ToMoCoMD-CARDD

a descriptors calculation tool

[Suite-Module]
QuBiLS-MIDAS v1.2

Software User Manual

ToMoCoMD-CARDD is a molecular descriptor (MD) calculating program comprised of two suites with parallel functionalities. One is a comprehensive collection of MD calculating modules based on the so called *relations frequency matrices*, molecular fingerprints and a pool of the most relevant MDs reported in the literature. The second suite is comprised of a set of modules derived from algebraic considerations, collectively known as **QuBiLS** (acronym for Quadratic, Bilinear and N-Linear MapS). This suite includes three modules: 1) ***QuBiLS-MAS*** (QuBiLS-based on Graph–Theoretic Electronic-Density Matrices and Atomic Weightings), 2) ***QuBiLS-MIDAS*** (QuBiLS-based on N-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings) and 3) ***QuBiLS-POMAS*** (QuBiLS-based on Molecular Surface-based Potential Matrices and Atomic Weightings). In this application, only ***QuBiLs-MIDAS*** module is included. ***QuBiLs-MIDAS*** constitutes a unique combination of methods for calculating 3D-MDs on a sound algebraic basis. These MDs can be used for a wide range of applications in all areas of chemistry, in particular in drug design, lead compound discovery and optimization, QSAR/QSPR studies, similarity searching, diversity assessment of compound libraries and prediction of ADMET properties.

USER MANUAL

ToMoCoMD-CARDD
QuBiLS-MIDAS v1.2

QuBiLS-MIDAS v1.2 is a program that calculates "novel 3D algebraic descriptors based on N-linear maps".

CENTRO DE ESTUDIO DE MATEMATICA COMPUTACIONAL (CEMC)
Grupo de Investigación de Bioinformática
Universidad de las Ciencias Informáticas
La Habana, Cuba.

ESCUELA DE MEDICINA
Colegio de Ciencias de la Salud, Edificio de Especialidades Médicas
Hospital de los Valles
Universidad San Francisco de Quito (USFQ)
Quito, Ecuador

March, **2016**

User's Manual
Authorization Memorandum

I have carefully assessed the User's Manual for **ToMoCoMD-CARDD (QuBiLS-MIDAS)**.

MANAGEMENT CERTIFICATION - Please check the appropriate statement.

_____ The document is accepted.

_____ The document is accepted pending the changes noted.

_____ The document is not accepted.

We fully accept the changes as needed improvements and authorize initiation of work to proceed. Based on our authority and judgment, the continued operation of this system is authorized.

_____	_____
NAME Project Leader	DATE
_____	_____
NAME Operations Division Director	DATE
_____	_____
NAME Program Area/Sponsor Representative	DATE
_____	_____
NAME Program Area/Sponsor Director	DATE

USER'S MANUAL

TABLE OF CONTENTS

Page #

1.0 GENERAL INFORMATION

1.0 GENERAL INFORMATION

System Overview

ToMoCoMD-CARDD is an interactive and user-friendly, free, fully cross-platform application designed to calculate 2/3-D numerical descriptors (indices) for molecular structures, with the objective of characterizing or discriminating among them. This software is comprised of two suites with parallel functionalities.

One suite is made up of a set of modules derived from algebraic considerations (**QuBiLS** suite). This suite contains the follow modules: **QuBiLs-MAS** (acronym for Quadratic, Bilinear and Linear Maps based on Graph–Theoretic Electronic-Density Matrices and Atomic weightings), **QuBiLs-MIDAS** (acronym for Quadratic, Bilinear and N-Linear Maps based on N-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings) and **QuBiLs-POMAS** (acronym for Quadratic, Bilinear and Linear Maps based on Molecular Surface-based Potential Matrices and Atomic Weightings).

The second **ToMoCoMD-CARDD** Suite is a comprehensive collection of MD calculating modules based on the so called relations frequency matrices, molecular fingerprints and a pool of the most relevant MDs reported in the literature. These modules include:

- DIVATI (acronym for DIscrete DeriVAtive Type Indices),
- GT-STAF (acronym for Graph Theoretical Thermodynamic STAte Functions),
- FREMESSA (acronym for FREquency-type Matrices Extended claSSical Algorithms),
- FREMXALF (acronym for FREquency-type MatriX-based ALgebraic Forms),
- MOLFIP (acronym for MOLecular FIngerPrints) and
- DESPOOL (acronym for DEScriptor POOLs).

In this application, only QuBiLS-MIDAS module is included, which is for the calculation of 3D molecular descriptors based on the two-linear (bilinear), three-lineal and four-linear (multi-linear or N-linear) algebraic forms. Thus, is the unique software that compute these kinds of indices, establishing relations among two, three and four atoms, applying several (dis)similarity metrics or multi-metrics, matrix transformations (simple-stochastic, double-stochastic and mutual probability), cut-offs, local calculations and aggregation operators. The QuBiLS-MIDAS software was developed in the Java programming language and employ the Chemical Development Kit (CDK) library for the manipulation of the chemical structures and the calculation of the atomic properties. This software is composed by a desktop user-friendly interface and an API library. The former was created to ease to the users the configuration of the different options of the molecular descriptors, while the library was designed to be easily integrated in other software for chemo-informatics applications. This module present functionalities for data cleaning tasks and for batch processing of the molecular indices. In addition, it have as feature the parallel calculation of the molecular descriptors through the use of all available processors in the modern computers.

This new version has some relevant features, such as:

1) Two chemical input formats (SDF and MOL MDL files)
2) Ten properties as atomic weightings: TPSA, ALogP, Softness, Hardness and so on.

3) Four new matrix representations, non-, simple- and double-stochastic and mutual probabilistic.
4) Two new matrix forms, by using cutoffs based on Lag K and Lag R values.
5) Seven local indices, H-Bond Donor, H-Bond Acceptor, Aromatic Atoms and so on.
6) Three output file formats, Space Delimited Text file, Weka ARFF file and Comma Separated Values file.
7) Notification and information on system error and program exceptions of JRE.
8) Real-time updated logging status (see History Tab windows).
9) Optional generation of Debug Report file.
10) Distance to molecule center are computed as diagonal elements.
11) Lone-pair electrons are taken into consideration as multiple loops.
12) Included thirty brand new invariants (aggregation operators) that generalize the initial form of obtaining indices from atomic (or fragment) contributions, the new indices are obtained by using these invariants on LOVIs (Local Vertex Invariants).
13) Atom-based indices.
14) Extended chiral indices.
15) Batch Mode execution.
16) New generalized indices by using *n*-tuples atom-interactions based matrices, for $n = 2$, 3 and 4.
17) H-Atoms addition or removal capability.
18) Six sets of commonly used molecular datasets are provided in the Example Data tool.
19) Curating functionality for the input data, remove inorganic, dative bond, organometallic, salt neutralization and find duplicated structures, etc.
20) Brand new Descriptor Search Tool,
21) Several improvements to create a more user friendly GUI.
22) Enhanced speed for descriptor calculation process with more stability and robustness.
23) Distributed high-throughput molecular descriptors calculation by using a multi-server distributed computing system named T-arenal.

System requirements

QuBiLS-MIDAS Software runs on a wide variety of operating systems and computers including multi-processor clusters, multi-processor or multi-core desktops (PC and MAC), high-performance scientific workstations, and laptops. This release can run either interactively or in batch mode, which permits sequential execution to be distributed across multiple processors (and/or cores) workstations, even in a heterogeneous computing environment. In general terms the minimal and recommended system requirements are:

Hardware:

Processor: All processors developed hereafter by Intel Corp. are supported on the assembly level optimization. All AMD current processors work as old Pentium with higher clock frequency (no special optimization).

Processor Clock Speed: minimum Intel(R) Celeron(R) M processor 1.40GHz or equivalent. Recommended Intel(R) Core2Quad processor 2.5GHz or above.

Memory: 256MB minimum, 512MB default tuning. We recommend 4096 MB or above in order to improve performance.

Software:

Operation system: **ToMoCoMD-CARDD** is designed to run on any UNIX/LINUX or MAC platforms, as well as on microcomputers running Windows 95, 98, ME, 2000 or XP, Vista, 7 and above. **ToMoCoMD-CARDD** is platform-independent software.

Operation system extensions: **ToMoCoMD-CARDD** requires Java(TM) 7 Runtime Environment on the target system. It runs under any host operating system, which supports Java(TM) 7 Runtime Environment and also works on X86 and X64 based architecture.

Points of Contact

Information

For all comments, suggestions, information and inquiries about **ToMoCoMD-CARDD QuBiLS-MIDAS** Software please contact:

Ph.D. Yovani Marrero Ponce
Colegio de Ciencias de la Salud, Edificio de Especialidades Médicas
Hospital de los Valles
Universidad San Francisco de Quito (USFQ)
Quito, Ecuador.
E-mail: ymarrero77@yahoo.es
URL: http://www.uv.es/yoma/

Technical Support

For technical support please contact to:

Ph.D. César Raúl García Jacas
Grupo de Investigación de Bioinformática. Centro de Estudio de Matemática Computacional
Universidad de las Ciencias Informáticas, La Habana
Cuba
E-mail: crjacas@uci.cu
URL: https://scholar.google.com/citations?user=ND_S0RgAAAAJ&hl=en

MSc. José Ricardo Valdés Martiní
E-mail: jricky31@gmail.com

2.0 SYSTEM SUMMARY

2.0 SYSTEM SUMMARY

System Configuration

The system is prepared to maintain its default configuration regardless of the platform on which is executed. It does not require any parameters or initial configuration file, so it fits natively over the Java™ virtual machine.

The configuration process to start performing calculations of algebraic form descriptors with this application begin on the "Algebraic Form" panel, located under the tabbed pane menu or simply can be loaded from a preconfigured QuBiLS-MIDAS's project file.

QuBiLS-MIDAS has been designed to process a broad range of chemistry. There are not limitations concerning the number of atoms of a molecule. Metal atoms and, especially transition metal atoms, can be processed but might cause problems in certain atom descriptor calculation routines due to the lack of parameterization. The number of records in an input structure file is unlimited. This "*unlimited*" upper boundary takes direct dependence for total of amount system memory available, so the maximum number of structure per file is not unlimited. As a matter of fact, several tests were accomplished with an up to **265 000** structures molecule SDF file and mixed descriptors configuration.

Installation of the program

The Java™ Runtime Environment version 7 are required on the target operating system, in a Microsoft Windows platform (i.e. XP, Vista, Windows 7) follows these steps:

1. Insert the **ToMoCoMD-CARDD** CD in the CD/DVD-ROM drive of your computer.
2. If the Windows *Autorun* feature is turned on, the installation options will be displayed automatically. Skip next instruction.
3. If you have downloaded **ToMoCoMD-CARDD** suite from our website on internet or obtained from a colleague recommendation, locate the Java™ based installer file and execute it, see Figure 1.
4. Follow the on screen instructions for installation. See below (Figures 2 - 10) guided sequence of screenshot for the step by step installation process on a Windows workstation.

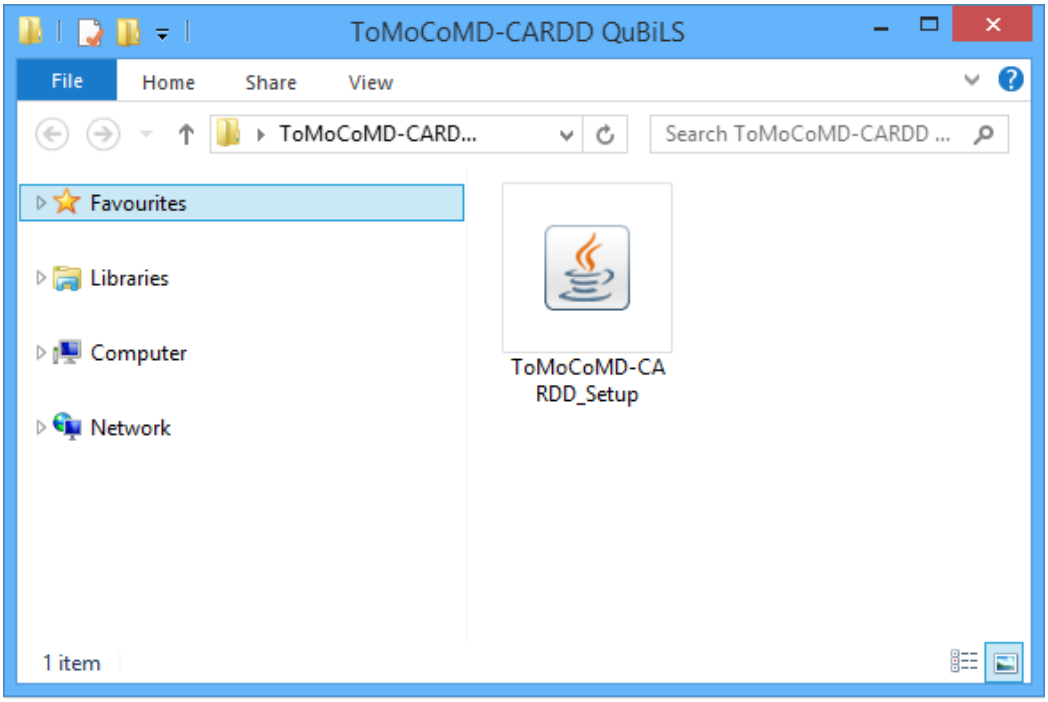

Figure 1. QuBiLS-MAS installer file.

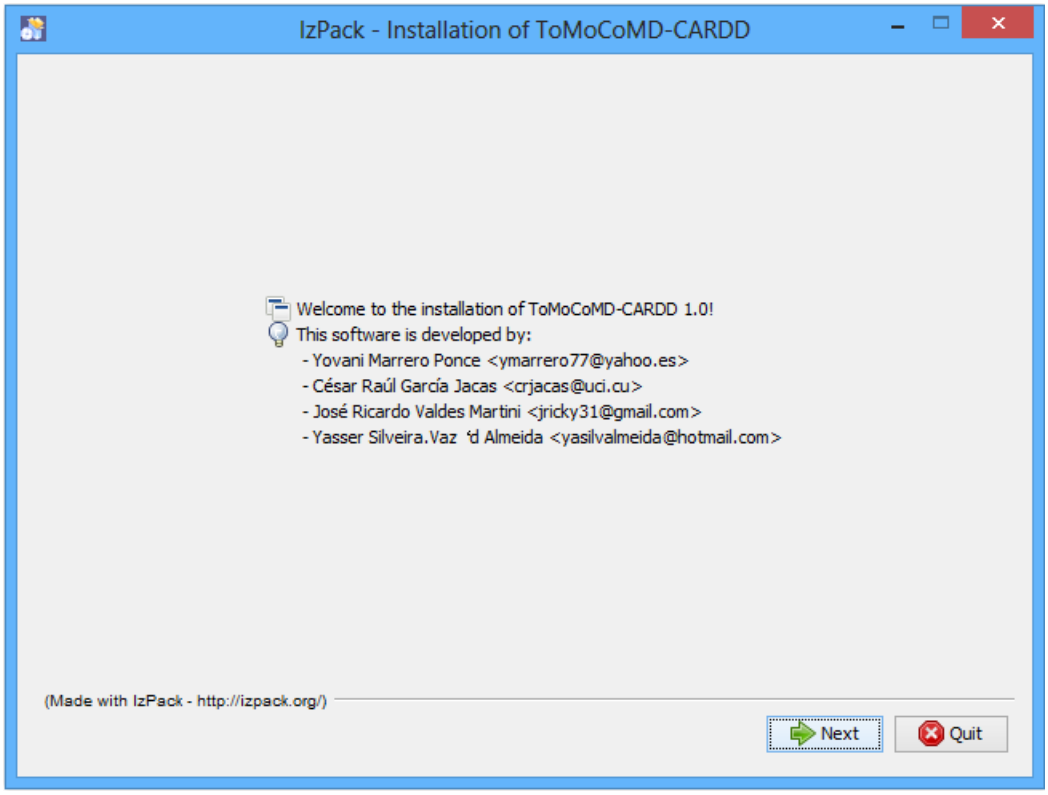

Figure 2. Welcome screen.

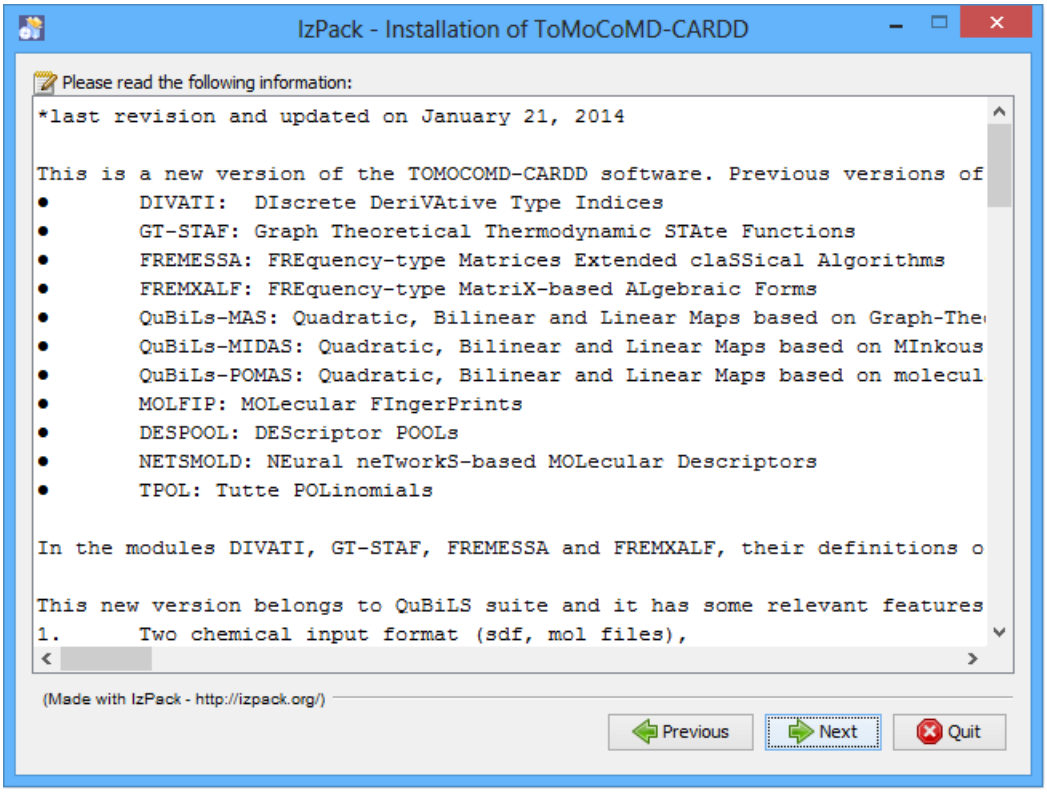

Figure 3. Summary information and notes.

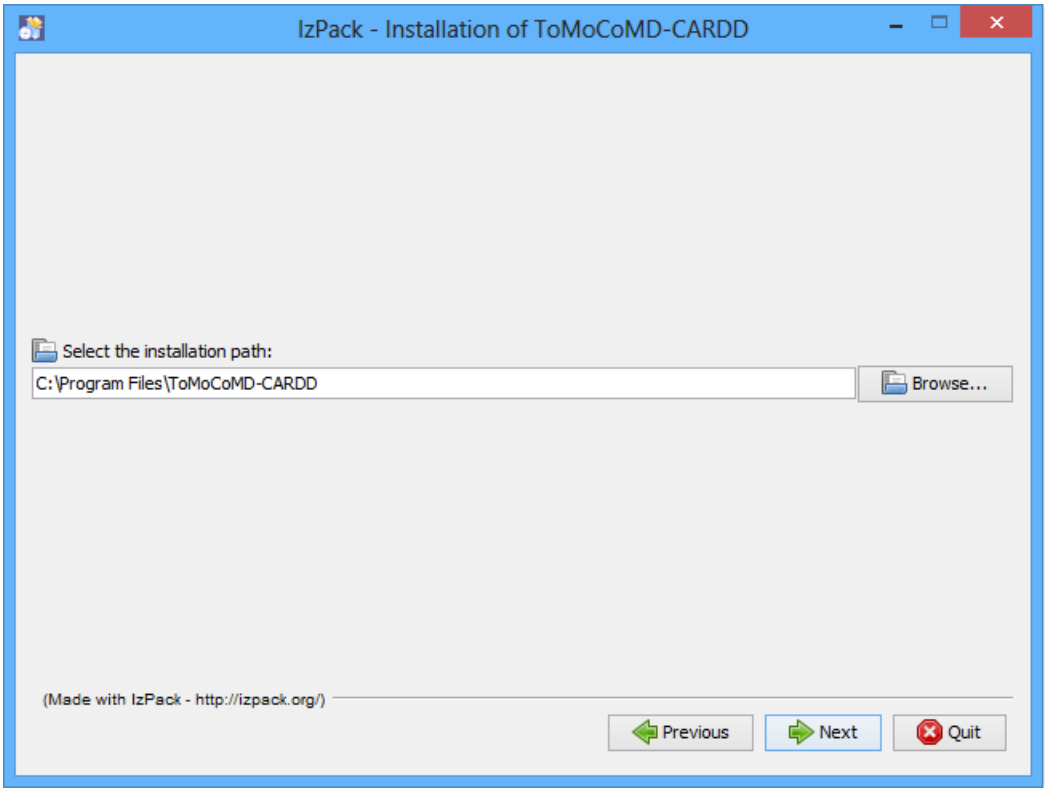

Figure 4. Browse installation path.

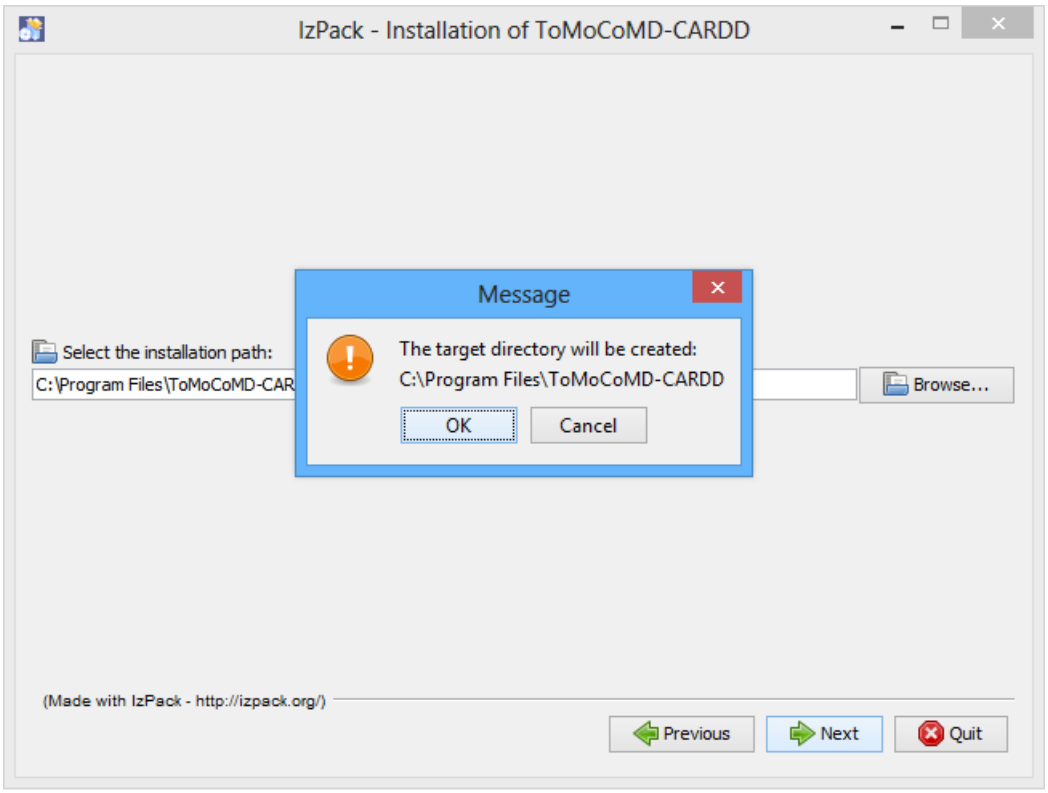

Figure 5. Confirmation for new folder creation.

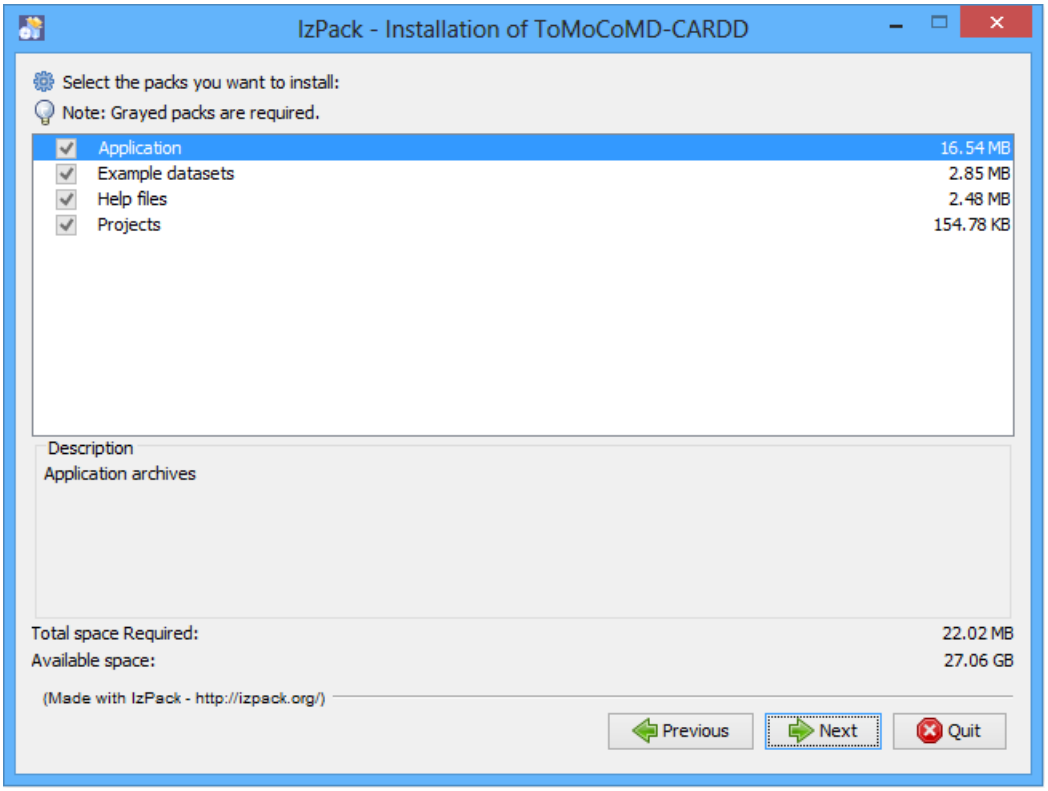

Figure 6. Another necessary files in the installation.

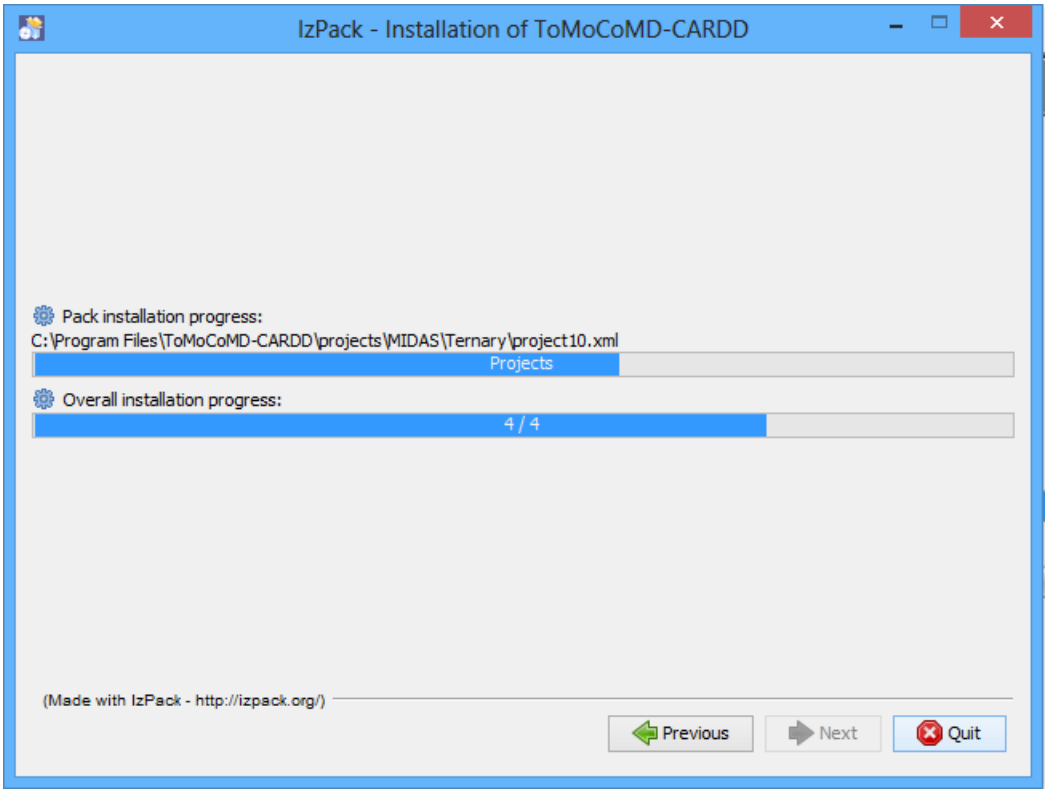

Figure 7. Installation process.

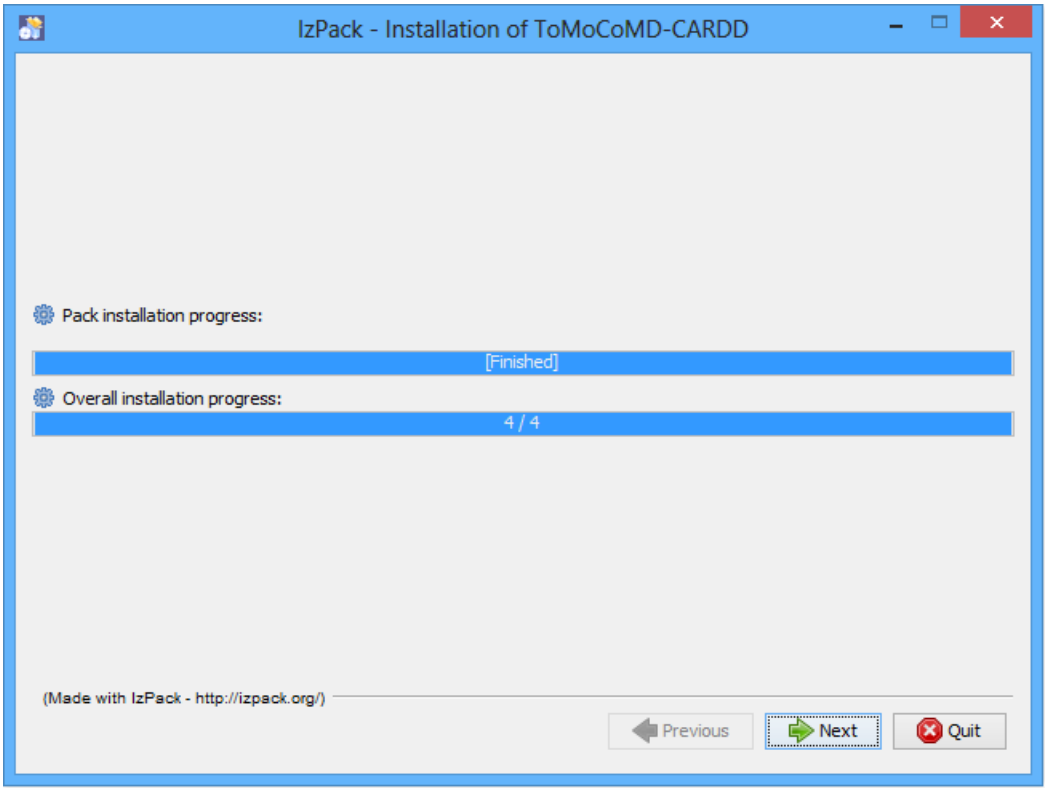

Figure 8. Finish copying files.

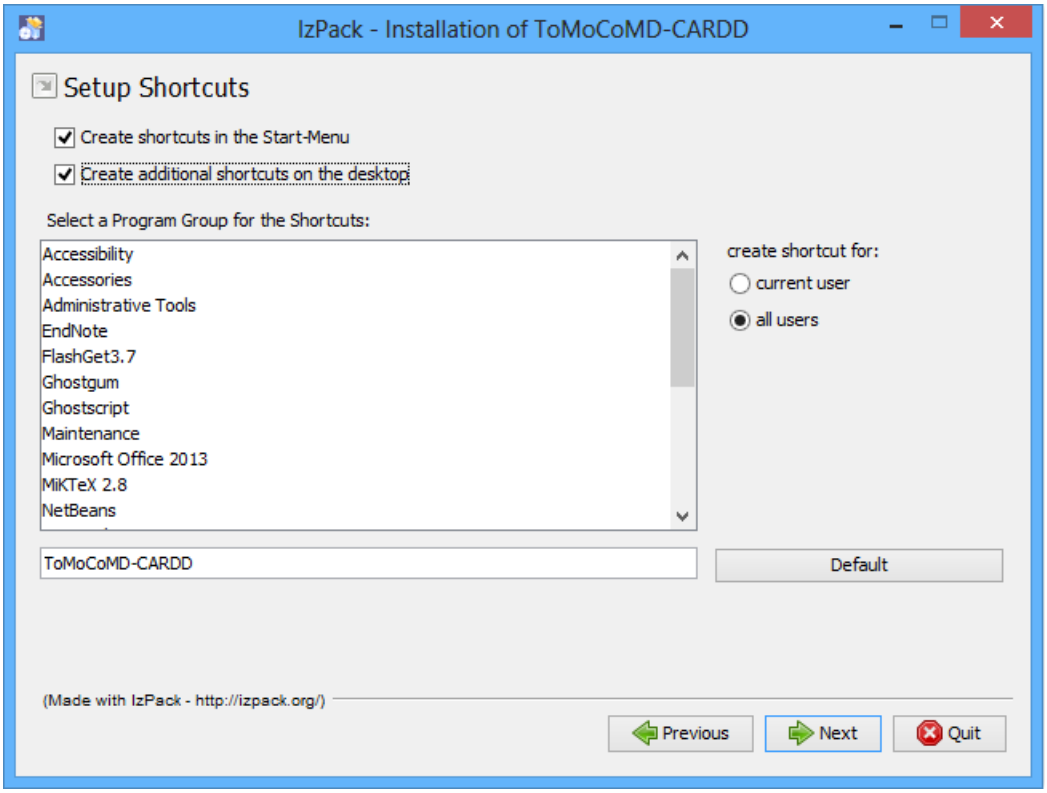

Figure 9. Configure desktop Icon and Start Menu items.

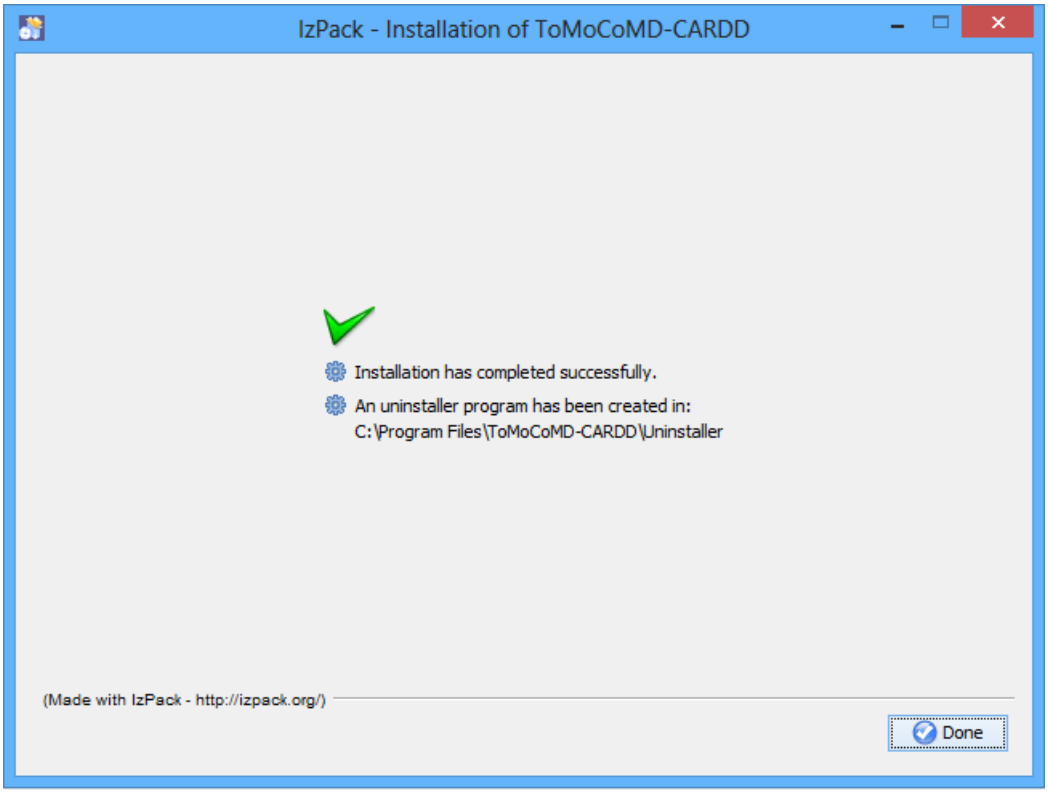

Figure 10. Installation and configuration has been completed successfully.

The ToMoCoMD-CARDD setup program will detect an existing previous version in your computer and offers the choice of uninstalling or keeping the older version of the program. Each version of this program could be uninstalled either by using the option:

"*Control Panel -> Add or Remove Programs*" or by clicking the uninstall shortcut in *"Uninstall" link located in "Start -> All Programs -> QUBILs-MIDAS -> Uninstall*".

If the target operating system is any UNIX/LINUX or MAC platforms, or if the Windows PATH environment variable does not recognize the executable *.jar* file, just execute the corresponding script for launch the installer application, or run directly the Java executable program from the command prompt: (assuming E: is your CD-ROM drive letter):

```
java – jar e:\installer\ToMoCoMD-CARDD_Setup.jar
```

ToMoCoMD-CARDD QuBiLS suite is available in three different forms, these are:

a) A guided installation program: *QuBiLS Setup.jar* *, which allows standard and common chemo-informatics users to deploy the whole program through a friendly step by step GUI, creating easy access shortcut icon in the user's profile Desktop and Start Menu (Windows platform).
b) A Java™ portable application: *QuBiLS Portable.jar,* for users that frequently use different workstations. No matter the operating system or workstation hardware configuration, ToMoCoMD-CARDD QuBiLS users always will have these software to one click away.
c) A self-extracting Windows executable pack: *QuBiLS Portable.exe,* for Microsoft Windows users that don't like installing programs. The self-extracting pack was created with WinRAR® program, also included a Java™ JRE release for both Windows platform x86(32-bits) and x64(64-btis), allowing the researcher to easily distribute for free the QuBiLS-MAS Software between others colleagues over internet and workgroup networks.

* *NOTICE:* Installation program also requires Java™ Runtime Environment on the target system and it is platform-independent. It runs under any host operating system, which supports Java(TM) 7 Runtime Environment, it's also works on X86 and X64 architecture.

3.0 GETTING STARTED

3.0 GETTING STARTED

This section provides a general walkthrough of the system from initiation through exit.

Loading application

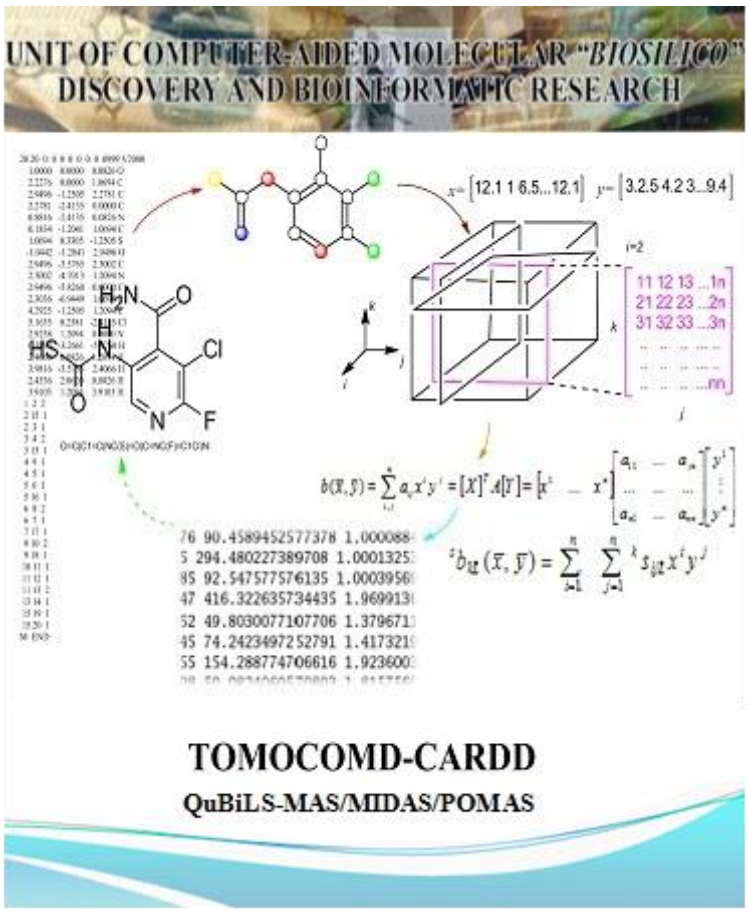

Figure 11. Loading SplashScreen for ToMoCoMD-CARDD (QuBiLS suite)

The software does not require any additional information to login or warm up, as soon as you execute the main program the Splash Screen is launched instantly.

QuBiLS-MIDAS Graphical Visual Interface (GUI)

*The **QuBiLS-MIDAS** GUI has the following screen areas*:

- **Title Bar: Bears the title of program,** ToMoCoMD-CARDD QuBiLS.
- **Menu Bar:** Menus related to different tasks performed by ToMoCoMD-CARDD QuBiLS-MIDAS.

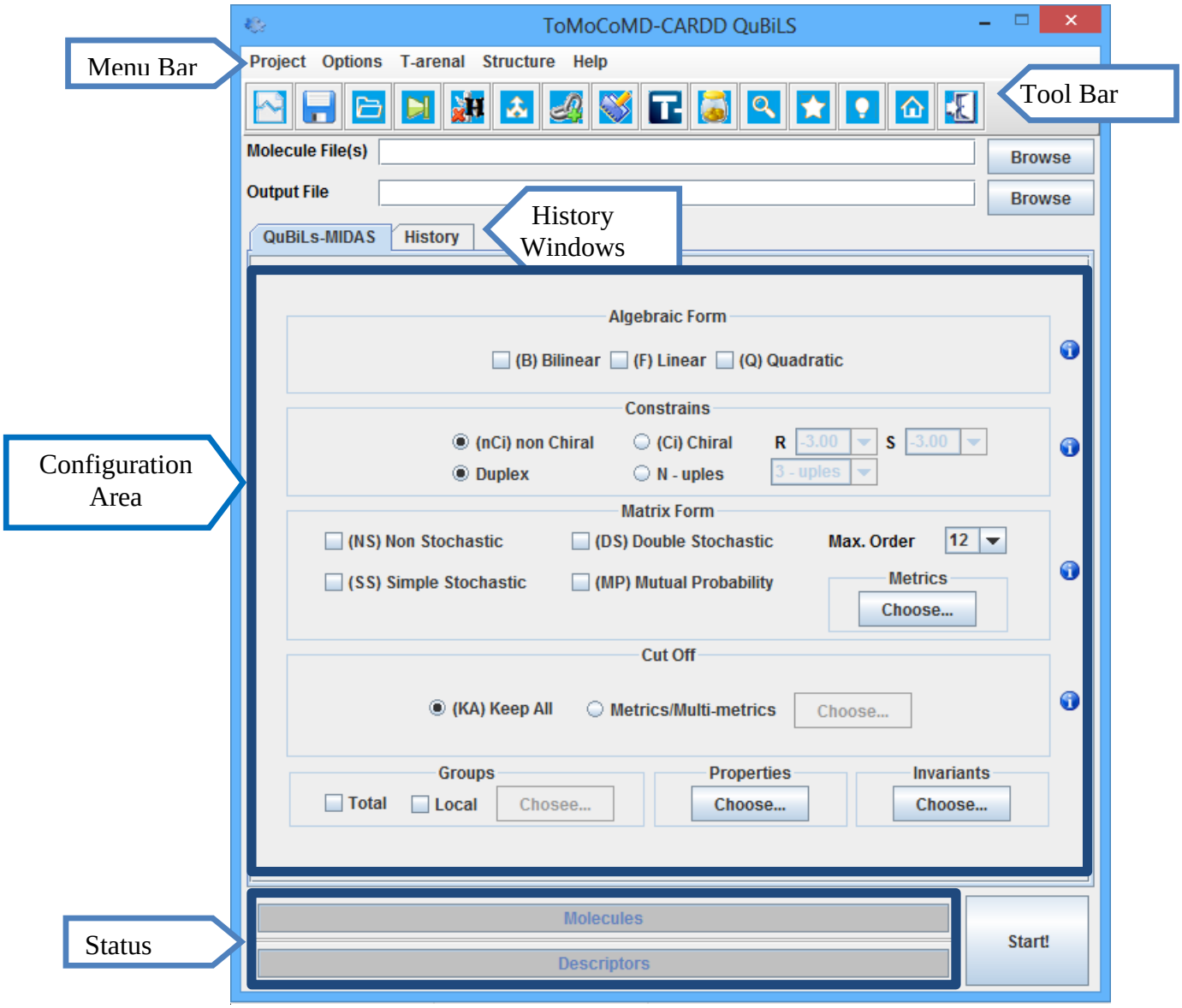

Figure 12. The QuBiLS-MIDAS main GUI

- **Tool Bar:** Quick access shortcuts to commonly performed tasks, displayed as graphical icons instead of classical menu items.
- **Status Bar:** Shows the current and remaining molecules and descriptors.
- **Configuration Area:** This is the *main client area*, which contain MDs pane (Algebraic Form descriptors configuration parameters).
- **History Window:** Logging windows for all operations and task.

System Menu Bar

This section describes in general terms the system menu first encountered by the user, as well as the navigation paths to functions noted on the screen. Each system function should be under a separate section header.

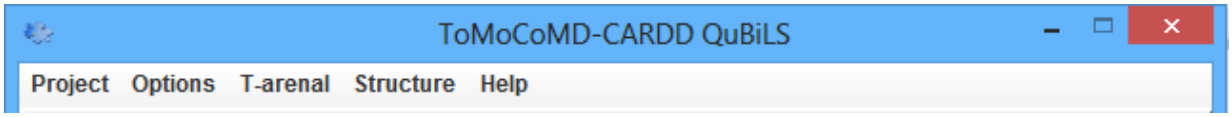

Figure 13. System Menu Bar

Project menu commands

Commands of the *Project menu* allow the user to create configurations and to open and edit existing project files.

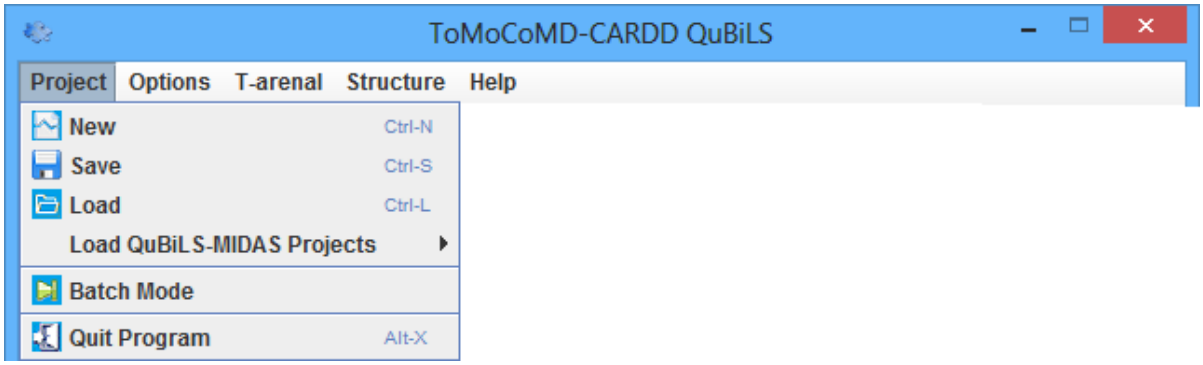

Figure 144. The Project menu

New

Creates a brand new empty configuration or Reset the currents configuration options of the Algebraic Form Panel.

Save as

Export the current configuration and options to a persistent Project Configuration File.

Load

Import configuration and options from a Project Configuration File. That is to say, opens and edit existing project files.

Load QuBiLS-MIDAS Projects

Import configuration and options from a Project Configuration File previously created for the best parameter combinations of Duplex, Ternary and Quaternary indices.

Batch Mode

Launch the Batch Process Manager Windows

Quit Program

Safely close application with saving option prompt (terminates the current **QuBiLS** session).

Options menu commands

Commands of the *Options menu* enable the user to set up the next molecular descriptors computation.

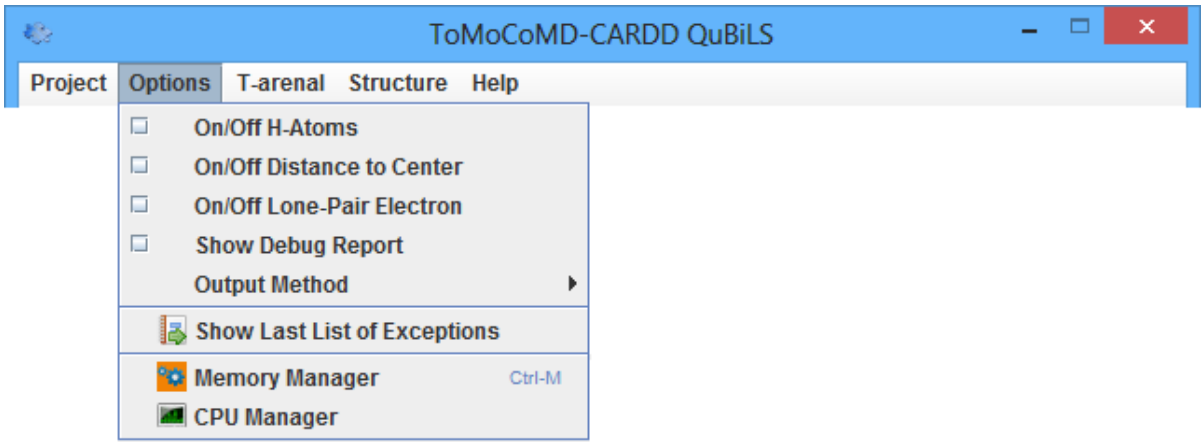

Figure 155. The Option menu

On/Off H-Atoms

Activates or deactivates the hydrogen atoms consideration option in the descriptors calculation process.

On/Off Distance to Center

Activates or deactivates the calculation of the distance of each atom to molecule center in the descriptors calculation process.

On/Off Lone Pair Electrons

Activates or deactivates the Lone Pair Electrons consideration option in the descriptors calculation process.

Show Report

If you check this option, the program generates a new text file with all information concerning the algebraic process that takes place in the calculation.

Output Method

Display the available options to format resulting file with calculations of indices, we can only select one option at a time.

Show Last List of Exceptions

Display the exceptions occurred during the calculation of the molecular indices.

Memory manager

Display a window where is shown the RAM employed by the program.

CPU manager

Display a window to set the amount CPU cores to use in the calculation of the molecular indices.

T-arenal menu commands

The commands of the *T-arenal menu* permit to the end user to perform the calculation of the QuBiLS-MIDAS molecular descriptors on the T-arenal distributed computing system, to monitor the progress of the distributed computation and download the solutions obtained.

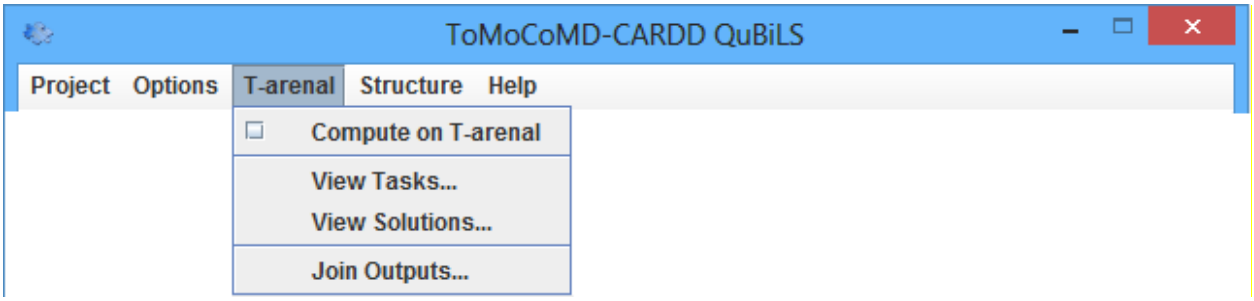

Figure 16. The T-arenal menu

Compute on T-arenal

Indicate to the QuBiLS-MIDAS software that the configured molecular indices will be computed on the T-arenal system.

View Tasks…

Show all calculations of molecular indices that are being performed on T-arenal system.

View Solutions…

Show all solutions obtained as from the distributed processing of the molecular descriptors.

Join Outputs…

Join into an only output file the independent outputs saved during the distributed processing time. This last is due to the fact that as T-arenal is a non-dedicated system, the results obtained as from the decomposition of the original calculation do not arrive in the same order in that the respective calculations were created and thus, each received result is saved of independent way.

Structure menu commands

The commands of the *Structure menu* permit to the client to perform data cleaning tasks, add or remove hydrogen atoms and convert to the MDL MOL/SDF formats.

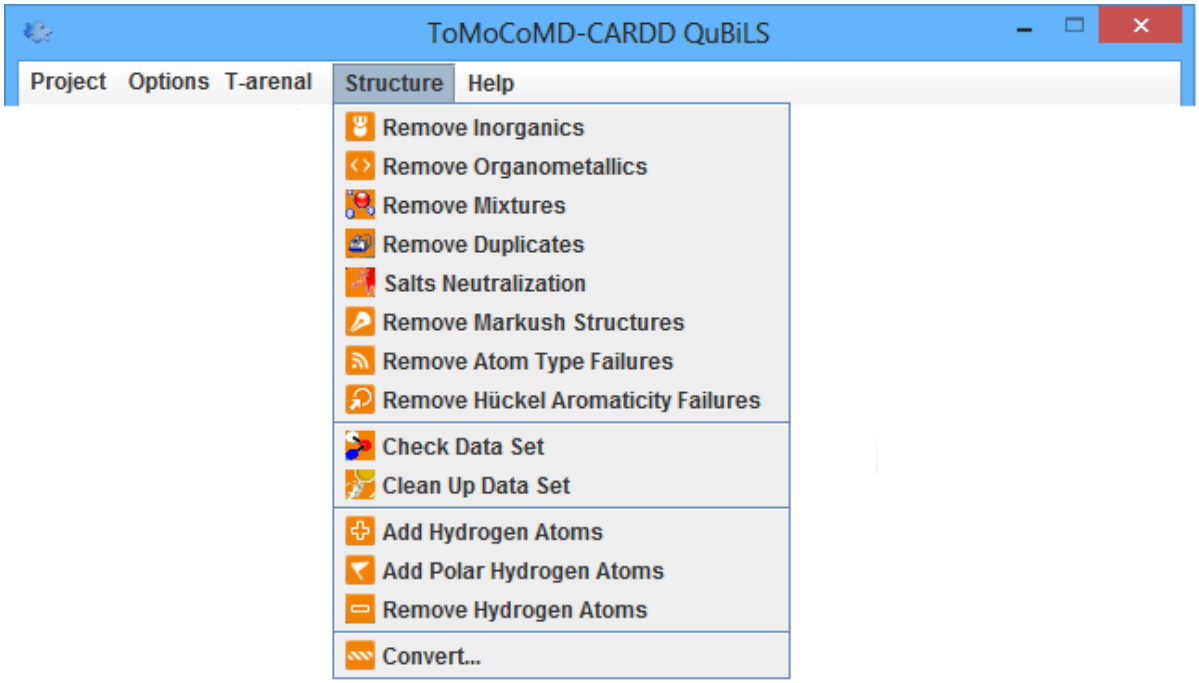

Figure 17. The Structure menu

Remove Inorganics

Remove, from the chemical structure datasets specified by the user, the compounds that not contain carbon atoms.

Remove Organometallics

Remove, from the chemical structure datasets specified by the user, the compounds that contain metallic elements.

Remove Mixtures

Detect, from the chemical structure datasets specified by the user, the compounds that constitute mixtures and then retain of this mixture the component with the highest molecular weight.

Remove Duplicates

Detect the compounds that have the same CANONNICAL SMILES and then these are removed from the chemical structure datasets specified by the user.

Salts Neutralization

Detect and neutralize, from the chemical structure datasets specified by the user, the compounds that contain salts.

Remove Markush Structure

Detect the *Markush* structure in the compounds and these are removed from the chemical structure datasets specified by the user.

Remove Atom Type Failures

Analyze if the atoms of each chemical structure are recognized by the program. If there is some error during this process then the corresponding compound is removed from the datasets specified by the user.

Remove Huckel Aromaticity Failures

Check the existing of aromatic rings in the chemical structures. If there is some during this process the troublesome compound is removed from the datasets specified by the user.

Check Data Set

Check the chemical structure datasets specified by the user through the application of the options previously mentioned. It is not mandatory the use of all data cleaning operations, so the user can select which of these options to apply. The results are saved in a log file.

Clean Up Data Set

Clean the chemical structure datasets through the application of the previous data cleaning options.

Add Hydrogen Atoms

Add the hydrogen atoms to the compounds belonging to the chemical structure datasets specified by the user.

Add Polar Hydrogen Atoms

Add the hydrogen atoms to the atoms different to the carbon in the compounds belonging to the chemical structure datasets specified by the user.

Remove Hydrogen Atoms

Remove the explicit hydrogen atoms in the compounds belonging to the chemical structure datasets specified by the user.

Convert…

Provide functionalities to convert:

- SMILES format to SDF/MOL format.
- MOL format to SMILES/SDF format
- SDF format to SMILES/MOL format
- PDB format to SMILES/SDF/MOL format
- MOL2 format to SMILES/SDF/MOL format
- XYZ format to SMILES/SDF/MOL format
- InChI format to SMILES/SDF/MOL format

Help menu commands

Moreover, the **QuBiLS-MIDAS** main window contains some icons which can be clicked in order to obtain specific information. The *Help menu* contains the following commands:

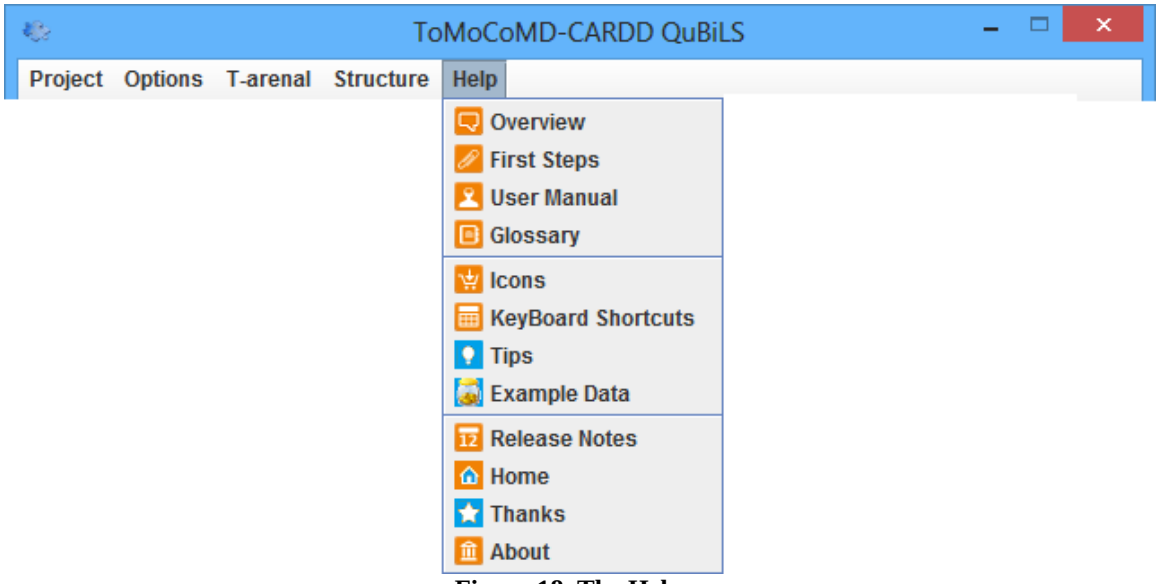

Figure 18. The Help menu

Overviews

Shows an illustrative procedure of how to execute, configure and use the program. QuBiLS module is based on Chemistry Development Kit (CDK) library.

About the Chemistry Development Kit (CDK).

The CDK an open-source library of algorithms for structural chemo- and bio-informatics, implemented in the programming language Java. It serves as a base for many other applications, including some parts of **QUBILs** software. For information about CDK, please visit the CDK home page. The CDK library is published under terms of the GNU Lesser General Public License. This project is hosted under http://cdk.sourceforge.net.

Copyright: The CDK is copyrighted by the CDK project, and has been written by Rich Apodaca, Ulrich Bauer, Miguel Rojas Cherto, Fabian Dortu, Martin Eklund, Matteo Floris, Dan Gezelter, Uli Fechner, Rajarshi Guha, Yonquan Han, Thierry Hanser, Tobias Helmus, Kai Hartmann, Christian Hoppe, Oliver Horlacher, Miguel Howard, Violeta Labarta, Nina Jeliazkova, Geert Josten, Anatoli Krassavine, Stefan Kuhn, Daniel Leidert, Edgar Luttmann, NathanaÃ«l Mazuir, Stephan Michels, Peter Murray-Rust, Irilenia Nobeli, Chris Pudney, Jonathan Rienstra-Kiracofe, David Robinson, Bhupinder Sandhu, Jean-Sebastien Senecal, Sulev Sild, Bradley Smith, Christoph Steinbeck, Stephan Tomkinson, Joerg Wegner, Stephane Werner, Egon Willighagen, and Yong Zhang.

First Steps

Introduce a first time **ToMoCoMD-CARDD** user into a general overview, system requirements and installation process, input and output file modes.

User Manual

Open the QuBiLS-MIDAS User Manual.

Glossary

Basic **QuBiLS** molecular descriptors terminology is provided in a tool. This terminology is provided in a window that the user can keep open to be supported through the **QuBiLS** indices setup.

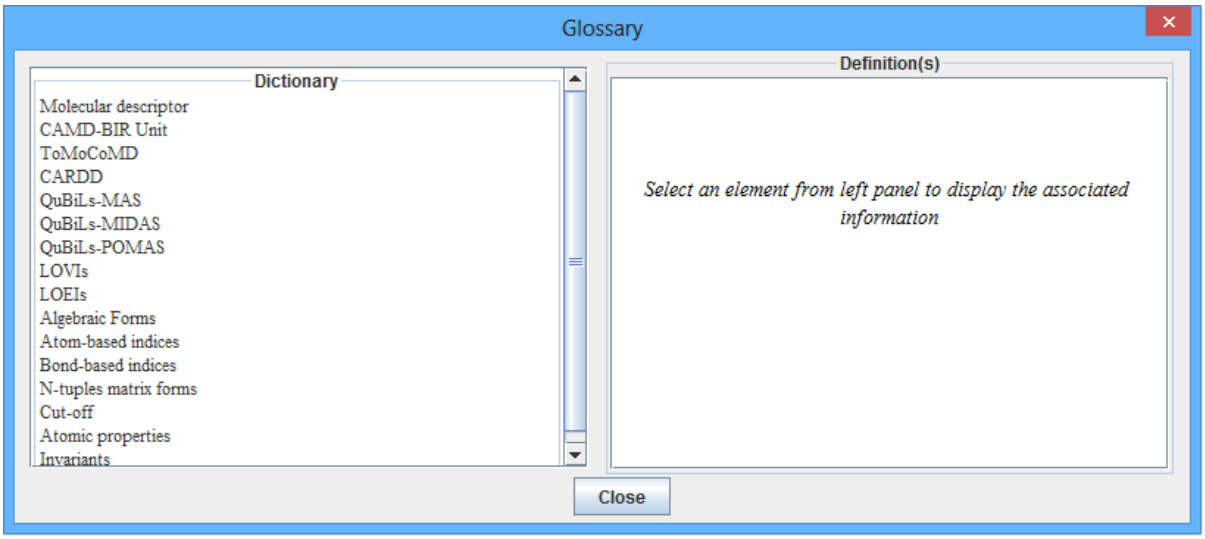

Figure 19. Terms and Concepts Glossary Tool.

Icons

The functionality of all Icon used in the program is described, so that the user learns the meaning of **ToMoCoMD-CARDD's** icons naturally.

Keyboard Shortcuts

Describe all keyboard accelerators used in the program. These keyboard shortcuts perform the specific commands or replace the equivalent menu items.

Tips

Most frequent first user questions are answered in this section, with 14 tips that provide instant technical support available any time you the program is run.

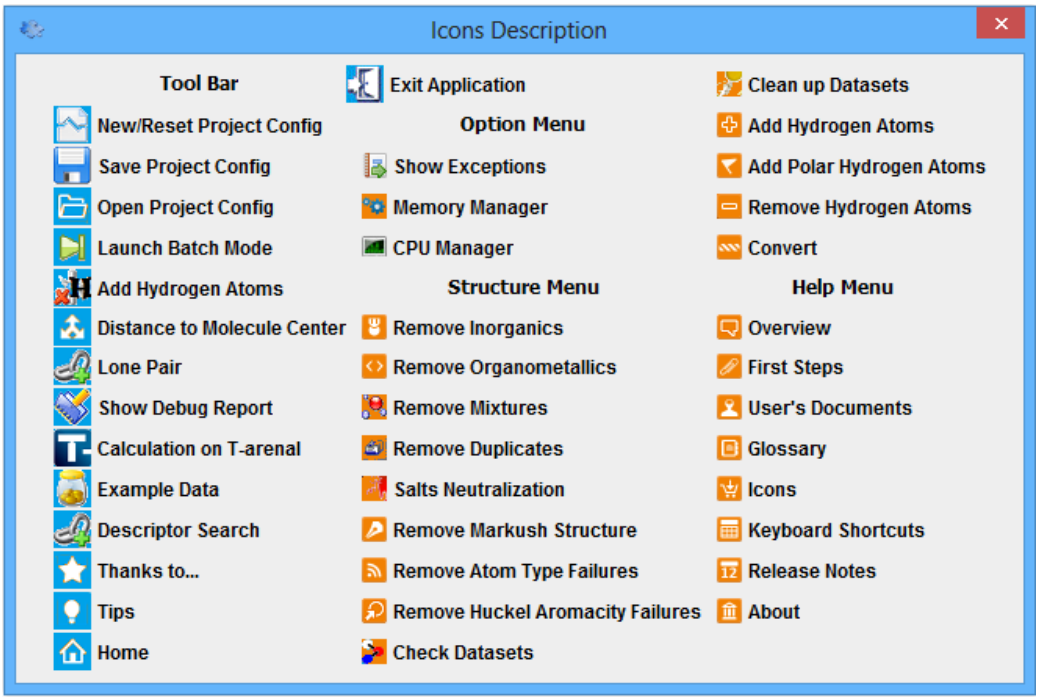

Figure 20. Icons description.

Example Data

The Example Data Tool is a key element for **ToMoCoMD-CARDD's** first users, in order to test the MDs calculated by this software, six molecular datasets of are provided. Click the example data icon in the tool bar to access these molecular datasets. This datasets will permit to make simple test calculations. To perform descriptor calculations, click the respective checkboxes to select the desired configuration options.

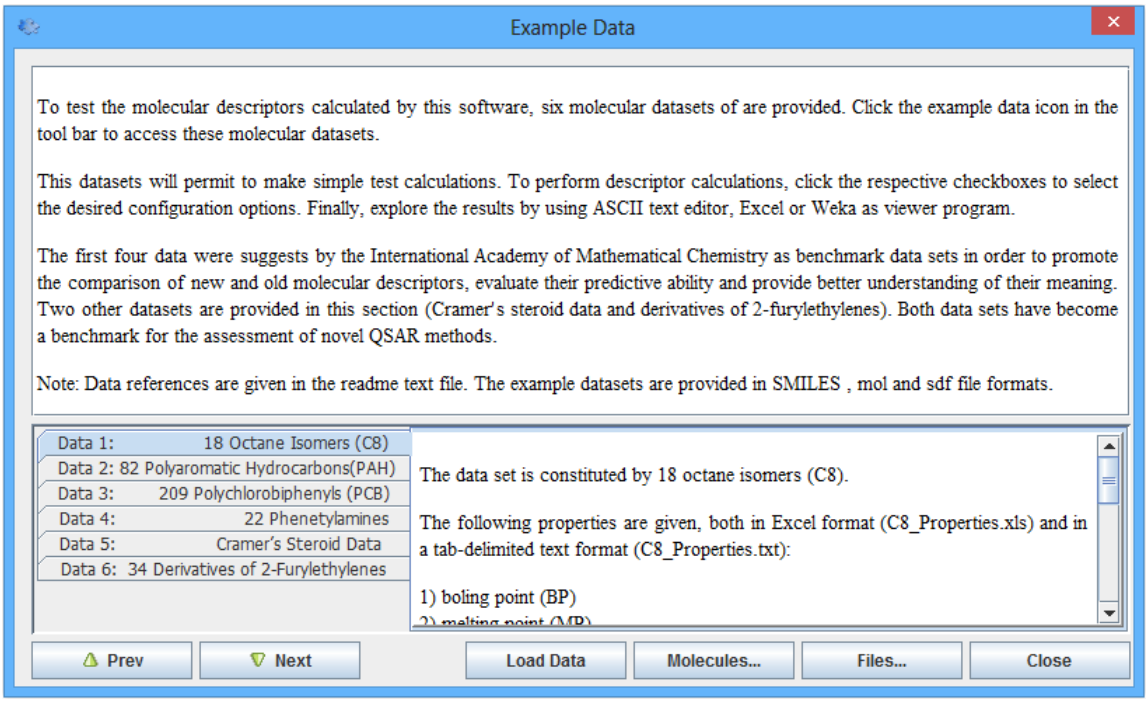

Figure 21. Example Data Tool.

Release Notes

Show the track changes since **QuBiLS-MIDAS** was a whiteboard idea.

Home

Bring useful information about **CAMD-BIR Unit**, how to contact us and cite.

Thanks

Recognition to different contributions to the success of this project is offered.

About

Several information about **QuBiLS-MIDAS** software and publications.

Tool Bar

Quick access shortcuts for most relevant option and tools. That is, the toolbar icons replace the most important and frequently used **QuBiLS** menu commands. Clicking on toolbar icons enables the user to perform the following commands:

Figure 22. Tool Bar elements.

1. New
2. Save
3. Open
4. Batch Process Manager
5. On/Off H Atoms
6. On/Off Distance to Molecule Center
7. On/Off Lone Pair Electron
8. On/Off Generate Debug Report
9. On/Off Distributed Calculation on T-arenal System
10. Launch Example data
11. Descriptor Search Tool
12. Thanks
13. Tips
14. Home
15. Exit Program

Descriptor Search

ToMoCoMD-CARDD includes an optional use tool developed to automatically decode headers assigned to each one of the molecular indices. *See the picture bellow*.

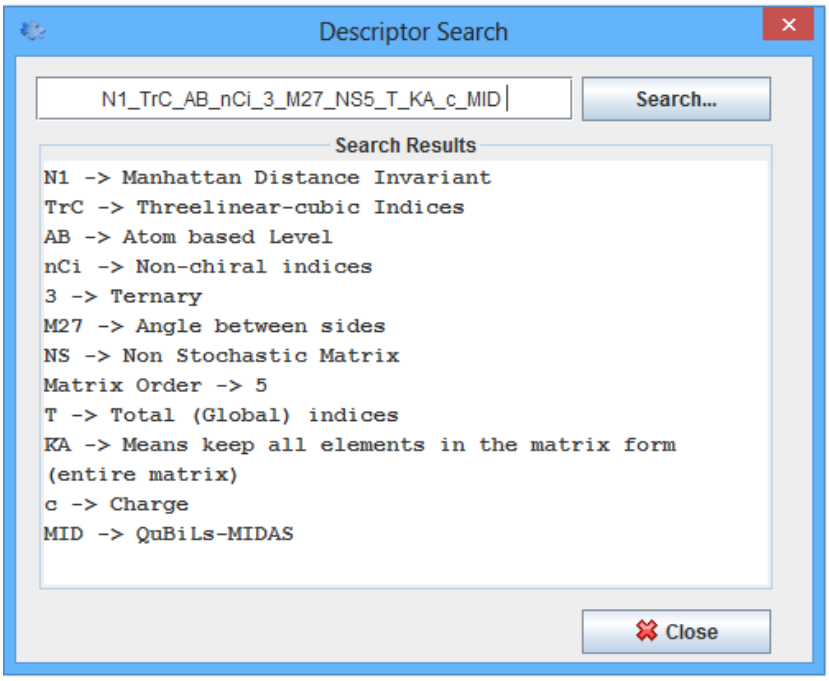

Figure 23. Dialog Search Window.

Status Bar

The status bar located at the bottom of main windows shows the molecule and the algebraic descriptor that is being calculated and the percentage of completion, also displays the name and number for molecules and descriptors.

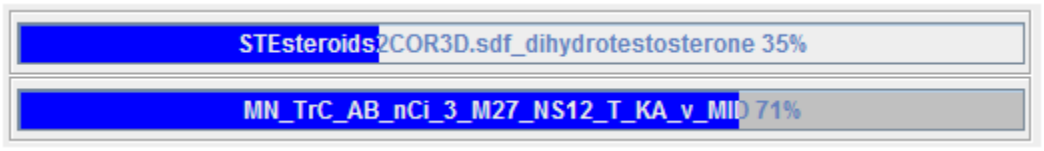

Figure 24. Status Bar.

Configuration Area

This area has seven sub-areas (panes) for MDs configuration. In each pane appears an info button (blue circle), where there exist a short explication regarding with theory include in each parts (for more details see **Starting QuBiLS-MIDAS** and **Configuring a project** sections).

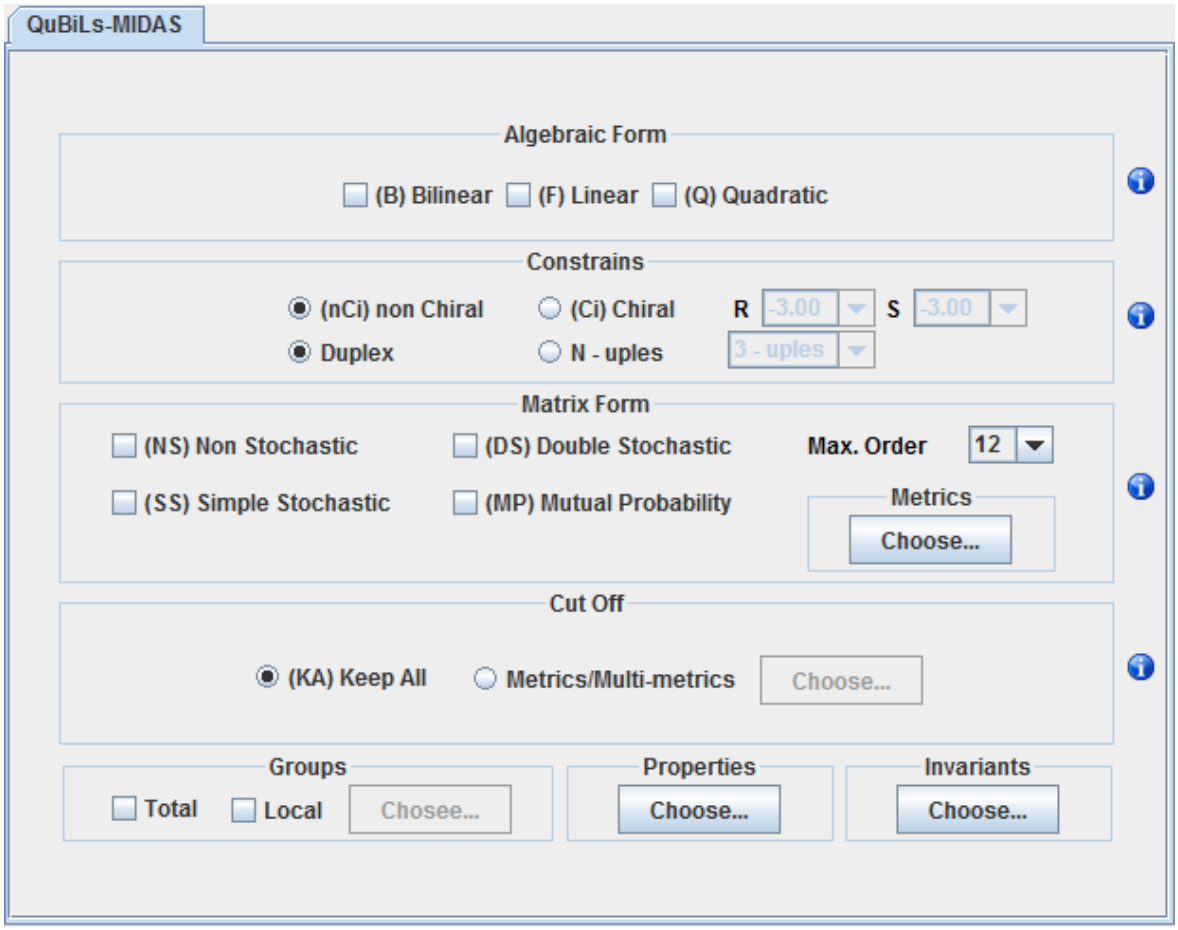

Figure 25. Algebraic Descriptors Configuration Area.

History Window

Logging windows for all operations and task. Besides, after the calculation is finished, the **tab History** (**log file**) shows some details and statistics of the calculation process.

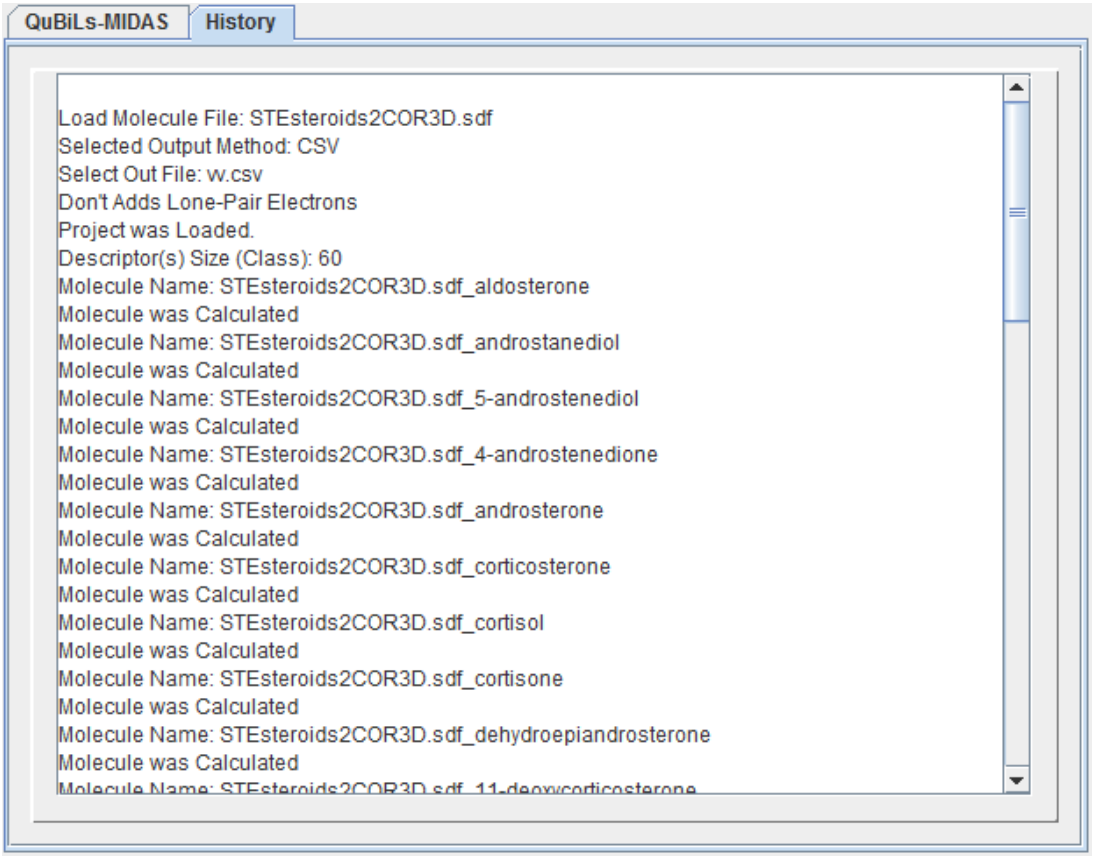

Figure 26. Logging (History) windows.

Exit System

Describe the actions necessary to properly exit the system.

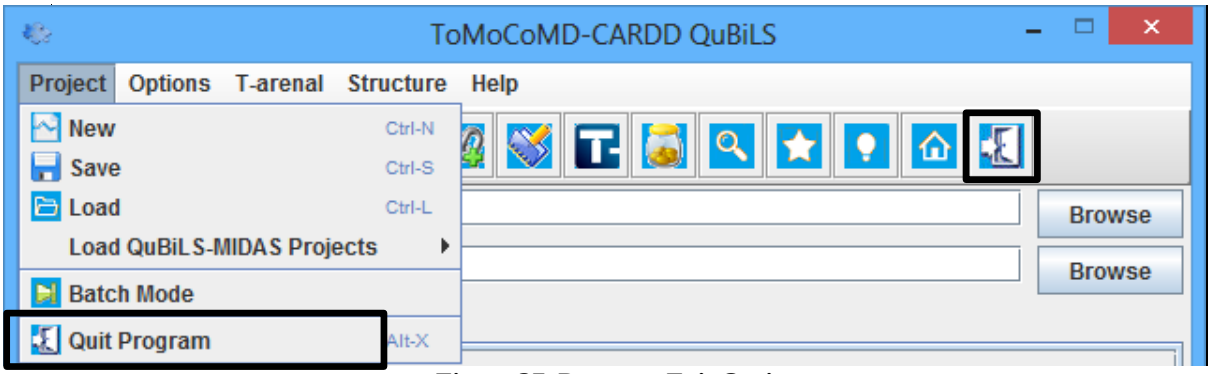

Figure 27. Program Exit Options.

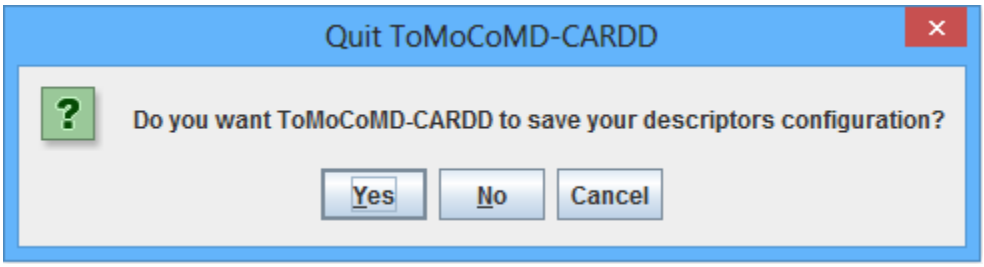

Figure 28. Safe Exit Action prompt.

4.0 USING THE SYSTEM

4.0 USING THE SYSTEM

This section provides a detailed description of the **QuBiLS-MIDAS** Software from the initial to the final steps, explaining in detail the characteristics of the required input and system-produced output. It covers both calculations of single and multiple molecular datasets and batch mode calculations. Each **QuBiLS-MIDAS** function is under a separate section header, and corresponds sequentially to the system functions (menu items) listed in subsections of chapters above.

What you need to know before using QuBiLS-MIDAS

If a molecule contains "exoteric" features (such as: some atom types not fixed, disconnected structure, radicals, and so on) it will not be recognized and therefore it will be rejected during descriptor calculation (see *File Created for QuBiLS-MIDAS* and/or *Special Instructions and Exceptions* sections). To make use of QuBiLS-MIDAS calculations, 3D optimized structures are required.

QuBiLS-MIDAS is not designed as QSAR software; it provides only molecular indices and does not perform QSAR analysis. However, by **QuBiLS-MIDAS** it is possible to merge calculated molecular descriptors and user-defined properties for a set of molecules, providing a complete output file which is easily loaded by any correlation analysis application. **QuBiLS-MIDAS** provides a total of **8 604 960** molecular descriptors based on atom-pairs relations, **138 640 320** molecular descriptors based on ternary relations among atoms and **286 191 360** molecular descriptors based on quaternary relations among atoms. These descriptors are not *chiral* and taking into consideration only *non-standardized Invariants*. In addition, in the **QuBiLS-MIDAS software** could be used N-tuple cut-offs to considerer the most important non-covalent interactions in a molecular structure according to criteria based on topological and/or geometric distances, as well as, to take into account the number of lone-pairs or the distance to molecule center as diagonal coefficients of the matrix forms. Also could be added the hydrogen atoms during the calculation of the molecular descriptors. Finally, several projects are provided, which were built from the best results obtained in *"in-house"* comparisons.

Starting QuBiLS-MIDAS

QuBiLS-MIDAS is launched by clicking on the configuration files (e.g. .bat files on Microsoft Windows platforms) that are provided. These files allow to improve the performance and speed. These are tweaks for the Java™ VM (JVM), that increase the maximum default limit of JVM heap memory. These preconfigured scripts are located in the root directory of QuBiLS-MIDAS program folder. The configurations of JVM heap memory limit are:

- 1 GB
- 2 GB
- 4 GB
- 8 GB
- 16 GB
- 32 GB

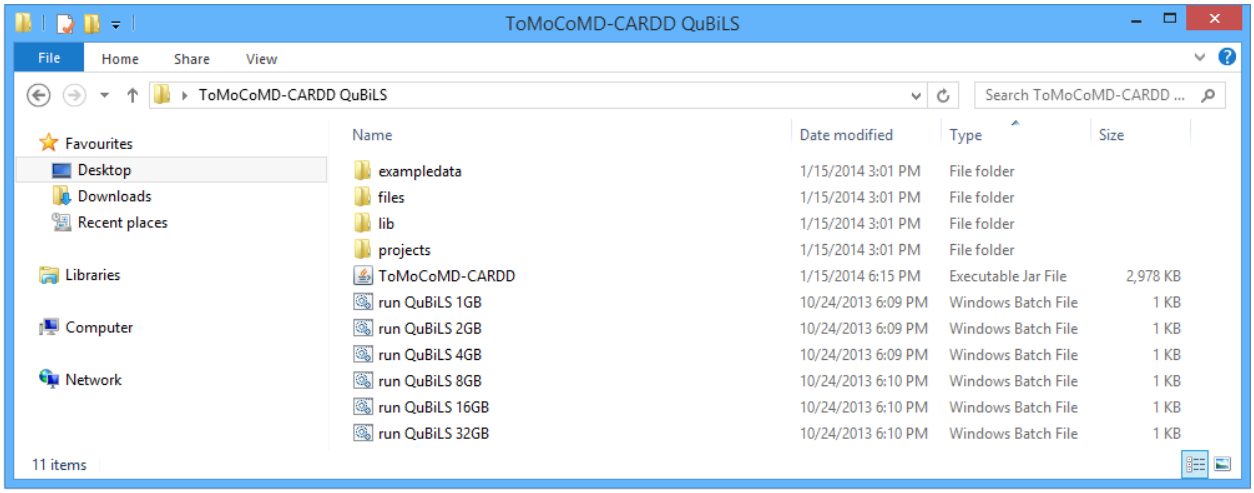

Figure 29. QuBiLS-MAS Windows Batch Files (.bat).

Indeed, for each heap memory limit, a command line scripts were targeted for two different kind of platform:

- Windows Command Script *(.cmd*) and Windows Batch File *(.bat*)
- Linux Shell Script *(.sh*).

Otherwise, if the preconfigured command line scripts do not suit your hardware preferences, users can modify the scripts for both platforms to adjust the program JVM heap memory limit according to their system hardware properties, editing these scripts with a text editor program, (i.e. *NotePad* or *WordPad* in Windows, and *GEdit* or *Vi* in Linux). The following example limits de JVM heap memory up to 1024 megabytes:

```
java -Xms256m -Xmx1024m -jar ToMoCoMD-CARDD QuBiLS.jar
```

After splash, the main window (GUI) will be displayed on the screen. The follow step is selecting the ***structure input*** and ***descriptors output*** files by pressing the browse button in the *Input and Output* section in the upper right part of the GUI. Next, the user can select the desired descriptors for calculation (see configuring a new project below). Finally, the button "**Starts!"** begins the calculation of the selected descriptors for the structures in the input file. After the calculation is finished, an external dialog window appears that shows a message about the successful calculation. This message can be closed by pressing the **Ok** button. Then, the **Exceptions Window** is come into view. This window depicts the list of molecules with structure errors. In addition, the **History Tab** (see *Logging Window*) shows some details and statistics of the run.

Starting a new project

The **New Option** clean all parameters that are used during a **QuBiLS-MIDAS** session, *e.g.*, algebraic forms, matrix forms, H-atoms, Lone-pair electrons and so on.

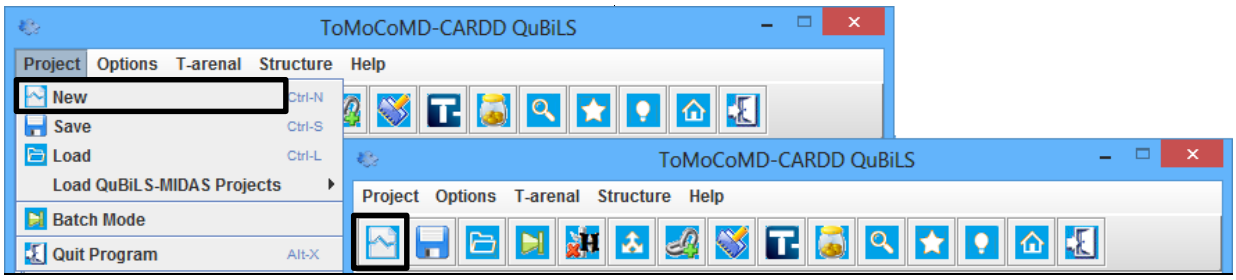

Figure 30. Creating a new project.

Saving a project file

A project file is saved in the menu **Project** in the main menu bar by the menu item **Save**. A dialog box appears where the path and the file name of the project file can be specified.

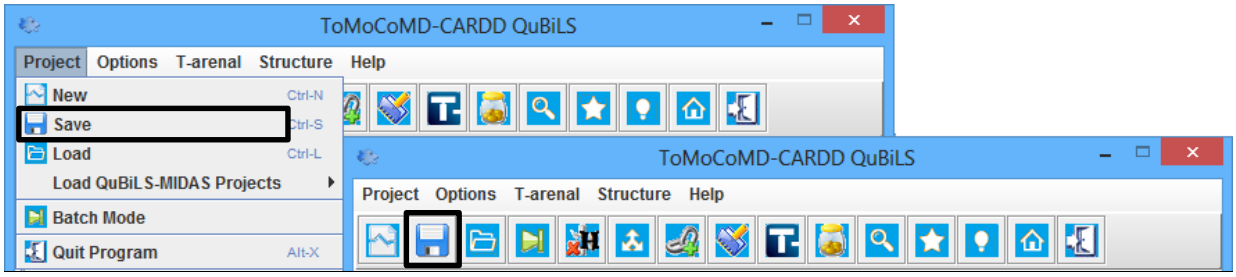

Figure 31. Saving project configuration.

Loading a Project File

Project files can be reloaded in order to restore all parameters in a later session or used to execute **QuBiLS-MIDAS** in batch mode.

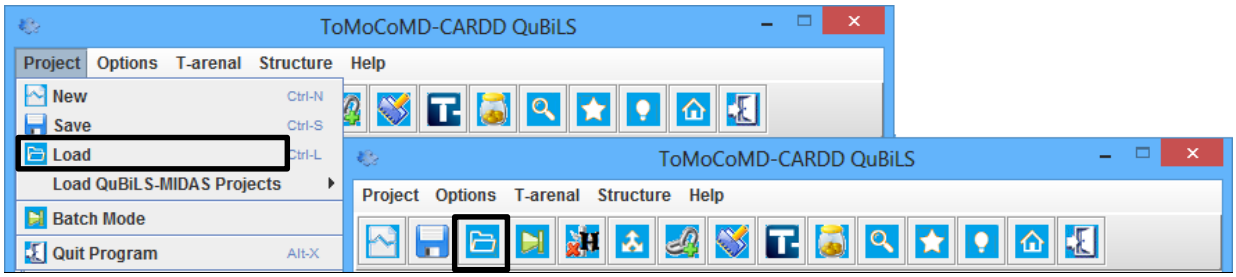

Figure 32. Loading projects button.

Running a saved project

The previously saved project can be reloaded in order to restore all parts and parameters of a previous session. Finally, the user can run the calculation by clicking the button **Start!** After the calculation is finished, an external dialog window appears that shows a message about successful completion of the calculation. This message can be closed by pressing the **Ok** button.

Program Run Options

The following is a description of the **Calculate** section of the **QuBiLS-MIDAS** GUI. The Calculate section consists of the three buttons **Start! / Cancel**, **Exception Window** and **History Tab**. The button **Start! / Cancel** begins the calculation of the selected descriptors. When the calculation is started the **Start!** button switches to **Cancel**, allowing it to stop the process. Unless the **Cancel** button is not pressed or the calculation is finished, all remaining buttons and menus are enabled, in case user needs to review the descriptor configuration or access the *Tool Bar* and *Menu Bar* options. In addition, a progress bar appears that shows the molecule and the algebraic descriptor that is being calculated at a given time and the percentage of completion, also the name and number for molecules and descriptors are displayed. When the calculation is finished a message appears on the screen that displays "***The Process was successfully finished.***"

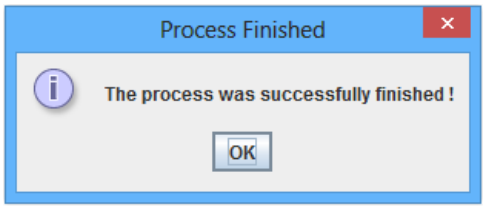

Figure 33. Task finished confirmation windows.

Distributed calculation on T-arenal system

Some molecular descriptors calculation are computationally complex due to the large compound datasets to analyze. In this sense, there are several computational alternatives to tackle this situation such as the distributed computation. Bearing this in mind to the QuBiLS-MIDAS software an option to perform distributed calculation of molecular descriptors was added.

Run calculation on T-arenal system

To perform a distributed calculation on the T-arenal system is only necessary to choose in the toolbar or in the T-arenal menu the corresponding option (see Figure 34). Subsequently, the button **Start!** is pressed to connect to the T-arenal system and send the configuration of the molecular indices (see next Section) and the compound datasets selected by the end users.

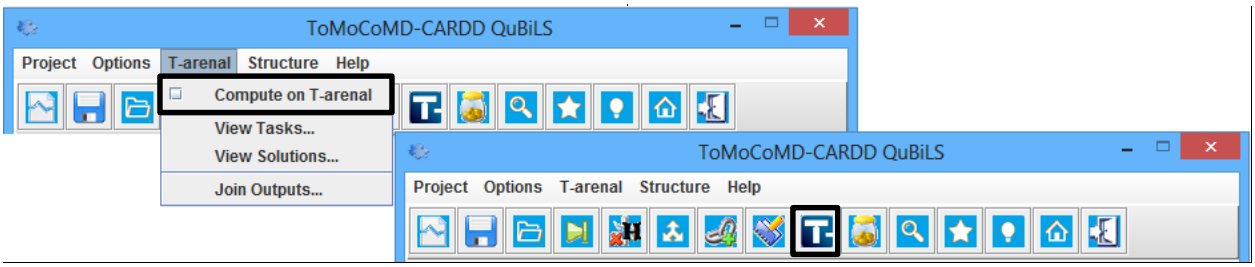

Figure 34. T-arenal button to perform distributed calculation.

If the end user is not authenticated into T-arenal system a login window is shown. In this window dialog the username, password, IP address and port of the T-arenal server are specified. Once the end user is authenticated a dialog window to set up the input parameters of the calculation

is shown as well. Finally, when the input parameters are stated the computation configuration is sent to T-arenal.

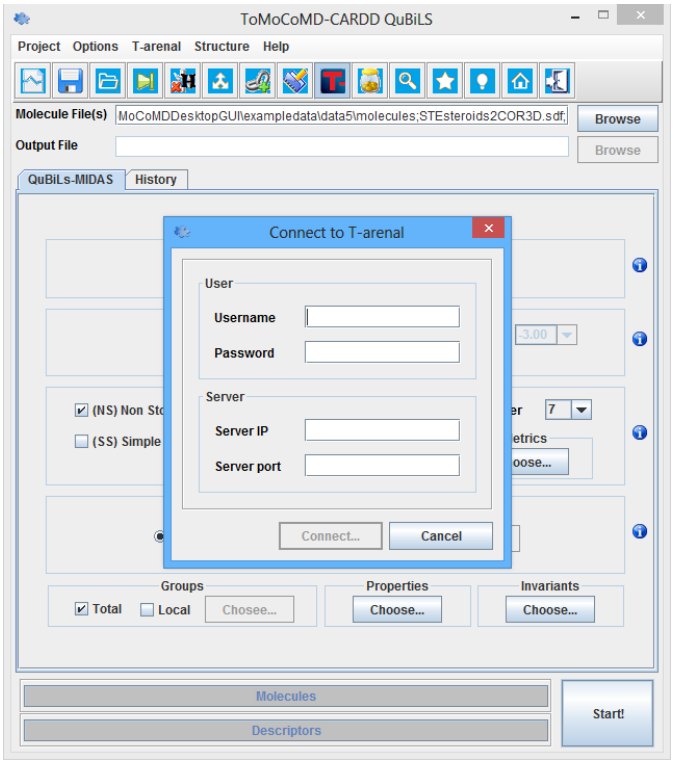

Figure 35. Dialog window to authenticate the end user into T-arenal system.

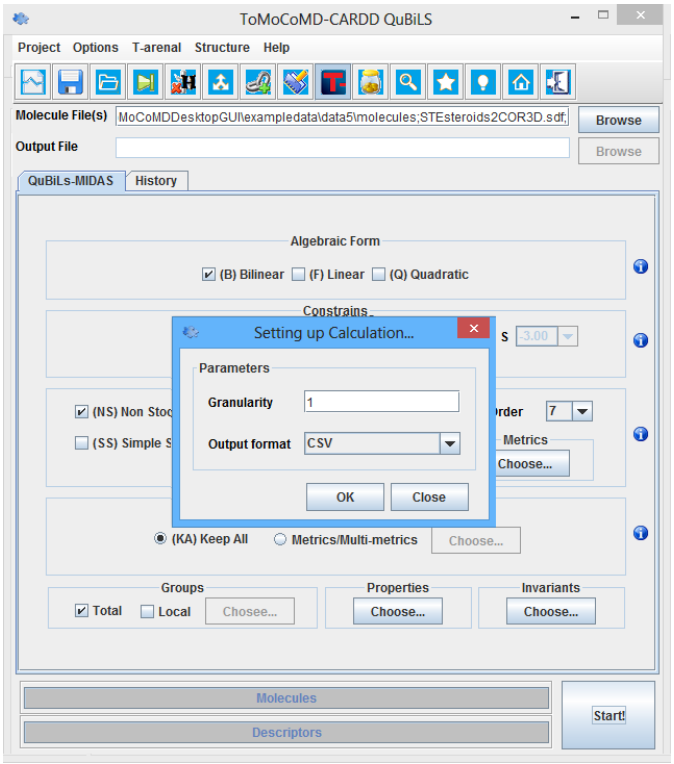

Figure 36. Dialog window to configure the input parameters of the distributed calculation.

Monitor progress of the distributed calculations on T-arenal system

When a calculation of molecular descriptors is sent to T-arenal, the end user can monitor its progress by using the menu item provided in the software. Firstly, the end user must show the tasks running into T-arenal (see Figure 37), choosing a task from the list shown and lastly to do click in the button Progress….

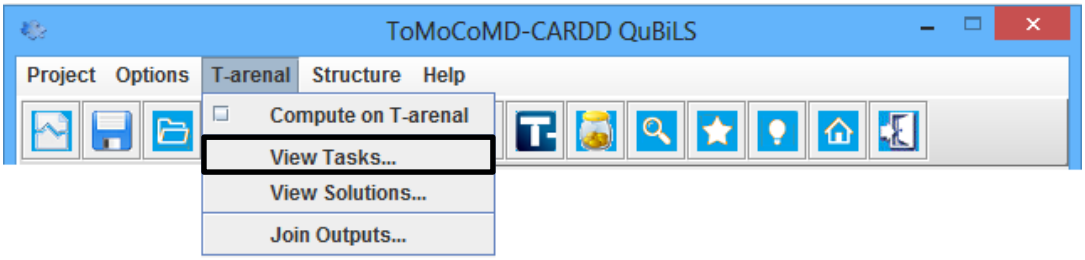

Figure 37. View Tasks menu item to show the distributed calculations into T-arenal.

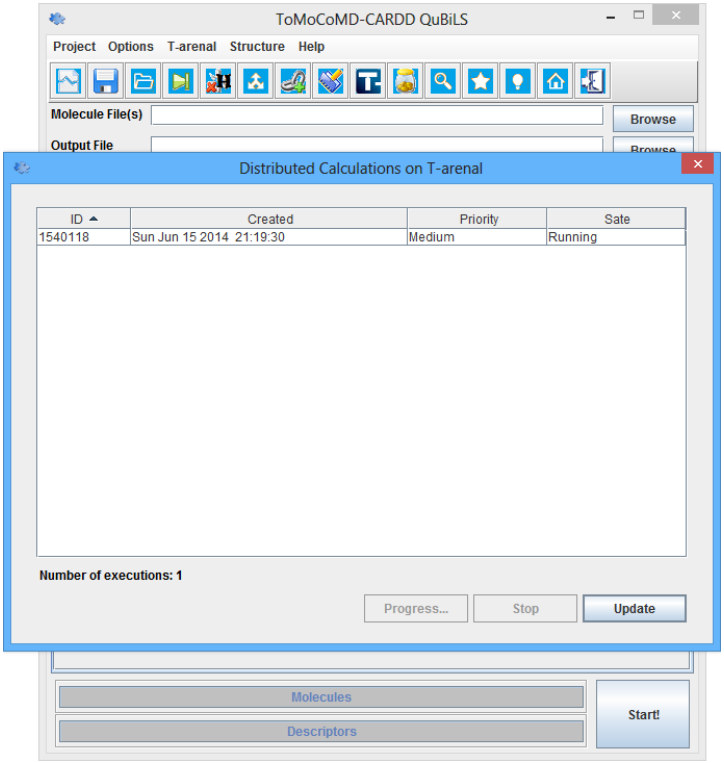

Figure 38. List of tasks running into T-arenal.

View solutions of the distributed calculations

Once the distributed calculations on the T-arenal system are performed, these are compressed into a .RAR file and put to arrangement of the end users. In this way, the results can be downloaded and subsequently analyzed.

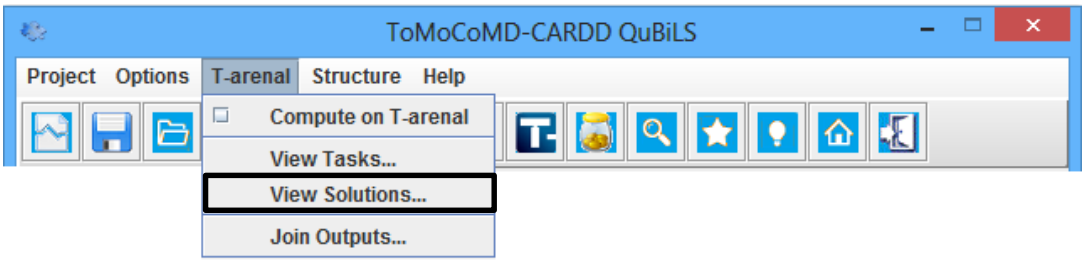

Figure 39. View Solutions menu item to show the solutions obtained.

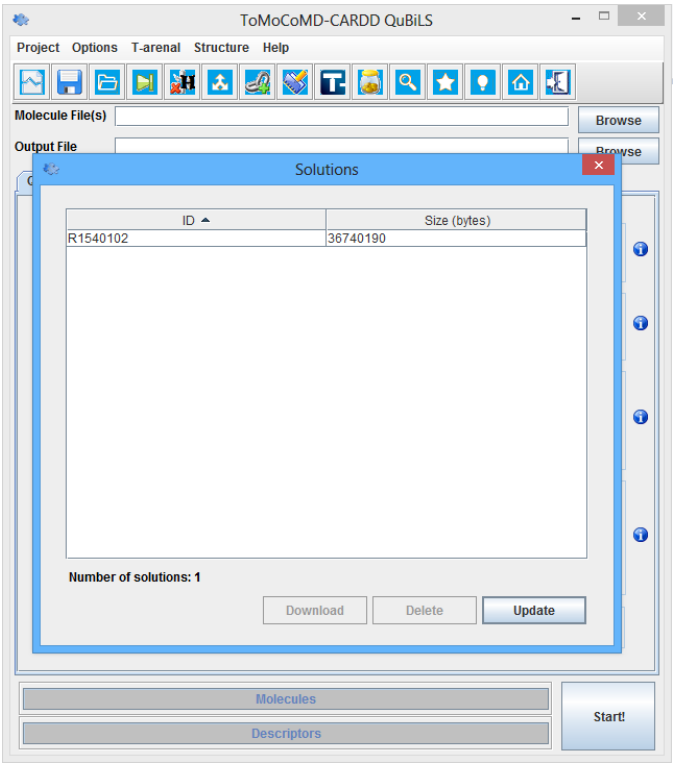

Figure 40. List of solutions obtained.

Configuring a project

The descriptors section *(Configuration Area)* in the middle part of the **QuBiLS-MIDAS** GUI consists of seven different sub-areas. Also, other three parameters can be configured by the user out of this area, these are:

- H-Atoms
- Distance to Molecule Center
- Lone-Pair Electrons

The previous options can be saved in the Project file and are likewise available in the Options Menu and Tool Bars. In each sub-area the user can select the possible parameters. Only in 4 sub-area is necessary to open a window to select the possible options, namely (dis-)similarity metrics (Figure 41), local indices (Figure 42), properties (Figure 43) and Invariants (Figure 44).

Figure 41. Dialog windows to set the (dis-)similarity metrics to compute the relations among "n" atoms. A) and B) (dis-)similarity metrics to compute the distance among two atoms. C) (dis-)similarity metrics to compute the relations among three atoms. D) (dis-)similarity metrics to compute the relations among four atoms.

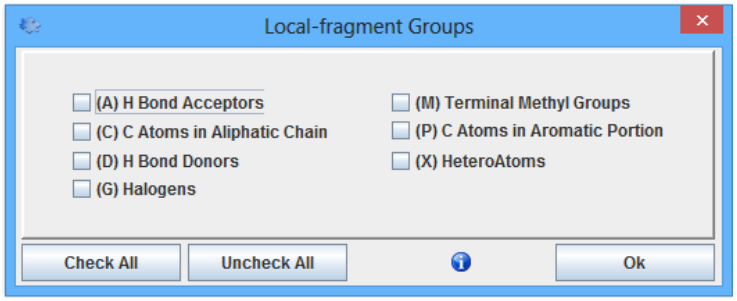

Figure 42. Locals (group-type) indices.

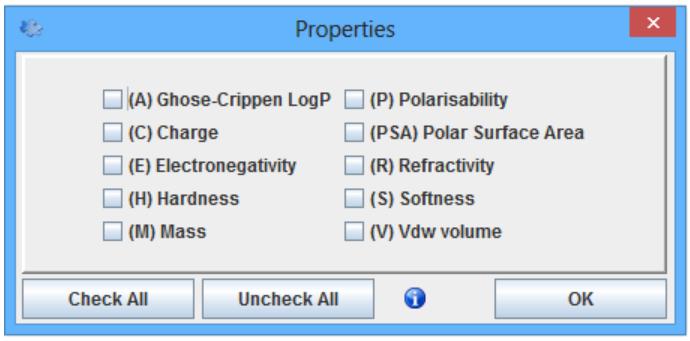

Figure 43. Atomic-properties (labels).

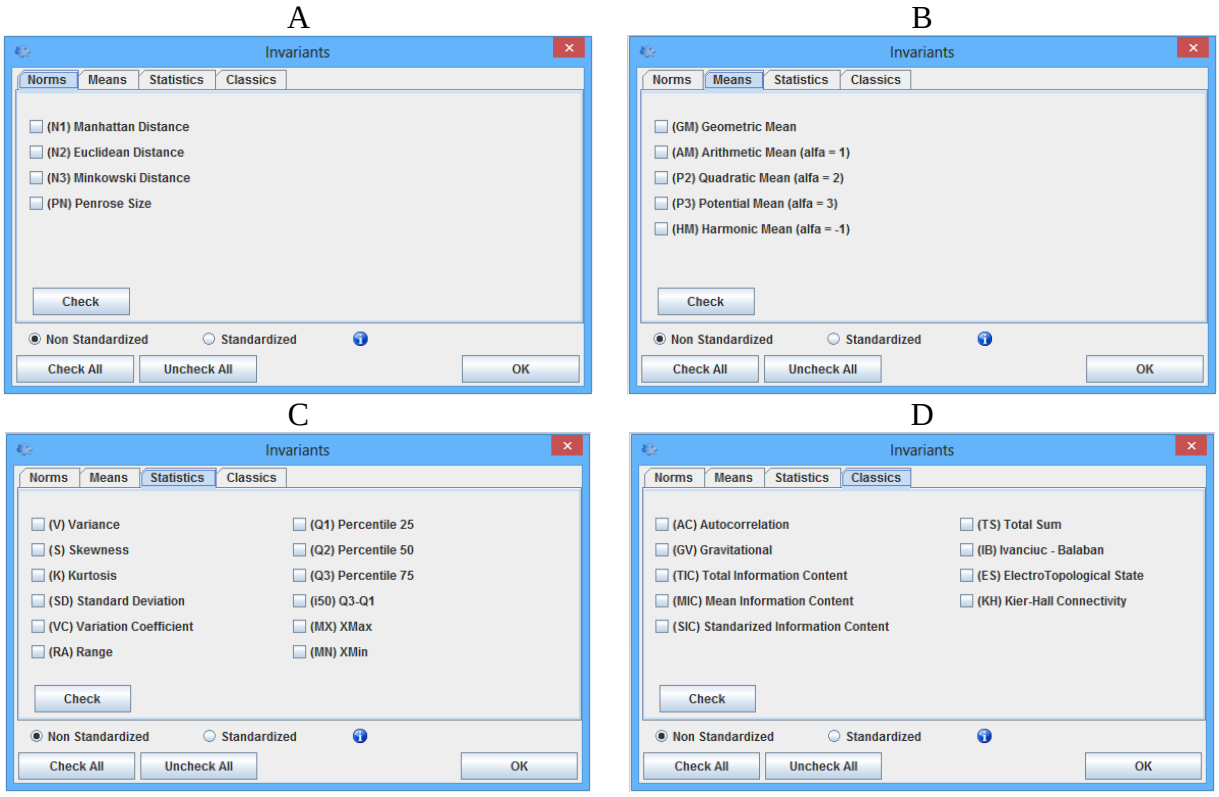

Figure 44. Dialog windows to set the invariant(s) to use to compute the molecular descriptors. A) Norm invariants. B) Mean invariants. C) Statistic invariants. D) Classic invariants.

Algebraic Forms

Three different algebraic forms can be calculated from input structures. Shortly, if a molecule consists of *n* atoms (*vector of* $\Re^n$), then the k^{th} *total (whole) two-linear*, *three-linear* and *four-linear* indices are calculated as *N-linear algebraic maps* (forms) in $\mathbb{R}^n$, in a canonical basis set. Specifically, the k^{th} *non-stochastic atom-based two-linear*, *three-linear* and *four-linear* indices for a molecule, $m^k(\bar{x},\bar{y})$, $tr^k(\bar{x},\bar{y},\bar{z})$ and $qu^k(\bar{x},\bar{y},\bar{z},\bar{w})$, respectively, are computed from the k^{th} *two-tuples*, *three-tuples* and *four-tuples spatial (dis-)similarity matrices*, [$\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$], as shown in Eqs. **1-3**, correspondingly:

$$m^k(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^k x^i y^j = [X]^T\,\mathbb{G}^k[Y] \quad (\mathbf{1})$$

$$tr^k(\bar{x},\bar{y},\bar{z}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} gt_{ijl}^k x^i y^j z^l = \mathbb{GT}^k \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \quad (\mathbf{2})$$

$$qu^k(\bar{x},\bar{y},\bar{z},\bar{w}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} gq_{ijl}^k x^i y^j z^l w^h = \mathbb{GQ}^k \cdot \bar{x} \cdot \bar{y} \cdot \bar{z} \cdot \bar{w} \quad (\mathbf{3})$$

where, *n* is the number of atoms in a molecule, g_{ij}^k, gt_{ijl}^k and gq_{ijl}^k are the coefficients of the matrices $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$ respectively, and and $x^1,\ldots,x^n$, $y^1,\ldots,y^n$, $z^1,\ldots,z^n$ and $w^1,\ldots,w^n$ are the coordinates or components of the molecular vectors $\bar{x}$, $\bar{y}$, $\bar{z}$ and $\bar{w}$ in a system of canonical ('natural') basis vectors of $\mathbb{R}^n$. As can be noticed, these molecular vectors are weighted with different atomic properties and thus several combinations are obtained. In this way, in the QuBiLS software are employed the algebraic forms shown in the Table 1.

Table 1. N-linear algebraic forms implemented in the QuBiLS program.

1. Two-linear [$m^k(\bar{x},\bar{y})$]
- Linear (X, Y = 1)
- Bilinear (X <> Y)
- Quadratic (X = Y)

Used symbols
1: Using the unitary vector
<>: Using different properties
=: Using equal properties

2. Three-linear [$tr^k(\bar{x},\bar{y},\bar{z})$]
- Threelinear (X <> Y <> Z)
- Threelinear-Quadratic-Bilinear ((X = Y) <> Z)
- Threelinear-Bilinear (X <> Y, Z = 1)
- Threelinear-Linear (X, Y = 1, Z = 1)
- Threelinear-Cubic (X = Y = Z)

3. Four-linear [$qu^k(\bar{x},\bar{y},\bar{z},\bar{w})$]
- Fourlinear (X <> Y <> Z <> W)
- Fourlinear-Quadratic-Threelinear ((X = Y) <> Z <> W)
- Fourlinear-Threelinear (X = 1, Y <> Z <> W)
- Fourlinear-Cubic-Bilinear ((X = Y = Z) <> W)
- Fourlinear-Bilinear (X = Y = 1, Z <> W)
- Fourlinear-Linear (X = Y = Z = 1, W)
- Fourlinear-Quadruple (X = Y = Z = W)

Constraints

Chiral Indices

The total and local n-linear indices, as defined in their "standard" form in the **QuBiLS-MIDAS** module not codify information about the chiral properties of the molecular structure. Thus, a *correction factor* is introduced in the molecular vectors $\bar{x}$, $\bar{y}$, $\bar{z}$ and $\bar{w}$, respectively, in

order to take into account this criterion. In this way, chirality-based molecular vectors are computed ($\overline{{}^{*}x}$, $\overline{{}^{*}y}$, $\overline{{}^{*}z}$ and $\overline{{}^{*}w}$), where each coefficient is equal to the sum of the considered atomic property and the *correction factor*. This idea in several works reported in the literature has been used, but only employing opposed integer numbers as *correction factor, i.e.*: 1 when an atom is labeled as *rectus* (R), -1 when an atom is labeled as *sinister* (S) and 0 when an atom does not present a specific chiral environment according to the Cahn-Ingold-Prelog rules. However, inspired on the No Free Lunch Theorem that could be interpreted as no single *correction factor* (*i.e.* 1 and -1) to perform chirality-based studies yields superior results than others when its result is averaged over all possible chemical datasets, then it may be stated that the use of several integer and rational numbers as *correction factor* could contribute to obtain better results in cheminformatics applications (*e.g.* QSAR studies).

Therefore, values in the range $[-3,3]$ with step equal to 0.25 are accounted for. This range of values was chosen with the purpose of considering a suitable spectra of numbers as *correction factor*. It is important to highlight that those atoms labeled as R or S can take any value, which is rather than to the reported where 1 and -1 constitute the only values for atoms classified as R and S, respectively. In this way, the *correction factor* for atoms labeled as R or S can have assigned negative and positive values (*e.g.* -3 for R atom and 1 for S atom, or vice versa), opposed values (e.g. -3 for R atom and 3 for S atom, or vice versa), positive values (*e.g.* 3 for R atom and 1 for S atom) or negative values (*e.g.* -3 for R atom and -1 for S atom). Lastly, it is valid to clarify that this *correction factor* has essentially a mathematical means and must not be cause of any misinterpretation.

N-tuples

For **Duplex option (2)**, the index calculations are performed over vertex pairs *i* and *j*. Here, the k^{th} *spatial (dis-)similarity matrix* of the geometric molecular structure, $\mathbb{G}^{k}$, is used like matrix forms. For duplex relation, *bilinear, linear* and *quadratic indices* can be obtained. However, a generalization of these indices can be obtained using *n-linear maps*. That is to say, for **3-uples option,** the index calculations are performed over vertices *i, j* and *k* using the k^{th} *three-tuples spatial (dis-)similarity matrix*, $\mathbb{GT}^{k}$; while for **4-uples option** the index calculations are performed over vertices *i, j, k* and *l* employing the k^{th} *four-tuples spatial (dis-)similarity matrix*, $\mathbb{GQ}^{k}$. For ternary and quaternary matrices, previously mentioned, several matrices types are used, such as:

- **3C**: All the elements of the 3-tuples matrix, with coordinates (*i, j, k*), are considered.
- **3nC**: Only elements of the triple matrix that satisfy the condition that all the 3 coordinates are different are selected.
- **4C**: All the elements of the quadruple matrix, with coordinates (*i, j, k, z*), are considered.
- **4nC**: Only elements of the quadruple matrix that satisfy the condition that all the 4 coordinates are different are considered

Matrix Forms

Four kinds of matrices (order k, *k* = 0-12) can be used in **QuBiLS-MIDAS** software, *namely* Non-stochastic (NS), Simple stochastic (SS), Double stochastic (DS) and Mutual Probabilistic (MP) matrices.

Non-stochastic (NS) matrix

The codification of 3D information of the non-covalent interactions of the molecular structure is fulfilled through rules among two, three and four atoms, and the values of these rules are represented in the *two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrix*, $\mathbb{G}$, $\mathbb{GT}$ and $\mathbb{GQ}$, respectively. The generalized expressions of these matrices are the *k^{th} two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, denoted by $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$, where superscript *k* indicates the power to which $\mathbb{G}$, $\mathbb{GT}$ and $\mathbb{GQ}$ are raised. Thus, for $k = 0$, all coefficients g_{ij}^0, gt_{ijl}^0 and gq_{ijlh}^0 corresponding to the matrices $\mathbb{G}^0$, $\mathbb{GT}^0$ and $\mathbb{GQ}^0$ have value 1; and for $k = 1$, the coefficients g_{ij}^1 of the matrix $\mathbb{G}^1$, gt_{ijl}^1 of the matrix $\mathbb{GT}^1$ and gq_{ijlh}^1 of the matrix $\mathbb{GQ}^1$, represent the information of the interactions among two, three and four atoms respectively. The formal definitions of these elements is shown below:

$$\begin{aligned} g_{ij}^1 &= D_{ij} \quad \textit{if atoms } i \textit{ and } j \textit{ are not equal} \\ &= L_{ij} \quad i = j \wedge \textit{lone-pairs are considered } (\textit{or } D_{io}) \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{1}$$

$$\begin{aligned} gt_{ijl}^1 &= T_{ijl} \quad \textit{if atoms } i, j \textit{ and } l \textit{ are not equal} \\ &= L_{ijl} \quad i = j = l \wedge \textit{lone-pairs are considered } (\textit{or } D_{io}) \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{2}$$

$$\begin{aligned} gq_{ijlh}^1 &= Q_{ijlh} \textit{ if atoms } i, j, l \textit{ and } h \textit{ are not equal} \\ &= L_{ijlh} \quad i = j = l = h \wedge \textit{lone-pairs are considered (or } D_{io}) \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{3}$$

where, D_{ij} is a (dis-)similarity metric between two atoms, T_{ijl} is an measure for ternary relations of atoms, Q_{ijlh} is an measure for quaternary relations of atoms, while L_{ij}, L_{ijl} and L_{ijlh} are the diagonal entries, which could have assigned two different values to achieve greater discrimination of the molecular structures: 1) representing the number of lone-pairs electrons for atoms, or 2) the spatial distance, D_{io} for each atom *i* and center of the molecule, *o*.

Simple-stochastic (SS) matrix

The *k^{th} simple-stochastic two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, ${}_{ss}\mathbb{G}^k$, ${}_{ss}\mathbb{GT}^k$ and ${}_{ss}\mathbb{GQ}^k$, can be directly obtained from $\mathbb{G}^k$, $\mathbb{GT}^k$ and $\mathbb{GQ}^k$, respectively. The coefficients of these matrices are computed as follows:

$${}_{ss}g_{ij}^k = \frac{{}_{ns}g_{ij}^k}{S_j} = \frac{{}_{ns}g_{ij}^k}{\sum_{j=1}^{n} {}_{ns}g_{ij}^k} \tag{4}$$

$${}_{ss}gt_{ijl}^k = \frac{{}_{ns}gt_{ijl}^k}{S_{jl}} = \frac{{}_{ns}gt_{ijl}^k}{\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt_{ijl}^k} \tag{5}$$

$$ {}_{ss}gq^{k}_{ijlh} = \frac{{}_{ns}gq^{k}_{ijlh}}{S_{jlh}} = \frac{{}_{ns}gq^{k}_{ijlh}}{\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq^{k}_{ijlh}} \quad (\mathbf{6}) $$

where, ${}_{ns}g^{k}_{ij}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gt^{k}_{ijl}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gq^{k}_{ijlh}$ are the elements of the k^{th} power of the matrix $\mathbb{GQ}$.

Double-stochastic (DS) matrix

The *k^{th} double-stochastic two-tuples spatial-(dis)similarity matrices*, ${}_{ds}\mathbb{G}^{k}$, can also be directly obtained from $\mathbb{G}^{k}$. Here, ${}_{ds}\mathbb{G}^{k} = [\,{}_{ds}g^{k}_{ij}\,]$, is a square matrix of order n (n = number of atomic nuclei). It should be remarked that the matrix ${}_{ds}\mathbb{G}^{k}$ has the property that *the sum of the elements in each row or in each column* is 1. Notice that ${}_{ss}\mathbb{G}^{k}$ matrix (simple stochastic) is not symmetric, therefore, with the aim of equalize the probabilities in both senses is employed a *double-stochastic scaling*. This scaling is performed by using Sinkhorn-Knopp algorithm.

Mutual probability (MP) matrix

The *k^{th} mutual probability two-tuples, three-tuples and four-tuples spatial-(dis)similarity matrices*, ${}_{mp}\mathbb{G}^{k}$, ${}_{mp}\mathbb{GT}^{k}$ and ${}_{mp}\mathbb{GQ}^{k}$, can be directly obtained from $\mathbb{G}^{k}$, $\mathbb{GT}^{k}$ and $\mathbb{GQ}^{k}$, respectively. The coefficients of these matrices are computed as follows:

$$ {}_{mp}g^{k}_{ij} = \frac{{}_{ns}g^{k}_{ij}}{S_{ij}} = \frac{{}_{ns}g^{k}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} {}_{ns}g^{k}_{ij}} \quad (\mathbf{7}) $$

$$ {}_{mp}gt^{k}_{ijl} = \frac{{}_{ns}gt^{k}_{ijl}}{S_{ijl}} = \frac{{}_{ns}gt^{k}_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}_{ns}gt^{k}_{ijl}} \quad (\mathbf{8}) $$

$$ {}_{mp}gq^{k}_{ijlh} = \frac{{}_{ns}gq^{k}_{ijlh}}{S_{ijlh}} = \frac{{}_{ns}gq^{k}_{ijlh}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{h=1}^{n} {}_{ns}gq^{k}_{ijlh}} \quad (\mathbf{9}) $$

where, ${}_{ns}g^{k}_{ij}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gt^{k}_{ijl}$ are the elements of the k^{th} power of the matrix $\mathbb{GT}$, ${}_{ns}gq^{k}_{ijlh}$ are the elements of the k^{th} power of the matrix $\mathbb{GQ}$.

Cut-Off Setting

The QuBiLS-MIDAS 3D-MDs codify information with respect to the relations between all atoms of a molecule or according to chemical fragments (*F*) of interest, *e.g.* carbon atoms in aliphatic chains. Therefore, with the purpose of establishing a relation between the topological and geometrical aspects for each group of *"N"* atoms considered and in this way take into account some short-, middle- and large-relations, two procedures are defined:

- *N-tuple Graph-theoretical cutoff* (*P*) known as "*path cutoff"*, based on the topological distance. These cutoffs are denoted as: ***lag P*** for $N=2$, ***lag 3P*** for $N=3$ and ***lag 4P*** for $N=4$.
- *N-tuple Euclidean-geometric cutoff* (*L*) known as "*length cutoff"*, based on the Euclidean distance. These cutoffs are denoted as: ***lag L*** for $N=2$, ***lag 3L*** for $N=3$ and ***lag 4L*** for $N=4$.

The application of one or both molecular cutoffs on the matrices $\mathbb{G}^1_{(F)}$, $\mathbb{GT}^1_{(F)}$ and $\mathbb{GQ}^1_{(F)}$ permits to compute the *two-, three- and four-tuple topological and geometric neighborhood quotient total (or local-fragment) spatial-(dis)similarity matrices*, $\mathbb{NQG}^1_{(F)}$, $\mathbb{NQGT}^1_{(F)}$ and $\mathbb{NQGQ}^1_{(F)}$, respectively. The coefficients of these novel matrix approaches are computed by multiplying the elements of the matrices $\mathbb{G}^1_{(F)}$, $\mathbb{GT}^1_{(F)}$ and $\mathbb{GQ}^1_{(F)}$ by a ratio obtained as the number of relations between the *N* considered atoms that present a topological and/or Euclidean-geometric distance smaller or equal to a predefined *p* and/or *l* thresholds. Then, the entries ${}^{NQ}g^1_{ij(F)}$, ${}^{NQ}gt^1_{ijl(F)}$ and ${}^{NQ}gq^1_{ijlh(F)}$ of the matrices $\mathbb{NQG}^1_{(F)}$, $\mathbb{NQGT}^1_{(F)}$ and $\mathbb{NQGQ}^1_{(F)}$ are mathematically defined as follows:

$$\begin{aligned} {}^{NQ}g^1_{ij(F)} &= g^1_{ij(F)} \text{ if } p_{min} \le p_{ij} \le p_{max} \quad \textit{and / or} \quad l_{min} \le l_{ij} \le l_{max} \\ &= 0 \text{ otherwise} \end{aligned} \tag{10}$$

$$\begin{aligned} {}^{NQ}gt^1_{ijl(F)} &= gt^1_{ijl(F)} \quad \textit{if } p_{min} \le p_{ij},p_{jl},p_{li} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij},l_{jl},l_{li} \le l_{max} \\ &= \frac{2}{3} gt^1_{ijl(F)} \quad \begin{cases} \textit{if } p_{min} \le p_{ij},p_{jl\,(li)} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij},l_{jl\,(li)} \le l_{max} \\ \textit{if } p_{min} \le p_{jl},p_{li} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{jl},l_{li} \le l_{max} \end{cases} \\ &= \frac{1}{3} gt^1_{ijl(F)} \quad \textit{if } p_{min} \le p_{ij\,(jl,li)} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij\,(jl,\,li)} \le l_{max} \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{11}$$

$$\begin{aligned} {}^{NQ}gq^1_{ijlh(F)} &= gq^1_{ijlh(F)} \quad \textit{if } p_{min} \le p_{ij},p_{jl},p_{lh},p_{hi} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij},l_{jl},l_{lh},l_{hi} \le l_{max} \\ &= \frac{3}{4} gq^1_{ijlh(F)} \quad \begin{cases} \textit{if } p_{min} \le p_{ij},p_{jl\,(lh)},p_{lh\,(hi)} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij},l_{jl\,(lh)},l_{lh\,(hi)} \le l_{max} \\ \textit{if } p_{min} \le p_{jl},p_{lh},p_{hi} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{jl},l_{lh},l_{hi} \le l_{max} \end{cases} \\ &= \frac{2}{4} gq^1_{ijlh(F)} \quad \begin{cases} \textit{if } p_{min} \le p_{ij},p_{jl\,(lh,\,hi)} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij},l_{jl\,(lh,\,hi)} \le l_{max} \\ \textit{if } p_{min} \le p_{jl},p_{lh\,(hi)} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{jl},l_{lh\,(hi)} \le l_{max} \\ \textit{if } p_{min} \le p_{lh},p_{hi} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{lh},l_{hi} \le l_{max} \end{cases} \\ &= \frac{1}{4} gq^1_{ijlh(F)} \quad \textit{if } p_{min} \le p_{ij\,(jl,\,li,\,hi)} \le p_{max} \quad \textit{and/or} \quad l_{min} \le l_{ij\,(jl,\,lh,\,hi)} \le l_{max} \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{12}$$

where, the coefficients g_{ij}^{1}, gt_{ijl}^{1}, gq_{ijlh}^{1} represents the relations between two, three and four atoms of a molecule and correspond to the total (or local-fragment) matrices $\mathbb{G}_{(F)}^{1}$, $\mathbb{GT}_{(F)}^{1}$ and $\mathbb{GQ}_{(F)}^{1}$, respectively. In addition, p_{xy} and l_{xy} represent the topological and Euclidean-geometric distance between two atoms of a molecule, while $[p_{\min}, p_{\max}]$ and $[l_{\min}, l_{\max}]$ constitute the used-defined topological and Euclidean-geometric intervals, respectively.

Also, other molecular cutoff procedures are proposed in order to only consider the ternary $(N=3)$ and quaternary $(N=4)$ relations between atoms of a molecule whose values are consistent with a specific multi-metric. These procedures are denominated as *N-tuple Geometric cutoff based on Multi-metrics* and its mathematical definition is as follows:

$$\begin{aligned} {}^{NQ}gt_{ijl(F)}^{1} &= gt_{ijl(F)}^{1} \text{ if } \mathrm{tv}_{\min} \le \mathrm{tv}_{\mathrm{ijl}} \le \mathrm{tv}_{\max} \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{13}$$

$$\begin{aligned} {}^{NQ}gq_{ijlh(F)}^{1} &= gq_{ijlh(F)}^{1} \text{ if } \mathrm{qv}_{\min} \le \mathrm{qv}_{\mathrm{ijlh}} \le \mathrm{qv}_{\max} \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{14}$$

where, tv_{ijl} and qv_{ijlh} are the values corresponding to the calculation of a ternary and quaternary multi-metric, respectively. In addition, $[tv_{\min}, tv_{\max}]$ and $[qv_{\min}, qv_{\max}]$ are the predefined intervals when cutoffs based on relations between three and four atoms are applied, respectively. Specifically, the ternary multi-metrics that may be used include the Triangle Area (***lag A***), Bond Angle (***lag BA***) and Ternary (or Triangle) Perimeter (***lag TP***); while the quaternary multi-metrics that may be used include Volume (***lag V***), Dihedral Angle (***lag DA***) and Quaternary (or Quadrilateral) Perimeter (***lag QP***).

It is important to highlight that the molecular cutoffs defined for a same number of atoms could be simultaneously applied, *e.g.*: in a relation between three distinct atoms $(i \neq j \neq l)$ if any permutation of *three-tuple cutoffs* (***lag 3P***, ***lag 3L***, ***lag A***, ***lag BA*** and ***lag TP***) is used, then all the considered criteria must be fulfilled. On the other hand, also the molecular cutoffs for relations between two, three and four atoms can be concurrently applied on the same matrix representation. Therefore, on *four-tuple matrix approaches* when four distinct atoms are analyzed $(i \neq j \neq l \neq h)$ then *four-tuple cutoffs* can be applied, if three distinct atoms are analyzed $[(i=j) \neq l \neq h]$ then *three-tuple cutoffs* can be applied, and if two distinct atoms are analyzed $[(i=j=l) \neq h]$ then *two-tuple cutoffs* can be applied. Likewise, this previous strategy is employed on *three-tuple matrix approaches* when *three-tuple cutoffs* and *two-tuple cutoffs* are computed for relations between three $(i \neq j \neq l)$ and two $[(i=j) \neq l]$ distinct atoms, respectively.

So far, only *topological and geometric neighborhood quotient total (or local-fragment) spatial-(dis)similarity matrices* for order 1 $(k=1)$ have been obtained ($\mathbb{NQQG}_{(F)}^{1}$, $\mathbb{NQQGT}_{(F)}^{1}$ and $\mathbb{NQQGQ}_{(F)}^{1}$). However, these matrices constitute classes of ***generalized matrices*** as well, where the coefficients for representations of higher orders $(k \geq 2)$ are computed through the Hadamard product. Thus, these are the basis for calculating *topological and geometric neighborhood quotient descriptors* using the Eqs. **1-3** but employing the matrices $\mathbb{NQQG}_{(F)}^{k}$, $\mathbb{NQQGT}_{(F)}^{k}$ and $\mathbb{NQQGQ}_{(F)}^{k}$ in

place of $\mathbb{G}_{(F)}^{k}$, $\mathbb{GT}_{(F)}^{k}$ and $\mathbb{GQ}_{(F)}^{k}$, respectively. To conclude, it may be stated that with the incorporation of these cutoffs to the QuBiLS-MIDAS formalism, chemical information regarding particular topological and geometric aspects of the molecules is codified. This constitutes an important advance in the QuBiLS-MIDAS framework since it is known that some chemical and/biological properties are more dependent on atomic interactions at particular distances, and thus considering specific atomic separations/relations should enhance the modeling capacity of the QuBiLS-MIDAS 3D-MDs.

Group Sub-Area

In addition to ***total*** *algebraic indices* computed for the whole molecule, a **local-fragment** formalism can be developed. In this way, the geometric matrix representations of the molecular structures can be transformed to considerer information related with groups or atom-types belonging to a specific molecular fragment (*F*). So, these *local-fragment matrices* are used as matrix form of the algebraic maps to compute the *local-fragment indices*. The molecular fragments employed in this software are:

- Hydrogen bond acceptors (A)
- Carbon atoms in aliphatic chains (C)
- Hydrogen bond donors (D)
- Halogens (G)
- Terminal methyl groups (M)
- Carbon atoms in aromatic portion (P)
- Heteroatoms (O, N and S in all valence states, denoted as X)

Properties (labels)

Atom-properties are used in **QUBILs-MIDAS** as **Atom-Weights.** The first four atom properties are standard values for each element, *namely*:

1. Polarizability (p),
2. Atomic Mass (m) [carbon atom scaled values],
3. VdW Volumen (v),
4. Electronegativity (e).

However, other six interesting atom properties also included as Atom-weights are:

5. Charge (c),
6. TPSA (psa),
7. ALogP (a),
8. Refractivity (r),
9. Softness (s)
10. Hardness (h).

These atomic properties were taken from CDK (Chemical Development kit) implementation.

Invariants to LOVIs Vector

The invariants are numerical quantities derived from the molecular structure and used to characterize local properties of a molecule; these numbers are calculated in such a way as to be independent of any arbitrary atom/bond numbering. Local invariants can be distinguished into LOcal Vertex Invariants (LOVIs) and LOcal Edge Invariants (LOEIs), depending on whether they refer to atoms or bonds.

LOVIs of a molecule are usually collected into an *N*-dimensional vector (*N* = number of atoms). LOVIs and LOEIs (also known as LOBIs) are used to calculate several molecular indices by applying different aggregation operators. *L* is here adopted as the general symbol for local invariants.

Over the years, it has been generally accepted that the definition of global (or local) indices from LOVIs (L_i) implies the summation of the contributions of the elements that constitute a given molecular structure. However, summation (Minskowski's first norm (N1) in our specific case) is just one of the many invariants capable of globally characterizing given LOVIs.

In this program, are employed a series of *invariants* that generalize the traditional method of obtaining global (or local) invariants by summation of the LOVIs. These are classified in four major groups:

1) **Norms (or Metrics)**:
   - **a)** Minkowski's norms (N1, N2, N3), and
   - **b)** Penrose's size (PN);
2) **Mean Invariants (first statistical moment)**:
   - **a)** Geometric Mean (G),
   - **b)** Arithmetic Mean (M),
   - **c)** Quadratic Mean (P2),
   - **d)** Potential Mean (P3) and
   - **e)** Harmonic Mean (A);
3) **Statistical Invariants (highest statistical moments):**
   - **a)** Variance (V),
   - **b)** Skewness (S),
   - **c)** Kurtosis (K),
   - **d)** Standard Deviation (DE),
   - **e)** Variation Coefficient (CV),
   - **f)** Range (R),
   - **g)** Percentile 25 (Q1),
   - **h)** Percentile 50 (Q2),
   - **i)** Percentile 75 (Q3),
   - **j)** Inter-quartile Range (I50),
   - **k)** X max (MX) and
   - **l)** X min (MN),
4) **"Classical algorithms" Invariants:**
   - **a)** Autocorrelations AC(i),
   - **b)** Gravitational (GI(i)),
   - **c)** Total sum at *k* lags (TSk(i)),
   - **d)** Kier-hall connectivity (CN(i)),

e) Mean information content (MI(i)),
f) Total information content (TI),
g) Standardized information content (SI),
h) Ivanciuc-Balaban operator (IB),
i) Electrotopological state (ES(i)).

Standardized tab

In the standardization procedure, all values of *original* LOVIs are replaced by standardized LOVI values which are computed as follows: *Std. LOVIs* = (Original LOVI – mean of LOVIs)/Std. deviation of original LOVIs. With this re-scaling, each new LOVI has a mean of 0 and a standard deviation of 1.

Table 1. Norms (Metrics) Invariant.

Name	ID	Formula (Equation)
Minkowski's norms (p = 1) Manhattan norm	N1	$\left\|\bar{x}\right\|_1 = \sum_{i=1}^{n} \left\|L_i\right\|$
Minkowski's norms (p = 2) Euclidean norm	N2	$\left\|\bar{x}\right\|_2 = \sqrt{\sum_{i=1}^{n} \left\|L_i\right\|^2}$
Minkowski's norms (p = 3)	N3	$\left\|\bar{x}\right\|_3 = \sqrt[3]{\sum_{i=1}^{n} \left\|L_i\right\|^3}$
Penrose's size	PN	$d_i = \sqrt{\frac{1}{n^2}\left[\sum_{i=1} (L_i)\right]^2}$

Note 1. The general equation of Minkowski's norms is, $\left\|\bar{x}\right\|_p = \sqrt[p]{\sum_{i=1}^{n} \left|x_i\right|^p}$.

Note 2. The formulae used in these invariants, are simplified forms of general equations given that the vector $\bar{y}$ is constituted of the coordinates of the origin. For example, in the case of the Euclidean norm (N2), the general formula is: $\left\|\bar{x}\right\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2 + (x_j - y_j)^2 + (x_z - y_z)^2}$. But given that $\bar{y}$ = (0, 0, 0), this formula reduces to $\left\|\bar{x}\right\|_2 = \sqrt{\sum_{i=1}^{n} \left|x_i\right|^2}$.

Table 2. Mean (First Statistical Moment) Invariants.

Name	ID	Formula (Equation)
Geometric Mean	G	$\bar{\xi} = \sqrt[n]{\prod_{i=1}^{n} L_i}$
Arithmetic Mean (power mean with degree α = 1)	M	$m_\alpha = \left(\frac{L_1^\alpha + L_2^\alpha + ... + L_n^\alpha}{n}\right)^{\frac{1}{\alpha}}$
Quadratic Mean (power mean with degree α = 2)	P2	
Power Mean with degree α = 3	P3	
Harmonic Mean (power mean with degree α = -1)	A	

Table 3 Statistical (Highest Statistical Moments) Invariants

Name	ID	Formula (Equation)
Variance	V	$V = \frac{\sum_{i=1}^{n} \left(L_i - \bar{L}\right)^2}{n-1}$
Skewness	S	$S = n*M_3 /[(n-1)*(n-2)*s^3]$ $M_3 = \sum_{i=1}^{n} \left(L_i - \bar{L}\right)^3$ s^3 *is the standard deviation raised to the 3rd power* *n is the number of atoms.*
Kurtosis	K	$K = [n*(n+1)*M_4 - 3*M_2*M_2*(n-1)] / [(n-1)*(n-2)*(n-3)*s^4]$ $M_j = \sum_{i=1}^{n} \left(L_i - \bar{L}\right)^j$ *n is the number of atoms.* s^4 *is the standard deviation raised to the fourth power*
Standard Deviation	DE	$\sigma = \sqrt{\frac{\left(\sum L_i - \bar{L}\right)^2}{n-1}}$
Variation Coefficient	CV	$c_v = s / \bar{L}$
Range	R	$R = L_{\max} - L_{\min}$
Percentile 25	Q1	$P25 = \left[\frac{N}{4} + \frac{1}{2}\right]$ *N is the number of values*
Percentile 50	Q2	$P50 = \left[\frac{N}{2} + \frac{1}{2}\right]$ *N is the number of values*
Percentile 75	Q3	$P75 = \left[\frac{3N}{4} + \frac{1}{2}\right]$ *N is the number of values*
Inter-quartile Range	I50	$I50 = P75 - P25$
X max	MX	L_i maximum
X min	MN	L_j minimum

Table 4 "Classical" (classical functions to derive MDs from LOVIs) Invariants

Name	ID	Formula (Equation)
Autocorrelation	$AC^k(i)$	$AUT_k = \sum_{i=1}^{n}\sum_{j\geq 1}^{n} L_i \times L_j \bullet (\delta(d_{ij}, k)), k = 1,2,..7$
Gravitational	$GI^k(i)$	$G_k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{L_i L_j}{{}^k d_{ij}} \bullet \delta(d_{ij}, k)), k = 1,2,..7$
Total sum at lag k	$TS^k(i)$	$TS_k = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij} \bullet \delta(d_{ij}, k)) = 1,2,...7$
Kier-Hall connectivity	$CN^m(i)$	${}^m Chi_t = \sum_{i=1}^{K} (\prod_{i=1}^{n_k} L_i, w)_k^{\lambda}$ *where, K is the number of sub-graphs, n_k is the number of atoms in a fragment, λ is equal to ½, m and t are the sub-graph order and type, respectively*

Mean Information Content	MI(*i*)	$MI = -\sum_{i=1}^{A} \frac{N_g}{N_o} \cdot \log_2 \frac{N_g}{N_o}$ *where, Ng is the number of atoms with the same LOVI value. N_o is the number of atoms in a molecule*
Total Information Content	TI	$TI = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g$
Standardized Information Content	SI	$SI = \frac{IT}{N_0 \cdot \log_2 N_0}$
Electrotopological state (E-state index)	ES(*i*)	$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^{n} \frac{I_i - I_j}{(d_{ij}+1)^2}$ where, I_i is the intrinsic state of the i^{th} atom and ΔI_i is the field effect on the ith atom calculated as perturbation of the I_i of i^{th} atom by all other atoms in the molecule, d_{ij} is the topological distance between the i^{th} and the j^{th} atoms, and *n* is the number of atoms. The exponent k is 2.
Ivanciuc-Balaban Type-Indices	IB(*i*)	$J_k = \frac{n^2 \cdot B}{n + C + 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \left[L_i \times L_j \right]^{-\frac{1}{2}}$ where, the summation goes over all pairs of atoms but only pairs of adjacent atoms are accounted for by means of the elements a_{ij} of the adjacency matrix. The *n*, *B*, and *C* are the number of atoms, bonds, and rings (cyclomatic number), respectively.

Additional Configuration Options

Hydrogen's Atoms

The H-atoms can be used or not for make the calculations. That is to say, this option ON (consider) or OFF (not consider) the H-atoms in each molecule in input file. This option can be configured in option menu.

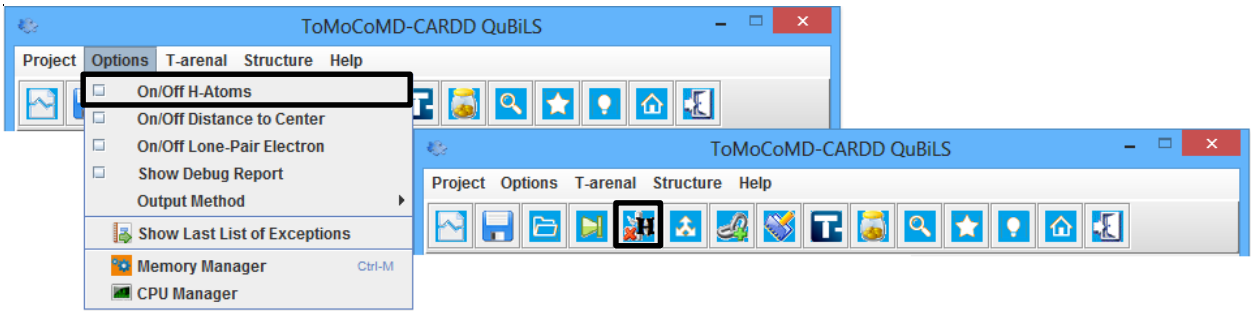

Figure 45. On/Off H-atoms button.

Distance to Molecule Center

The computations of the distance of each atom to molecule center can be used or not to perform the calculations. That is to say, ON (compute) while OFF (does not compute) the distance to molecule center. This option can be configured in option menu.

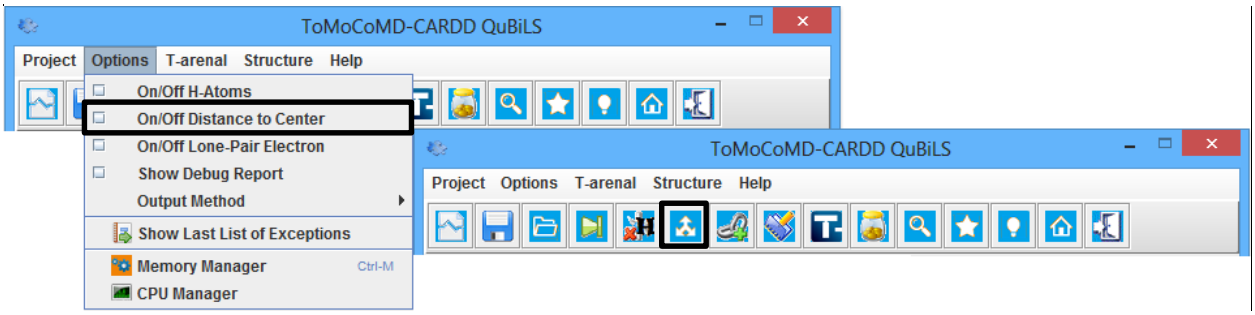

Figure 46. On/Off H-atoms button.

Lone Pairs Electrons

The lone-pair electrons on heteroatoms can be used or not to perform the calculations. That is to say, ON (considers) while OFF (does not consider) the so-called *n*-electron for heteroatoms in each molecule in input file. This option can be configured in option menu.

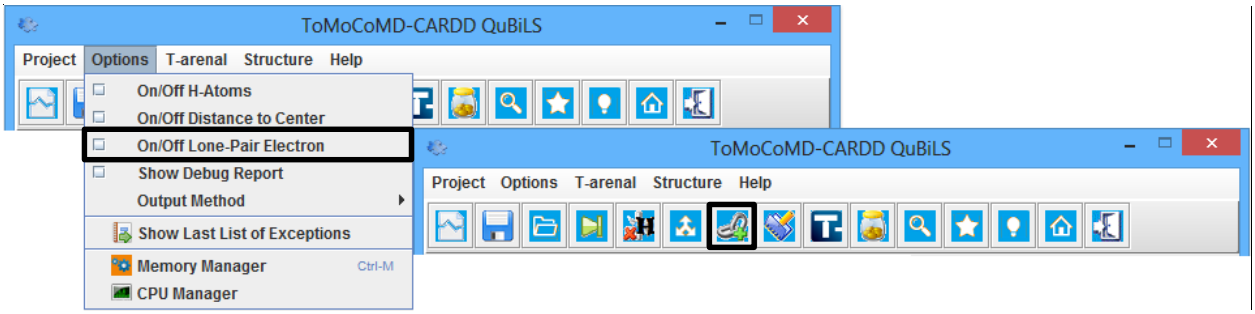

Figure 47. On/off lone-pair electrons button.

Input and Output Files

The following is a description of the **Input** and **Output** section of the **QuBiLs-MIDAS** GUI. In these sections, input structure files can be loaded and the output descriptor files can be chosen.

Structure input file are selected and loaded by clicking the *Browse* button in the **Input Molecules** section in the upper right part of the GUI. A dialog box appears displaying the directory that is specified in the input file. The last input folder path is remembered for QuBiLS-MIDAS, so you can easy locate your structures files. When the input file is selected and loaded the file format is recognized automatically. The current version of QuBiLS-MIDAS supports MDL MOL and SDF files.

The name and path of the descriptor output file is selected by clicking on the *Browse* button in the **Output** section in the upper right part of the GUI. QuBiLS-MIDAS supports, CSV format (comma separated value), TSV format (space-separated values file) and ARFF (Attribute-Relation File Format) Weka file.

Supported File Formats

MDL Molfile (MOL)

An MDL Molfile is a file format created by MDL, for holding information about the atoms, bonds, connectivity and coordinates of a molecule. The Molfile consists of some header

information, the Connection Table (CT) containing atom info, then bond connections and types, followed by sections for more complex information.

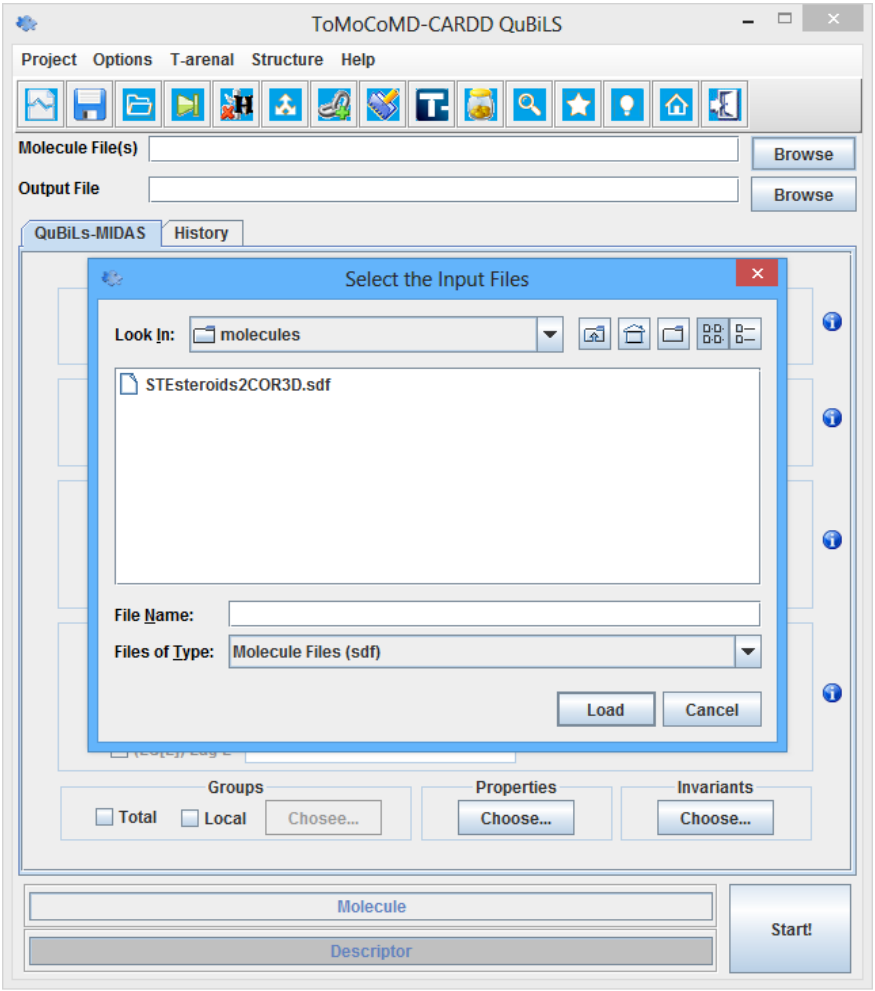

Figure 48. Browsing Input and Output files.

The Molfile is sufficiently common that most, if not all, chemo-informatics software systems/applications are able to read this format, though not always to the same degree. It is also supported by some computational software such as Mathematica (Wolfram-Research). **QuBiLS-MIDAS** reads V2000 and V3000 versions.

MDL Structure Data File (SDF)

SDF is one of a family of chemical-data file formats developed by MDL, it is intended especially for structural information. "SDF" stands for structure-data file, and SDF files actually wrap the Molfile (MDL Molfile) format. Multiple compounds are delimited by lines consisting of four dollar signs ($$$$). A feature of the SDF format is its ability to include associated data.

Space and Comma Separated Value Files (TXT, CSV)

A **space-separated values** file is a simple text format for a database table. Each record in the table is one line of the text file. Each field value of a record is separated from the next by a space (or blank) character, it is a form of the more general delimiter-separated values format.

As file extension for this output file we choose TXT, because it is a simple file format that is widely supported, so it is often used to move spaced data between different computer programs that support the format. For example, a space-separated file might be used to transfer information from a database program to a spreadsheet.

TXT is an alternative to the common comma-separated values (CSV) format, which often causes difficulties because of the need to escape commas. Literal commas are very common in text data.

A **comma-separated value (CSV)** file stores tabular data (numbers and text) in plain-text form. A plain text form means that the file is a sequence of characters, with no data that has to be interpreted instead, as binary numbers. A CSV file consists of any number of records, separated by line breaks of some kind; each record consists of fields, separated by some other character or string, most commonly a literal comma or tab. Usually, all records have an identical sequence of fields.

CSV is a common, relatively simple file format that is widely supported by consumer, business, and scientific applications. Among its most common uses is moving tabular data between programs that natively operate on incompatible (often proprietary and/or undocumented) formats. This works because so many programs support some variation of CSV at least as an alternative import/export format.

Weka Attribute-Relation File Format (ARFF)

An **ARFF** (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software (Waikato). A complete specification of ARFF files can be found at http://weka.wikispaces.com/ARFF.

Files Created for QuBiLS-MIDAS

QuBiLS-MIDAS produce an output file containing the values of the calculated and selected MDs, together with the additional information imported by the user. The Output File is selected by clicking on the *Browse* button in the *Output* section and QuBiLS-MIDAS supports, CSV format (comma separated value), TXT format (space-separated values file) and ARFF (Attribute-Relation File Format) Weka files.

The error and warning messages given below are printed to the Exception Windows (see Exception Windows section) and can be saved by clicking the *Save* button.

The "missing values" is represented by a constant sequence of characters: "*NaN*", stands for *Not a Number* value. For instance, three errors or exceptions types are possible (see Special Instructions and Exceptions section):

1. Errors in checking or cleaning the molecular structure.
2. Errors in calculating the algebraic form descriptors.
3. Unexpected Error in calculating a descriptor.

The **standard QuBiLS-MIDAS format** for the output file (.csv) is organized as follows (this format, namely, array of MDs blocks, cannot be changed by the user, see Table 6 for a simple example):

- The *first record* (column) contains the name of molecules, *that is*: the "name of each file" plus "_" plus the name of each molecule within the MOL (or SDF) MDL file. If inside of each .mol file, the names do not exist, then a **MolID** is generated, by using a **sequential number of molecules,** together with a **molecule identifier** (i.e. *mol#*).
- The following records (columns) contain the **variable labels** *(descriptor headers), i.e.* **N1_TrC_AB_nCi_3_M27_NS4_T_KA_c_MID** (see Descriptor Search Tool in order to automatic decodify each header).

Table 5. An example Output file.

molecules	N1_TrC_AB_nCi_3_M27_NS0_T_KA_c_MID	N1_TrC_AB_nCi_3_M27_NS1_T_KA_c_MID
STEsteroids2COR3D.sdf_aldosterone	-3.38E-18	0.529577
STEsteroids2COR3D.sdf_androstanediol	-3.32E-18	-0.07715
STEsteroids2COR3D.sdf_5-androstenediol	-5.82E-19	-0.10126
STEsteroids2COR3D.sdf_4-androstenedione	-1.49E-17	-0.14713
STEsteroids2COR3D.sdf_androsterone	-2.79E-18	-0.10328
STEsteroids2COR3D.sdf_corticosterone	-3.76E-19	0.585171
STEsteroids2COR3D.sdf_cortisol	9.03E-18	-0.88704

Example Data

Click the example data icon in the tool bar to access these molecular datasets. These datasets will permit simple test calculations to be made.

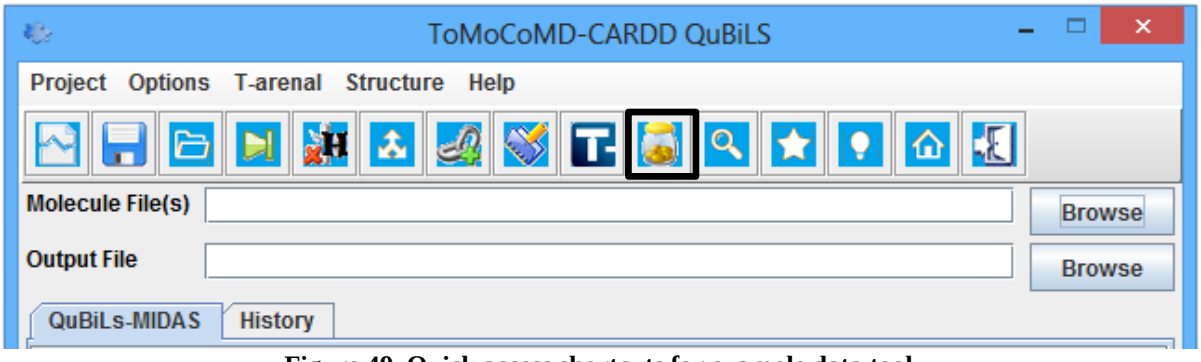

Figure 49. Quick access shortcuts for example data tool.

Searching for Descriptors Headers

By clicking the button 'Descriptor search' a window will appear where the user can enter the symbol (descriptor headers) of an unknown descriptor. A short definition of each part will be returned.

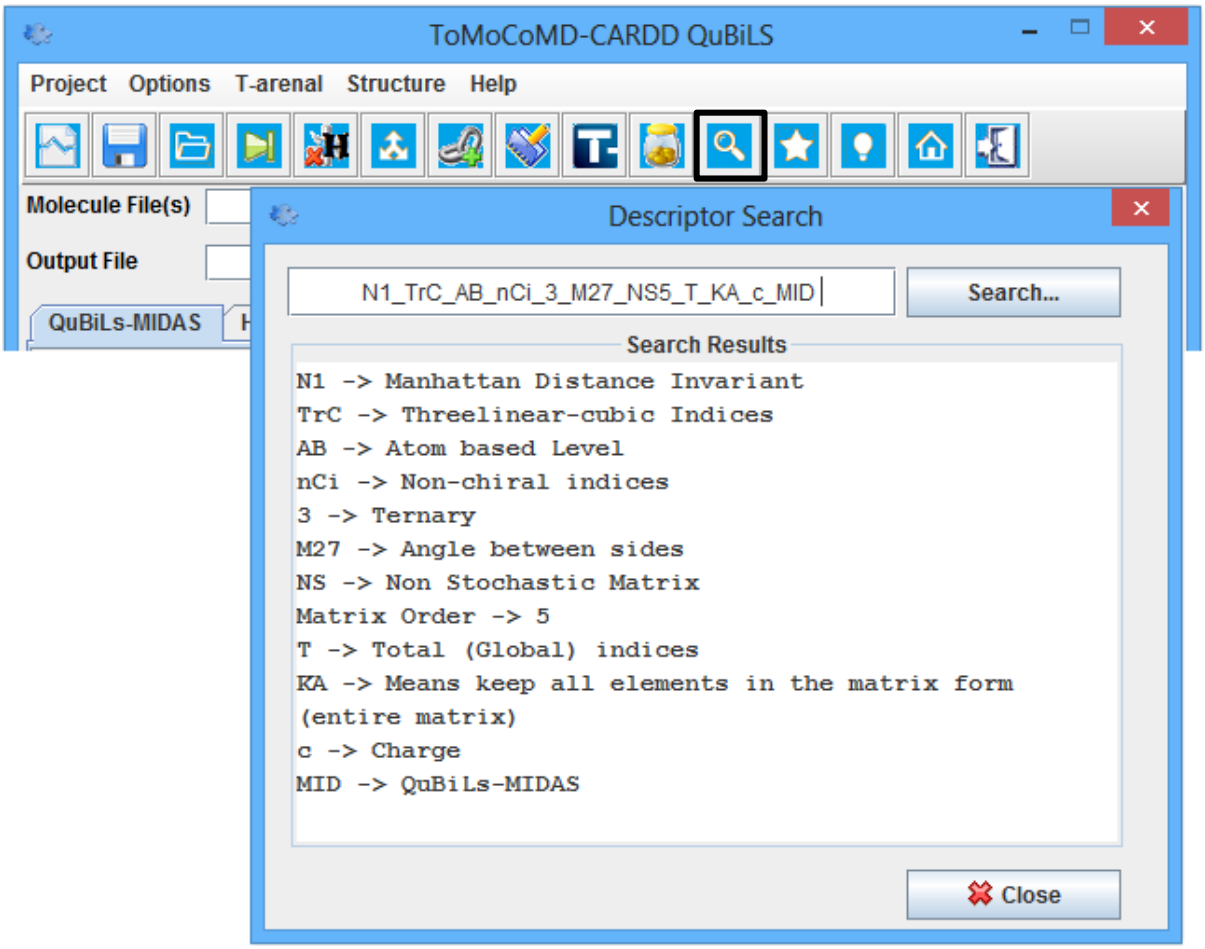

Figure 50. Descriptor Search Tool windows.

Debug Report Capability

This option permits, for each molecule in the dataset, the user to save a txt file with the atom-property (or two properties for bilinear forms) vector, the *k* order of molecular graph matrices (**NS**, **SS**, **DS** or **MP**), and LOVIs vectors (for each *k* order).

Property Vector X
Property Vector Y (if bilinear form is present)

Matrix order 0 (for a matrices forms, *namely* **NS**, **SS**, **DS** or **MP)**
LOVIS Vector order 0 (for a algebraic forms, *namely* **Linear**, **Bilinear**, or **Quadratic)**

Matrix order 1 (for a matrices forms, *namely* **NS**, **SS**, **DS** or **MP)**
LOVIS Vector order 1 (for a algebraic forms, *namely* **Linear**, **Bilinear**, or **Quadratic)**

Matrix order 2 (for a matrices forms, *namely* **NS**, **SS**, **DS** or **MP)**
LOVIS Vector order 2 (for a algebraic forms, *namely* **Linear**, **Bilinear**, or **Quadratic)**
…
Matrix order 12 (for a matrices forms, *namely* **NS**, **SS**, **DS** or **MP)**

LOVIS Vector order 12 (for a algebraic forms, *namely* **Linear**, **Bilinear**, or **Quadratic)**

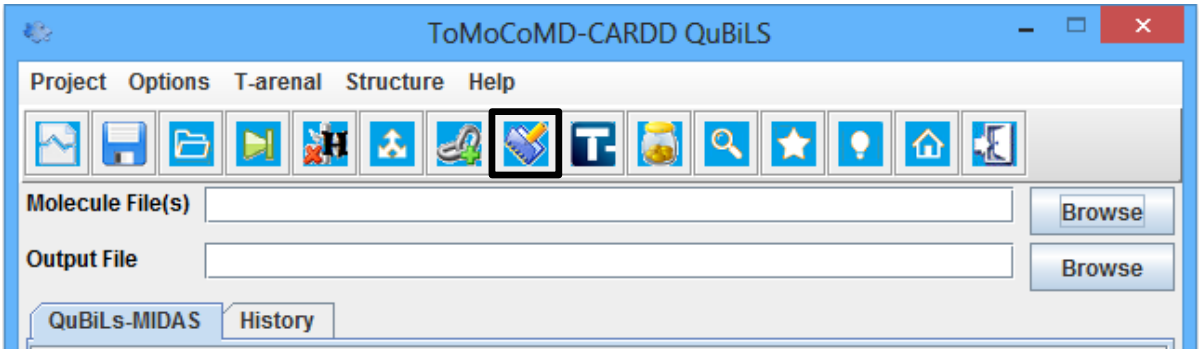

Figure 51. On/Off Generate Debug Report button.

Batch Mode Execution

QuBiLS-MIDAS can be executed in the interactive mode by using the GUI (as described in the previous sections) or in batch mode. For the execution in batch mode a project file (or several) that can be created by using the GUI is needed.

The executions in batch mode allow an easy integration for high-throughput and routinely carried out calculations. The batch version can be started at project menu. The simplest way to run **QuBiLS-MIDAS** in batch mode is to use one or several (eight as maximum) project files in which the all MDs and parameter are stored. Here, one or several datasets can be use (eight as maximum). It is important highlight that in the batch mode also can be sent the calculations of the molecular descriptors over large chemical datasets toward the T-arenal system.

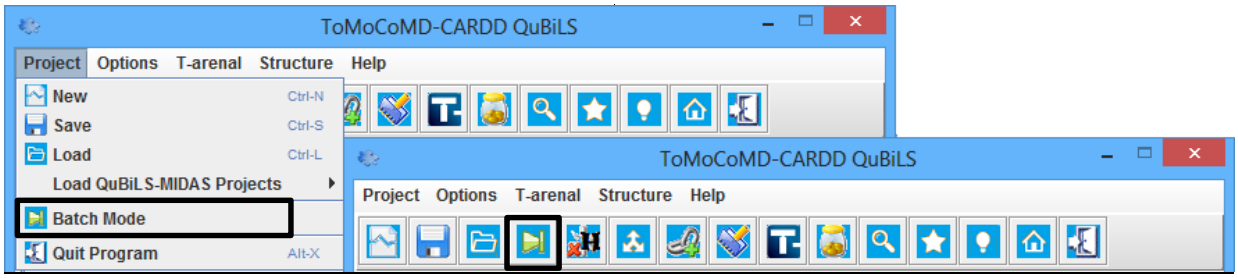

Figure 52. Batch Mode button.

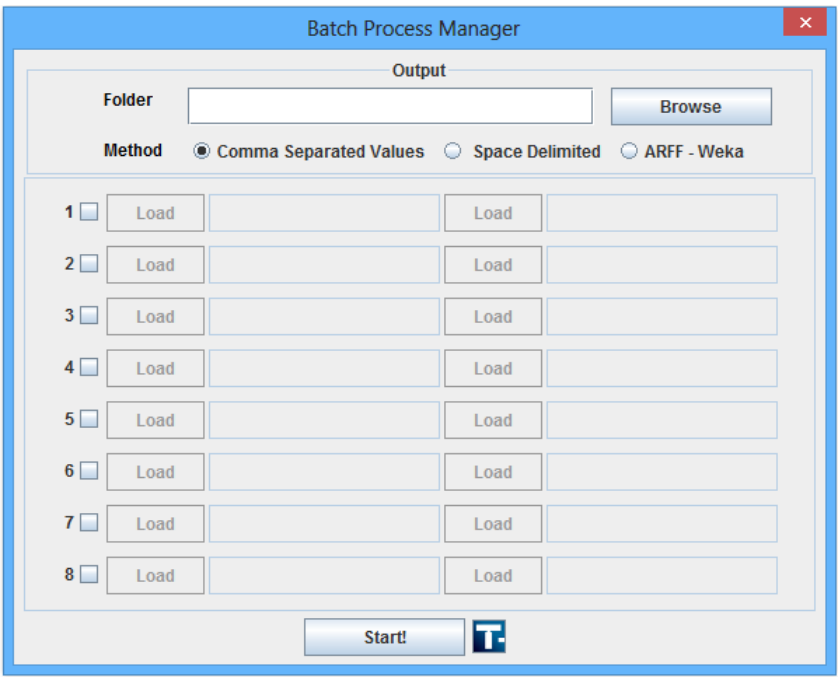

Figure 53. Batch Mode window.

Special Instructions and Exceptions

During the descriptor calculation **QuBiLS-MIDAS** writes out a log file that shows some statistics on the program run (History tab-window) and summarizes the errors and critical situations (warnings) encountered during the processing of the input structures (**Exception file**). That is, by checking the tab 'History' in the Configuration frame, a new window will appear during descriptor calculation where the main information concerning the batch in progress is progressively shown. If errors in the structural checking occur for a molecule, the molecule will be automatically *skipped* and the descriptors for this molecule are not calculated. In addition, if an error in descriptor calculation occurs for a molecule, then all its descriptor values will be missing in the final output file (*NaN*). Three errors or exceptions types are possible:

1. Errors in checking or cleaning the molecular structure. The molecule will be automatically skipped and is not considered for calculation results if some of the following take place:

    a. Connectivity Checker failure: if it has more than two parts, assumes it's a salt and just works with the larger part. Ideally a check should be made to ensure that the smaller part is a metal, or halogen etc.
    b. Markush structure found.
    c. Atom typing not recognized.
    d. Hückel Aromaticity Detection failure.
    e. Adding Implicit Hydrogen failure.
    f. Problem setting up 3D coordinates for implicit hydrogen(s) added.
    g. Atomic Covalent Radius not found.
    h. No 3D coordinates detected. Required for Hardness and Softness Properties

2. Errors calculating the algebraic form descriptors. The missing values label *NaN* is placed as descriptor value for each invalid entry. For example while using AlogP and Refractivity properties (see Properties Section) if the implicit hydrogen were added to the current molecule, and this molecule does not pass the Atom Type recognition process, the selected properties cannot be calculated.
3. Unexpected Error calculating descriptor. Any other error and exception will be notified through the Exception window, and the output file shows only the name of the molecule while the rows for the corresponding descriptor calculations are empty.

The list of molecules with problems for error in calculation is shown in the 'Exception file' window together with the error type. The error and warning messages given below are printed in this log file that can be saved by clicking the **save** button.

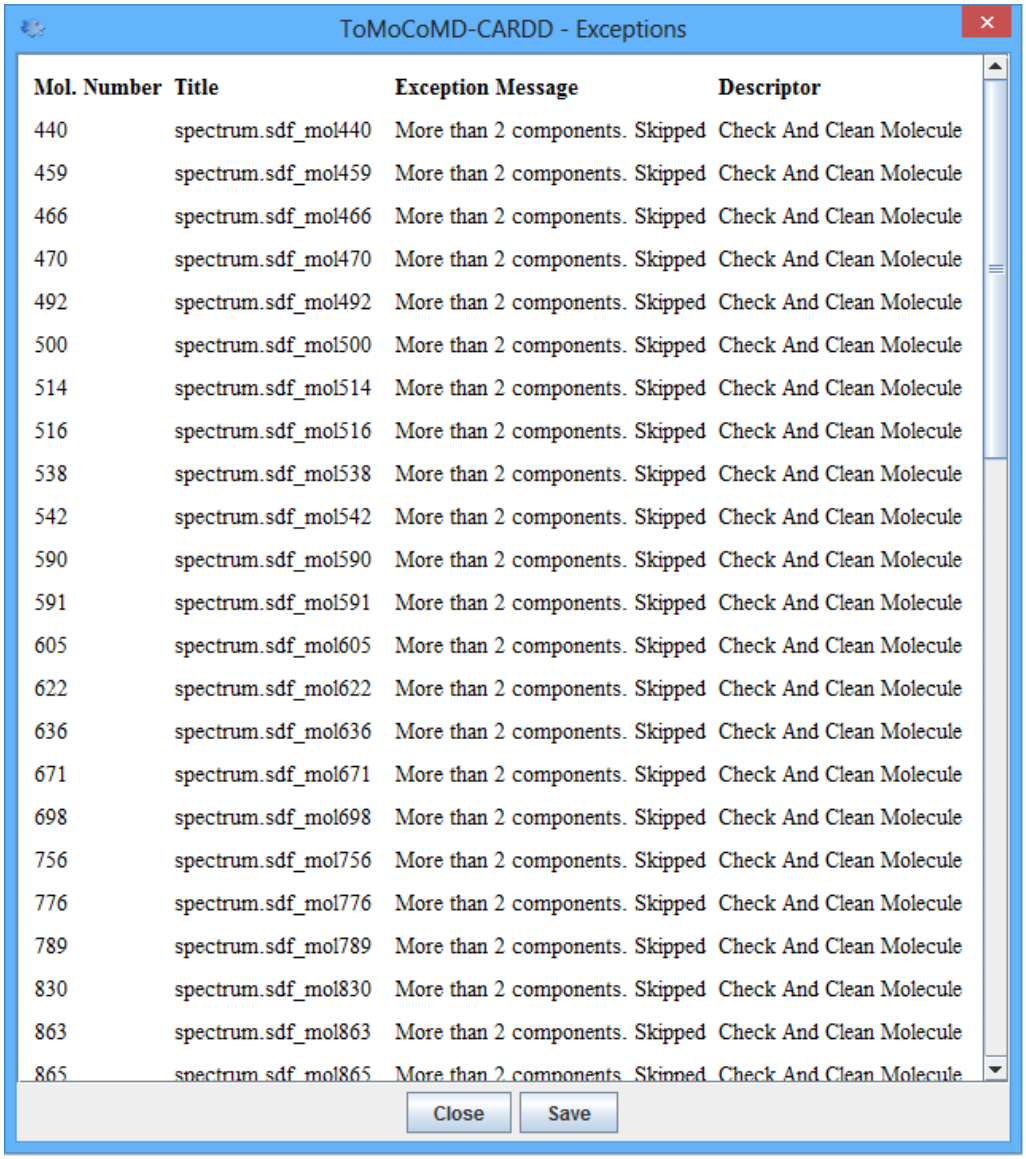

Figure 54. Exception window.