File Schematic Diagram

File Naming Rules: SDF files and Excel files must have identical names. Extensions: Use .sdf for SDF files, and .xlsx or .xls for Excel files, as shown in the figure below.



1.sdf	2025/9/29 15:41	MDL SDFiles	1 KB
1.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
2.sdf	2025/9/29 15:41	MDL SDFiles	3 KB
2.xlsx	2025/9/29 15:41	XLSX 工作表	6 KB
3.sdf	2025/9/29 15:41	MDL SDFiles	1 KB
3.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
4.sdf	2025/9/29 15:41	MDL SDFiles	1 KB
4.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
5.sdf	2025/9/29 15:41	MDL SDFiles	4 KB
5.xlsx	2025/9/29 15:41	XLSX 工作表	6 KB
6.sdf	2025/9/29 15:41	MDL SDFiles	2 KB
6.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
8.sdf	2025/9/29 15:41	MDL SDFiles	2 KB
8.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
9.sdf	2025/9/29 15:41	MDL SDFiles	1 KB
9.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
10.sdf	2025/9/29 15:41	MDL SDFiles	1 KB
10.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB
11.sdf	2025/9/29 15:41	MDL SDFiles	1 KB
11.xlsx	2025/9/29 15:41	XLSX 工作表	5 KB

Figure 5. Folder Structure for Database Generation

Excel File Content Format Requirements:

Column 1: Serial number (customizable, with header)

Column 2: Carbon atom index (integer corresponding to the C-atom index in the molecular file, with header)

Column 3: Chemical shift value (floating-point number, with header)

Data must start from the first row (the first row is the header, customizable).

Atom numbering must correspond to the SDF file, and chemical shift values must match the atom indices, as shown in the figure below.

	A	B	C
1	num	Index	13C_shift
2	1	1	136.4
3	2	2	124.3
4	3	3	149.95
5	4	5	149.95
6	5	6	124.3

Figure 3. Chemical Shift Assignment Table

Output Result: The integrated SDF file generates the following data for each molecule: ID, FW, Predicted ^{13}C shifts, Quaternaries, Tertiaries, Secondaries, and Primaries. This file can be directly used as a searchable database input.

3.2. File Requirements for Similarity and Substructure Extraction Plugin

The input should be a folder containing files in standard SDF format, supporting either 2D or 3D coordinates (3D recommended).

A single SDF file may contain one or multiple molecules, and the folder can contain multiple SDF files.

If the SDF files include NMR data, the extraction results will also contain this NMR data.

It is recommended that the files include property fields such as ID and Name.

3.3. Format Requirements for the ^{13}C NMR Molecular Database as Input

The database is a single SDF file containing multiple database molecules. Each molecule in the database must contain accurate molecular structure information along with the specified property fields: ID, FW, Predicted ^{13}C shifts, Quaternaries, Tertiaries, Secondaries, Primaries.

The NMR data and numbering should be delimited by tab characters. The numeric values within square brackets represent the atom index values incremented by 1. The specific format is as shown in the figure below.

```
0Chemicalbook2548-87-0.MOL
  RDKit          2D

 9  8  0  0  0  0  0  0  0  0999 V2000
  -2.1434    0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.4289   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.7145    0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.0000   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.7145    0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.4289   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.1434    0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.8579   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -2.8579   -0.2062    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
 1  2  1  0
 2  3  2  0
 3  4  1  0
 4  5  1  0
 5  6  1  0
 6  7  1  0



>  <ID>  (7)
108

>  <Predicted 13C shifts>  (7)
0[6]        31.34
1[5]        27.57
2[7]        22.43
3[4]        32.69
4[8]        13.92
5[3]        158.92
6[2]        133.02
7[1]        194.0

>  <FW>  (7)
126.104465068

>  <Formula>  (7)
C8H14O

>  <Quaternaries>  (7)


>  <Tertiaries>  (7)
1           194.0
2           133.02
3           158.92

>  <Secondaries>  (7)
4           32.69
5           27.57
6           31.34
7           22.43

>  <Primaries>  (7)
8           13.92

$$$$
```

Figure 7. Molecular File Diagram with Required PropertiesExperimental

3.4. ^{13}C NMR Input Data Format Requirements

All_C: Full carbon spectrum data. Example: "167.90,195.7,89.4,169,95.6"

Up_DEPT135: Signals for CH and CH_3 groups (positive peaks). Example: "105.8,34.2,91.2"

Down_DEPT135: Signals for CH_2 groups (negative peaks). Example: "40.1,44,53"

DEPT90: Signals for CH groups. Example: "105.8,34.2,91.2"

Input Requirements: Values must be comma-separated. It is recommended to remove noise signals with a signal-to-noise ratio below 3:1.

IV. Detailed Explanation of Parameter Settings

4.1. Basic Filtering Parameters

Carbon Count Range: Default 0-100, recommended 10-40 for plant extracts

Molecular Weight Range: Default 0-1000, recommended 100-600 for natural products

Primary Tolerance: Default 2.0 ppm, defines the full match region (adjust based on database precision)

Secondary Tolerance: Default 4.0 ppm, defines the partial match region (adjust based on database precision)

Scaffold Similarity: Default 0.85, when the system finds a maximum substructure score exceeding this value, it will boost the weight

V. Operational Procedures

5.1. Standard Spectral Analysis Workflow

Launch the software → Click "**Spectrum Search**".

Select a database: **System Databases**: Experimental DB, Simulated DB, Predicted DB. **Custom Database (Recommended)**: Click "**Select Custom Library**". Set Parameters: Configure Carbon Count, Molecular Weight, Tolerance values, etc. Input Data: Enter the corresponding chemical shifts in the respective text boxes. Execute Analysis: Click "**Search**" to begin matching. View Results: Review the list of candidate compounds, ranked by score.

5.2. Database Generation Workflow

Access Database Tools: Click the "**Database Tool**" menu.

Select Folder: Click "**select a folder**" and choose the folder containing the paired SDF and Excel files.

Validate Files: Click "**Check Folder**" to verify the file formats are correct.

Generate Database: Click "**Generate DataBase**" to create the standardized SDF file.

5.3. Similarity Search Workflow

Select Folder: Click "**select a folder**".

Set Parameters: Similarity Threshold (0.0-1.0), Radius (1-5).

Input Target: Enter the SMILES expression in the "**target_smiles**" box.

Execute Search: Click "**Search**". Results are saved as **similar_molecules.sdf**.

5.4. Substructure Search Workflow

Input Substructure: Enter the substructure SMILES in the "**sub_smiles**" box.

Execute Search: Click "**Search**". Results are saved as **molecules_with_substructure.sdf**.

VI. Result Interpretation and Analysis

6.1. Scoring System

0.9-1.0: Nearly identical structures

0.8-0.9: Highly similar structures

0.6-0.7: Related structures

< 0.6: Structures with significant differences

6.2. Result File Description

Spectral Matching: Results are displayed directly in the interface, with candidate compounds sorted by score.

Similarity Search: Outputs the `similar_molecules.sdf` file, containing similar molecules and their data.

Substructure Search: Outputs the `molecules_with_substructure.sdf` file, containing all molecules matching the substructure.

ReCQC 2.0 User Guide

I. Software Overview

ReCQC 2.0 is a professional natural product mixture analysis software designed for rapid compound identification using carbon NMR and HSQC spectral data. Developed with Python, it employs advanced matching algorithms and scoring strategies to provide natural product researchers with an efficient and accurate analytical tool.

II. Installation and Launch

2.1. System Requirements

Operating System: Windows 10/11, macOS 10.14+, or Linux Ubuntu 18.04+

Memory: 8GB (16GB recommended)

Storage Space: 2GB available space

Python Environment: 3.9.x (if choosing source code installation)

2.2. Installation Methods

One-Click Installation (Recommended): Download the compressed package for your system, extract and run directly. Windows users should execute the .exe file.

Source Code Installation: First create a Python virtual environment, then install dependencies (rdkit, matplotlib, pillow, numpy, tqdm), finally download the source code from GitHub and run the gui.py file.

2.3. First Run

After installation, double-click the desktop shortcut to launch the software. The main interface consists of four main areas: mode selection, parameter settings, data input, and result display. New users are recommended to practice first using the sample data.

III. Database File Format

The software uses the SDF (Structure-Data File) format for database files. Molecular structure files must include hydrogen atoms (if using HSQC data). Each compound record must contain the following fields:

Required Field Descriptions:

ID: Unique compound identifier (e.g., "Nicotine" or "CHEM_001")

FW: Molecular weight, must be a numerical value (e.g., "162.12")

Predicted ^{13}C shifts: Predicted or experimental carbon-13 chemical shift data

Quaternaries: Quaternary carbon (C) shifts

Tertiaries: Tertiary carbon (CH) shifts

Secondaries: Secondary carbon (CH2) shifts

Primaries: Primary carbon (CH3) shifts

HydrogenShifts: Predicted hydrogen chemical shift data, formatted the same as "Predicted 13C shifts"

IV. Detailed Parameter Settings

4.1. Basic Filtering Parameters

Carbon Count Range: Filters molecules in the database based on the specified number of carbon atoms.

For plant extract analysis and microbial metabolite analysis, typically set to **10-40** carbon atoms.

This parameter effectively narrows the search scope and improves analysis efficiency.

Molecular Weight Range: Filters molecules based on exact molecular weight. Low molecular weight range (**100-300 Da**): Suitable for volatile components. Medium molecular weight range (3**00-600 Da**): Covers most natural products. High molecular weight range (**600-1500 Da**): Suitable for peptides and macrolides.

Note: Molecular weight calculation is based on neutral molecules; adjust the range appropriately for ionic compounds.

4.2 Carbon Spectrum Analysis Parameters

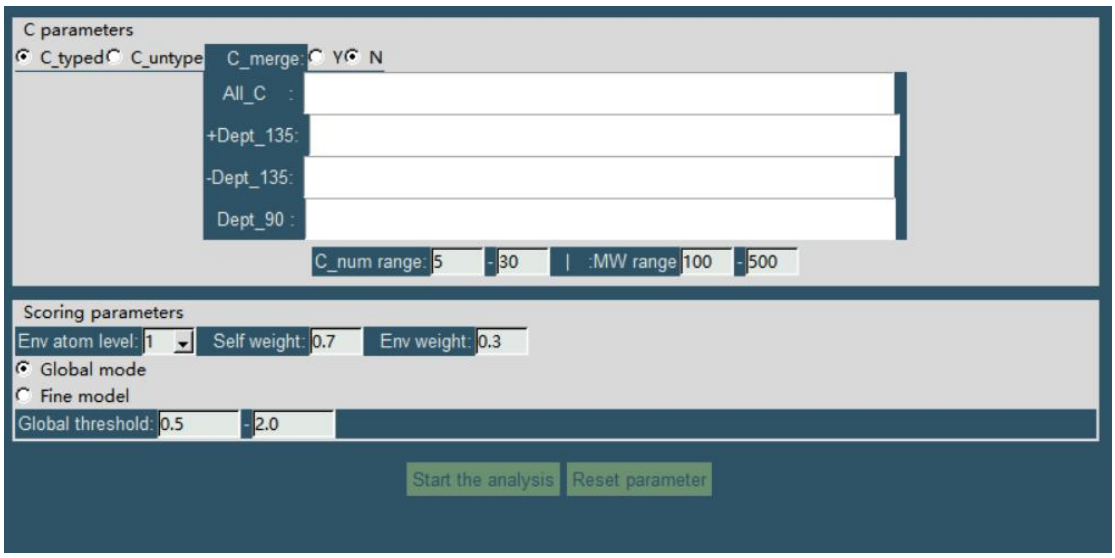


Figure 1: ^{13}C Parameter Settings Interface

Carbon Spectrum Mode Selection: Includes Classified and Unclassified modes. Classified Mode: Utilizes carbon type information to improve matching accuracy; suitable for high-precision identification with complete DEPT data. Unclassified Mode: Simple input and fast computation; ideal for rapid screening or preliminary analysis.

Equivalent Carbon Handling: Used for processing highly symmetric molecules. When chemically equivalent carbon atoms exist in a molecule, enabling this function treats them as signal groups, preventing matching errors caused by signal overlap. For asymmetric molecules or scenarios requiring atomic-level precision matching, it is recommended to disable this option.

Atomic Environment Level: Controls the depth of atomic environment consideration. **Level 1**: Considers only directly bonded atoms; fast computation speed.

Level 2: Considers the atomic environment within two bonds; achieves the best balance between accuracy and efficiency, recommended for most situations. **Level 3**: Considers the complete environment within three bonds; provides the highest precision but also the highest computational complexity.

Weight Assignment System: Includes Self Weight and Environment Weight.

Self Weight: Controls the contribution of the target atom's chemical shift match.

Environment Weight: Reflects the degree of match of the surrounding atomic environment.The sum of both weights must be 1.0. For precise matching scenarios, recommended: Self Weight **0.7**, Environment Weight **0.3**. For analog screening scenarios, an equal weight of **0.5** and **0.5** can be used.

Dual Threshold Scoring System: Offers two configurations: Global Mode and Fine Mode. Global Mode: Sets uniform tolerance thresholds for all chemical shifts. First Threshold (Green): Defines the full match region; typical setting 1.0-2.0 ppm. Second Threshold (Yellow): Defines the partial match region; typical setting 2.0-4.0 ppm.

Fine Mode: Allows setting different thresholds for different chemical shift ranges. For example, looser thresholds can be set for the aliphatic region (**0-50 ppm**), and stricter thresholds for the aromatic region (**100-160 ppm**).

4.3 HSQC Analysis Parameters

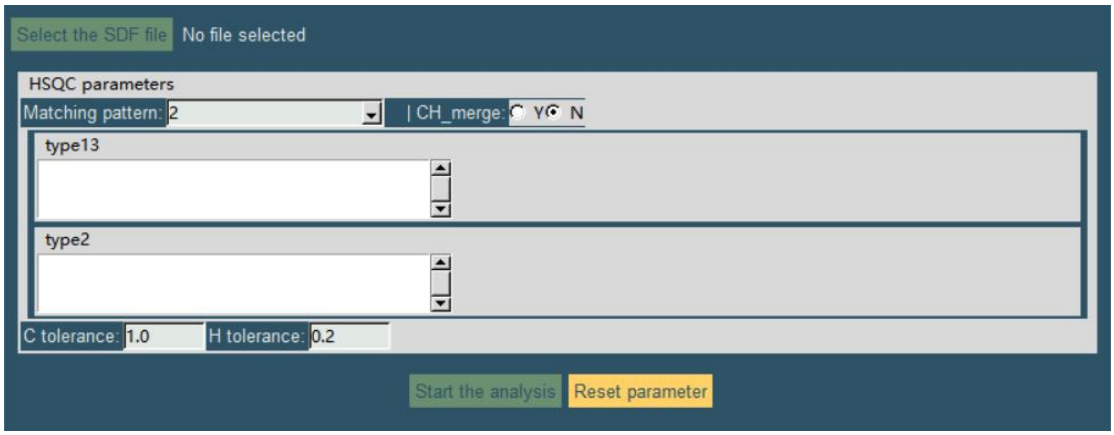


Figure 2. HSQC Parameter Configuration Interface

Matching Mode Selection: Includes three options. **Mode 1**: Processes all CH types uniformly; fastest calculation speed. **Mode 2**: Groups CH/CH_3 and CH_2 for processing; balances accuracy and efficiency. **Mode 3**: Fully separates into CH, CH_2, and CH_3 categories; provides the highest matching precision. The choice requires balancing data quality, precision requirements, and computational resources.

Tolerance Settings: Include Carbon Chemical Shift Tolerance and Hydrogen Chemical Shift Tolerance. Carbon Tolerance: Routine setting 1-3 ppm; strict setting can be 0.5-1 ppm. Hydrogen Tolerance: Routine setting 0.15-0.25 ppm; strict setting can be 0.10-0.15 ppm. Settings should consider instrument precision, data quality, and compound type. Smaller tolerances can be set for high-field instruments and rigid structures.

4.4 Joint Analysis Parameters

Weight Assignment: Balances the contribution of Carbon Spectrum and HSQC data to the final score.

Carbon Spectrum Weight: Reflects the reliability of the carbon spectrum data.

HSQC Weight: Reflects the quality of the HSQC data. When the quality of both data types is comparable, a weight assignment of **0.6** and **0.4** is recommended. If the carbon spectrum data quality is significantly better, increase the Carbon Spectrum Weight to **0.7-0.8**; conversely, increase the HSQC weight.

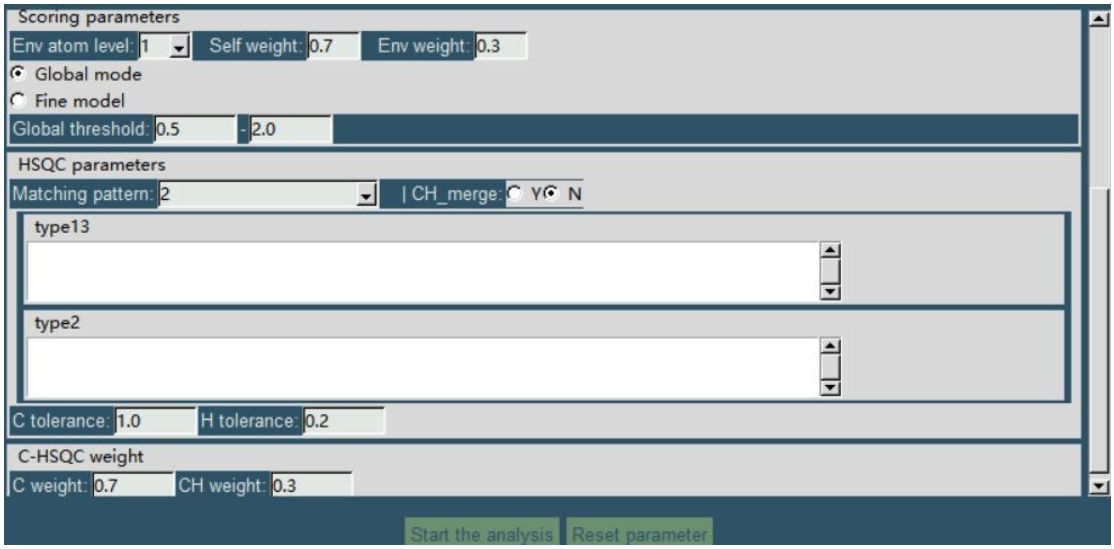


Figure 3: Joint Retrieval Interface Parameters

V. Analysis Workflow

After launching the software, first select the appropriate working mode based on the available data types: If both carbon spectrum and HSQC data are available, select the Joint Matching Mode. If only carbon spectrum data is available, select the Carbon

Spectrum Matching Mode. If only HSQC data is available, select the CH Matching Mode. The main interface is shown in the figure below.

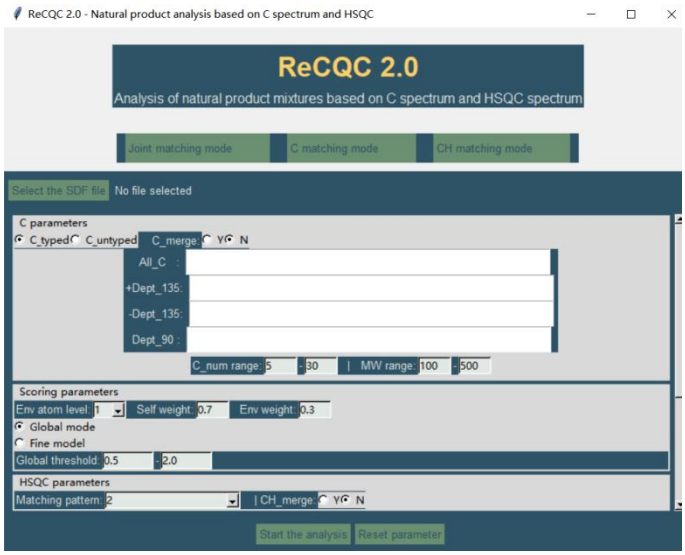


Figure 4: ReCQC 2.0 Main Workspace Interface

Click the **"Select SDF File(s)"** button to load the database file(s). Batch loading of multiple files is supported. During runtime, the software automatically verifies the file format and content integrity; if any issues are detected, corresponding prompts will be displayed.

Set various parameters according to the sample characteristics. Beginners can use the recommended values provided by the software, while experienced users can adjust them based on specific needs.

After completing the parameter settings, input the experimental data according to the format requirements: Input carbon spectrum data into the corresponding text box(es) based on the selected mode. Input HSQC data in the coordinate pair format. Click the **"Start Analysis"** button to execute the analysis. The software will display

real-time progress, including steps such as database filtering, molecular matching, and score calculation. The analysis time varies depending on the database size and parameter complexity, typically ranging from a few minutes to several tens of minutes.

VI. Interpretation and Analysis of Results

After the analysis is complete, the results interface displays candidate molecules in descending order of matching score. Matches with scores above **0.8** are considered highly reliable and are recommended for priority verification. Clicking on a specific molecule reveals a detailed matching report, including: A chemical shift comparison for each atom. The individual matching score. Annotations for mismatched atoms.

The software provides molecular structure visualization, allowing users to intuitively see the positions of matched atoms within the molecular structure. This is particularly helpful for understanding the structural basis of the match, especially in the identification of complex natural products.

For significant results, it is recommended to perform **multi-faceted verification**: Check the matching status of key structural features. Analyze the reasons for mismatched atoms. Consider the influence of stereochemical factors. If possible, compare the results with those obtained from known standards or other identification methods.

VII. Advanced Features and Tips

7.1. Performance Optimization Suggestions

When processing large databases, setting appropriate **Carbon Count** and **Molecular Weight** ranges can significantly improve efficiency. Selecting the appropriate **Atomic Environment Level** is also crucial; **Level 2** offers the best performance balance in most cases. Closing unnecessary background applications can also improve the runtime experience.

7.2. Result Verification Strategies

Internal Verification: Includes checking the self-consistency of matches and the reasonableness of the scoring system. **External Verification**: Can involve comparison with known standards or verification of results using other identification methods.
It is recommended to perform **replicate analyses** on important findings to ensure the reliability of the results.

If the software fails to load an SDF file, check that the file format meets the requirements, especially whether all required fields are complete. File paths containing Chinese characters might also cause issues. When errors occur during the analysis process, first check the error message prompt. Common causes include incorrect input data format, unreasonable parameter settings, or insufficient memory. Adjusting the relevant settings based on the error message usually resolves the problem.

Unsatisfactory matching results might be due to inadequate database coverage, poor data quality, or inappropriate parameter settings. Trying to adjust the tolerance thresholds, optimize weight assignments, or use different matching modes may improve the results.