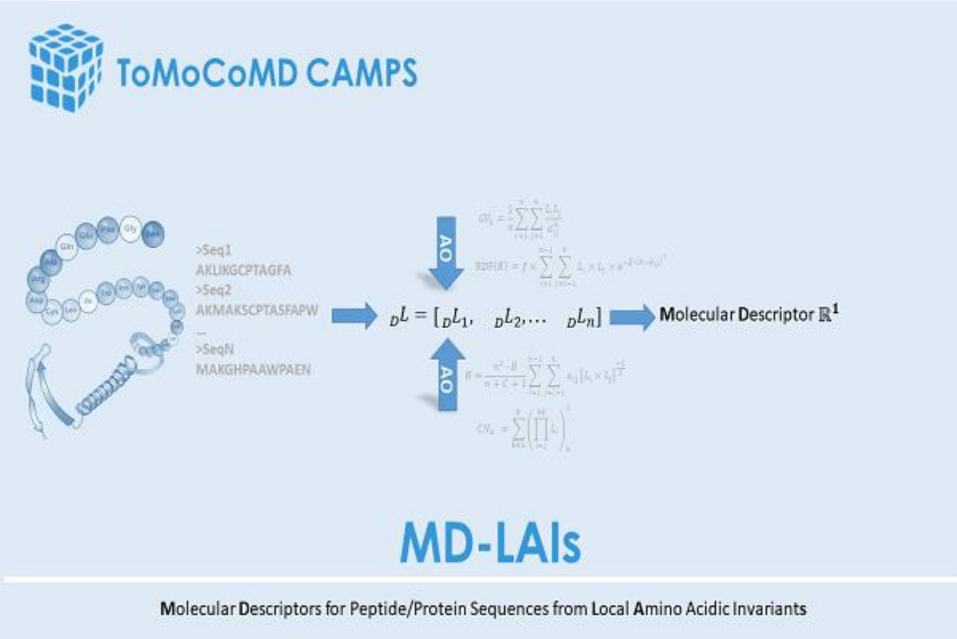
ToMoCoMD-CAMPS

a protein/peptide descriptor computation tool

[Suite-Module]
MD-LAIs v1.0


ToMoCoMD CAMPS
>Seq1
AKLIKGCPTAGFA
>Seq2
AKMAKSCPTASFAPW
...
>SeqN
MAKGHPAAWPAEN
AO
AO
$_{D}L = [_{D}L_1, \quad _{D}L_2, \dots \quad _{D}L_n]$
Molecular Descriptor $\mathbb{R}^1$
MD-LAIs
Molecular Descriptors for Peptide/Protein Sequences from Local Amino Acidic Invariants

Software User Manual

MD-LAIs (acronym of Molecular Descriptors from Amino Acid Invariants) is a biopolymer (protein and peptide) descriptors computing software developed under the umbrella of the **ToMoCoMD-CAMPS** suite. ***MD-LAIs*** constitutes a method for calculating MDs for peptides and proteins based on aggregating the local amino acidic contributions. These MDs can be used for a wide range of applications in bioinformatics, in particular, QSAM studies at sequence and residue levels, and similarity/dissimilarity analyses.

USER'S MANUAL

ToMoCoMD-CAMPS
MD-LAIs v1.0

MD-LAIs v1.0 is a program that calculates sequence-based protein/peptide descriptors based on Aggregation Operators.

COLEGIO DE MEDICINA
Universidad San Francisco de Quito
Quito, Ecuador

January, **2023**

User’s Manual
Authorization Memorandum

I have carefully assessed the User’s Manual for **ToMoCoMD-CAMPS (MD-LAIs)**.

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-CAMPS MD-LAIs is an interactive and user-friendly *free* multi-platform application designed to calculate 0/1-D numerical descriptors for proteins and peptides, with the objective of characterizing or discriminating among them. The MDs are obtained by applying aggregation operators, group-based calculations and fuzzy membership-based weights. The MD-LAIs software was developed in the Java programming language and employs the Chemical Development Kit (CDK) and BioJava libraries for the manipulation of the PDB and FASTA formats, respectively. 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 descriptors, while the library was designed to be integrated effortlessly with other software. This program performs the parallel calculation of the descriptors through the use of all available processors.

This version has some relevant features, such as:

1) One input format (FASTA).
2) A module for the curation of the polypeptide datasets prior to descriptor calculations. This module offers capabilities to repair FASTA files with wrong format, as well to handle those files containing non-standard amino acids.
3) Remove duplicates capabilities.
4) Format conversions (e.g., Multi-FASTA to FASTA, PDB to FASTA, Multi-Fasta to CSV and so on).
5) Sequence edition allowing the extraction of relevant subsequences (e.g., N-terminal, Middle and C-Terminal).
6) Eight datasets are provided as Example Data.
7) Forty-one aggregation operators (invariants) that generalize the way of obtaining MDs from amino acid (or fragment) contributions. The proposed MDs are obtained by applying several invariants to LAIs (Local Amino Acid Invariants).
8) Thirty-nine properties as amino acid labels: Mass, Volume, Z-scales, MD-LOVIS scales, QUBILS-MIDAS scales and son on.
9) A new amino acid position ponderation based on Fuzzy Membership Functions.
10) Thirty group-based MDs classified into in three categories: *Chemical-structural*: Apolar, Polar positively charged, Polar negatively charged, Alpha-helix favoring amino-Acids, Beta-Sheets favoring amino-Acids and so on and *R group* (one for each α-amino acid) including Alanine, Arginine and so on. 8400 *K*-mer based groups (400 dipeptides and 8000 tripeptides).
11) Amino acid-level MD calculation using a sliding-window centered at each amino acid.
12) Configuration/calculation of molecular descriptors via XML files.
13) Calculation of molecular descriptors from their headings.
14) Three output file formats: Space Delimited Text file, Weka ARFF file and Comma Separated Values file.
15) A module for descriptor selection including two filters based on Shannon-Entropy and Pearson (or Spearman) Correlation.
16) Notification and information on system error and program exceptions of JRE.

17) Real-time updated logging status (see Logs Tab).
18) Optional generation of Debug Report file.
19) Descriptor Search Tool.
20) CPU/Memory Manager.
21) Enhanced speed for descriptor calculation process with more stability and robustness.

System requirements

MD-LAIs 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. 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 strongly recommend 4096 MB or above for an optimal performance.

Software:

Operating system: **ToMoCoMD-CAMPS MD-LAIs** 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, 8, 10 and 11. **ToMoCoMD-CAMPS** is platform-independent software.

Operating system extensions: **ToMoCoMD-CAMPS** requires Java (TM) 8 Runtime Environment or above on the target system. It runs under any host operating system, which supports Java(TM) 8 Runtime Environment and also works on x86 and x64 based architectures.

Points of Contact

Information

For all comments, suggestions, information, and inquiries about **ToMoCoMD-CAMPS MD-LAIs** software please contact:

Prof. Yovani Marrero Ponce, PhD
Ecuador de Medicina
Universidad de San Francisco de Quito, Quito
Ecuador.
E-mail: ymarrero77@yahoo.es
URL: http://www.uv.es/yoma/
ORCID: https://orcid.org/0000-0003-2721-1142

Technical Support

For technical support please contact to:

Ernesto Contreras Torres, M.Sc.
BCAM-Basque Center for Applied Mathematics.
Basque Country, Bilbao
Spain.
E-mail: econtrerastorres88@gmail.com
ORCID: https://orcid.org/0000-0003-4761-1784

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 MDs with this application begin on the "Projects” panel, located under the tabbed pane menu or simply can be loaded from a preconfigured MD-LAIs’s project file or just selecting a predefined list of headings withing the tab “List”.

Installation of the program

ToMoCoMD-CAMPS MD-LAIs suite is available as a Java^{TM} standalone and portable application: *MD-LAIs.jar*, for users that frequently use different workstations. No matter the operating system or workstation hardware configuration, ToMoCoMD-CAMPS MD-LAIs users always will have this software at one click away.

3.0 GETTING STARTED

3.0 GETTING STARTED

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

Loading application

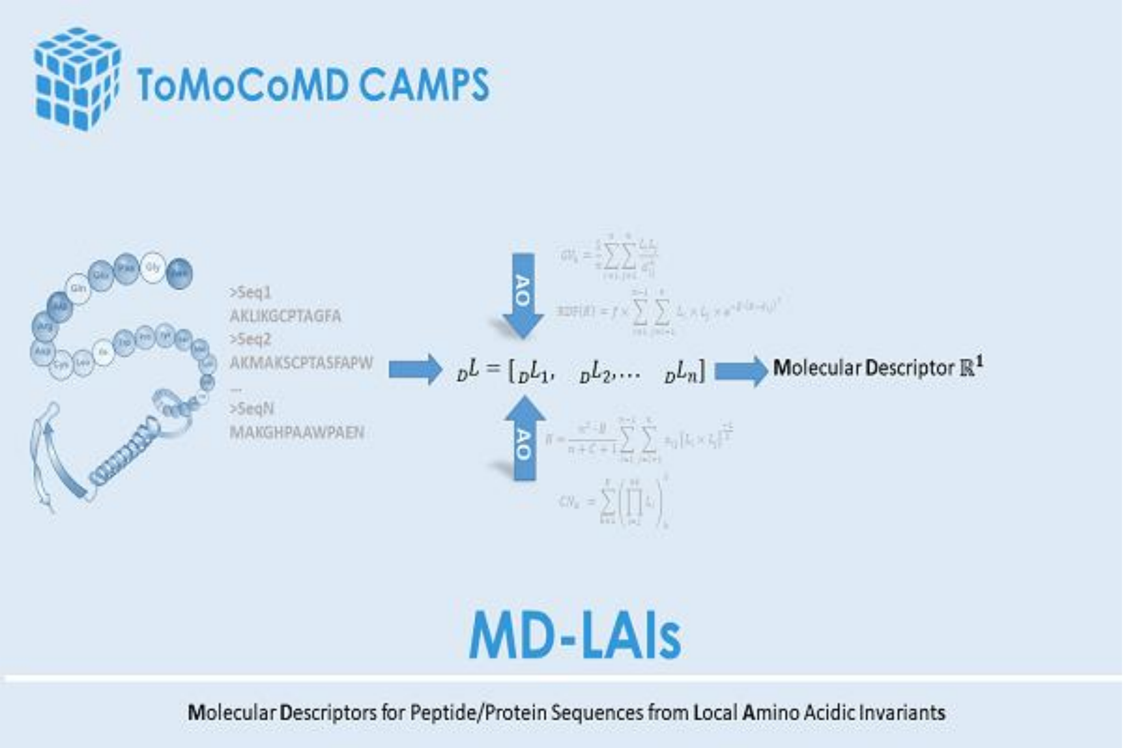

ToMoCoMD CAMPS
>Seq1
AKLIKGCPTAGFA
>Seq2
AKMAKSCPTASFAPW
...
>SeqN
MAKGHPAAWPAEN
AO
AO
$_{D}L = [_{D}L_1, \ _{D}L_2, \dots \ _{D}L_n]$
Molecular Descriptor $\mathbb{R}^1$
MD-LAIs
Molecular Descriptors for Peptide/Protein Sequences from Local Amino Acidic Invariants

Figure 1. Loading SplashScreen for ToMoCoMD-CAMPS (MD-LAIs)

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 immediately.

MD-LAIs Graphical User Interface (GUI)

*The **MD-LAIs** GUI has the following screen areas*:

- **Title Bar: Bears the title of program,** ToMoCoMD-CAMPS MD-LAIs.
- **Menu Bar:** Menus related to different tasks performed by ToMoCoMD-CAMPS MD-LAIs.

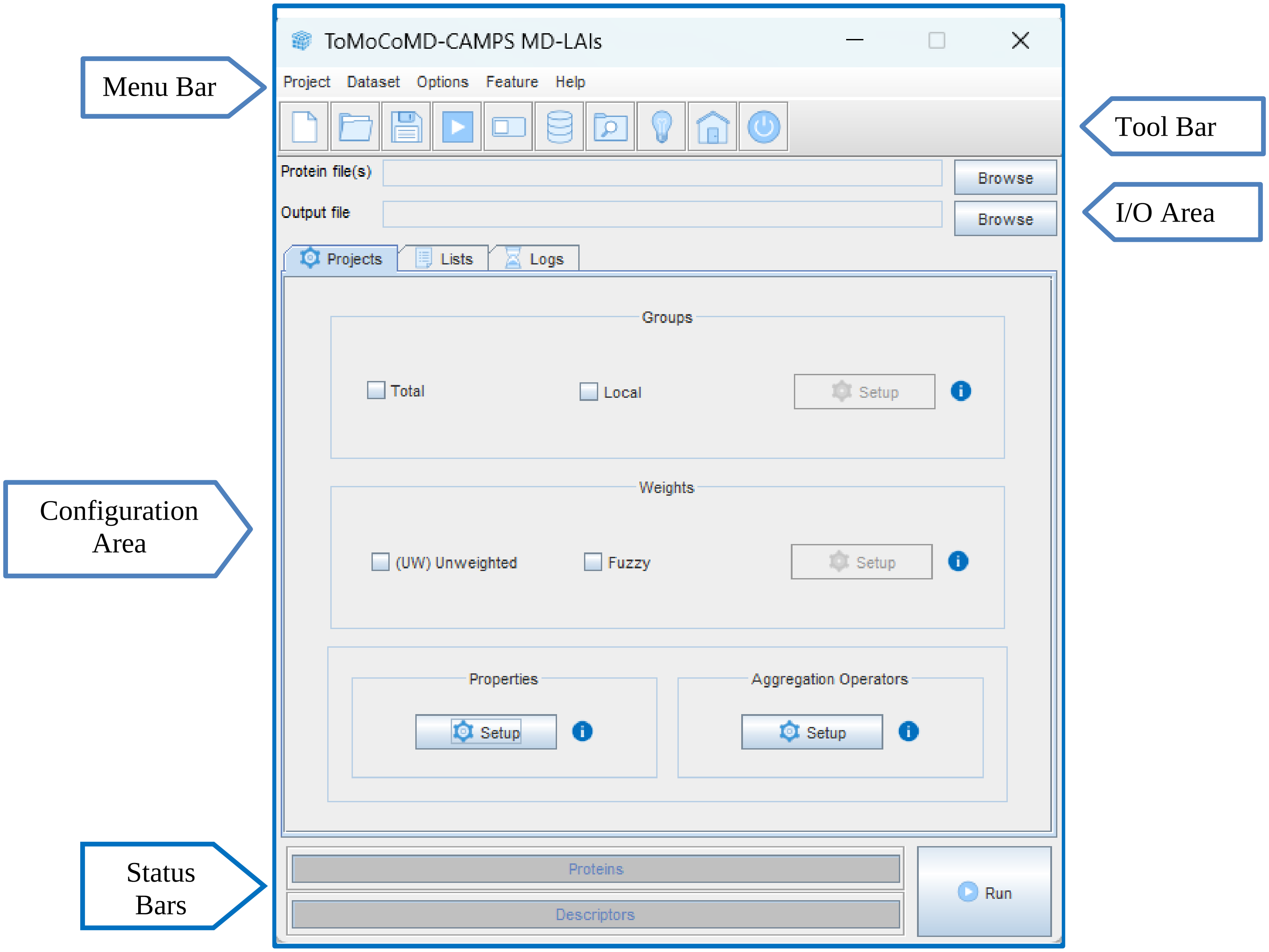


Figure 2. The MD-LAIs main GUI

- **Tool Bar:** Quick access shortcuts to commonly performed tasks, displayed as graphical icons instead of classical menu items.
- **Input/Output Area:** Configure input/output files.
- **Status Bar:** Shows the current and remaining proteins and descriptors.
- **Configuration Area:** This is the *main client area*, which contains the tabs Projects and Lists. The tab *Project*s allow to configure the MDs and the tab *Lists* allows to compute MDs from their headings.
- **Logs Tab:** Record of all operations and tasks.

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.

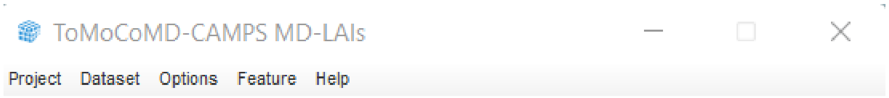


Figure 3. System Menu Bar

Project menu commands

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

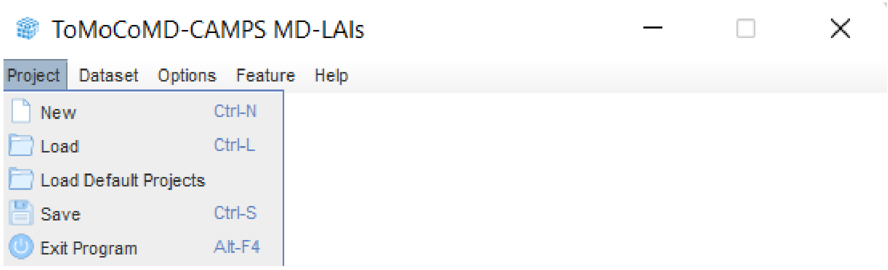


Figure 4. The Project menu

New

Creates a brand-new empty configuration or resets the current configuration options of the Project Manager Tab Panel.

Load

Import configuration and options from a Project Configuration File.

Load Default Projects

Import configuration and options from default (suggested) Project Configuration Files.

Save

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

Exit Program

Close the application.

Dataset menu commands

Commands of the *Dataset menu* allow the user to open/edit existing dataset files.

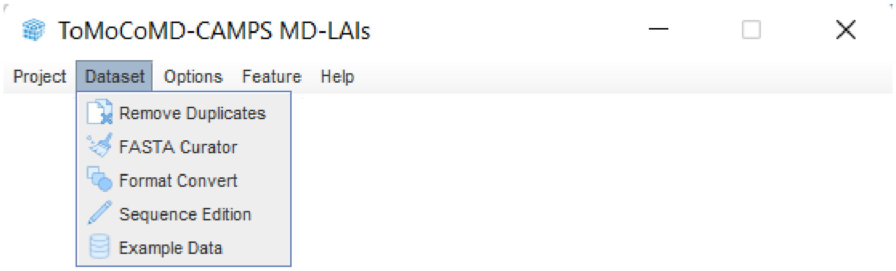


Figure 5. The Dataset menu

Remove Duplicates

Remove duplicated sequences from the specified datasets either by the identifier or sequence.

FASTA Curator

Provide options to work with non-standard amino acids or files with erroneous format.

Format Convert

Options to convert among the most common formats.

Sequence Edition

Options to extract specific segments of the sequence.

Example Data

Show the example datasets.

Options menu commands

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

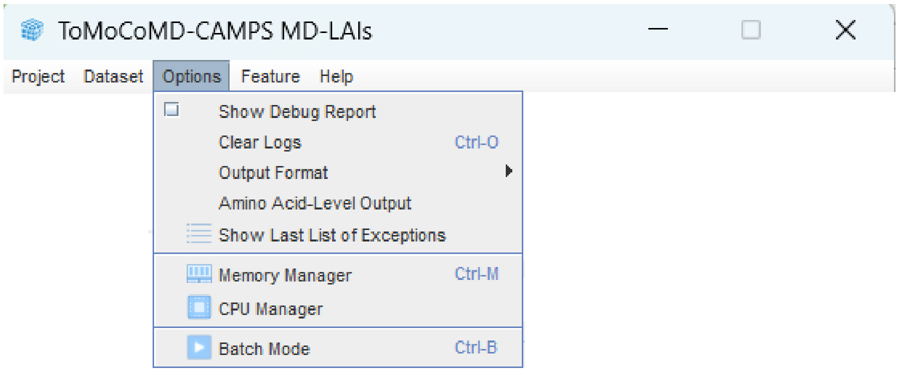


Figure 4. The Option menu

Show Debug Report

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

Clear Logs

Cleans the Logs window.

Output Method

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

Amino Acid-Level Output

Shows a dialog to configure the amino acid-level output.

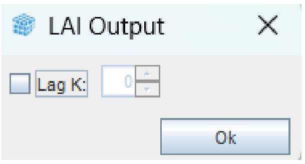


Figure 6. LAI Output Window

Show Last List of Exceptions

Displays the exceptions occurred during the calculation of the MDs.

Memory manager

Displays a window that shows Random Access Memory (RAM) (measured in Megabytes (MB)) employed by the program.

CPU manager

Displays a window to set the number of CPU cores to use in the calculation of the MDs.

Batch manager

Shows a window to work in batch mode.

Help menu commands

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

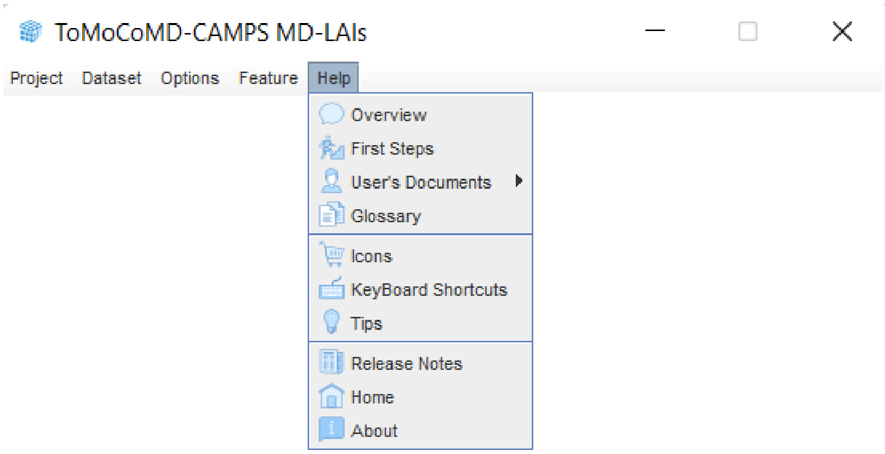


Figure 7. The Help menu

Overviews

Shows an illustrative procedure of how to execute, configure and use the program. MD-LAIs is based on the 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 **MD-LAIs** 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

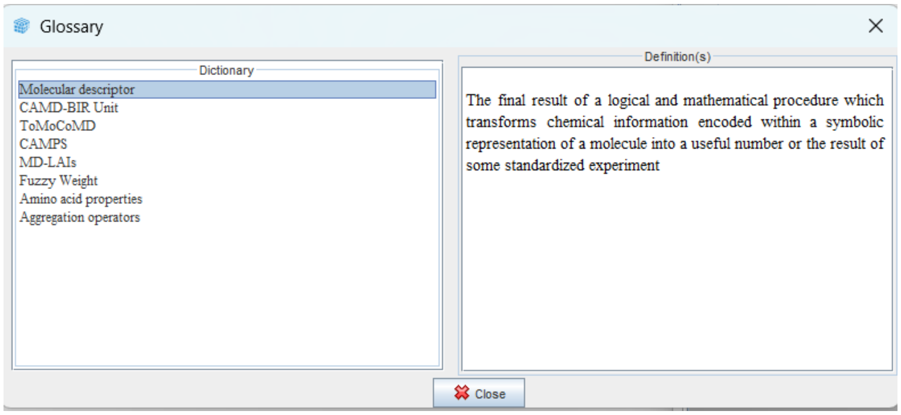
Introduce a first time **ToMoCoMD-CAMPS** user into a general overview, system requirements and installation process, input and output file modes.

User's Documents

Present three sub-menus related to the **MD-LAIs** descriptors: Acronyms, Theory and the User's Manual corresponding to this application.

Glossary

Basic **MD-LAIs** terminology is provided in a tool. This terminology is provided in a window that the user can keep open to be supported through the **MD-LAIs** indices setup.



Dictionary

Molecular descriptor
CAMD-BIR Unit
ToMoCoMD
CAMPS
MD-LAIs
Fuzzy Weight
Amino acid properties
Aggregation operators

Definition(s)

The final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment

Figure 8. 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-CAMPS'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 9 tips that provide instant technical information any time you execute this program.

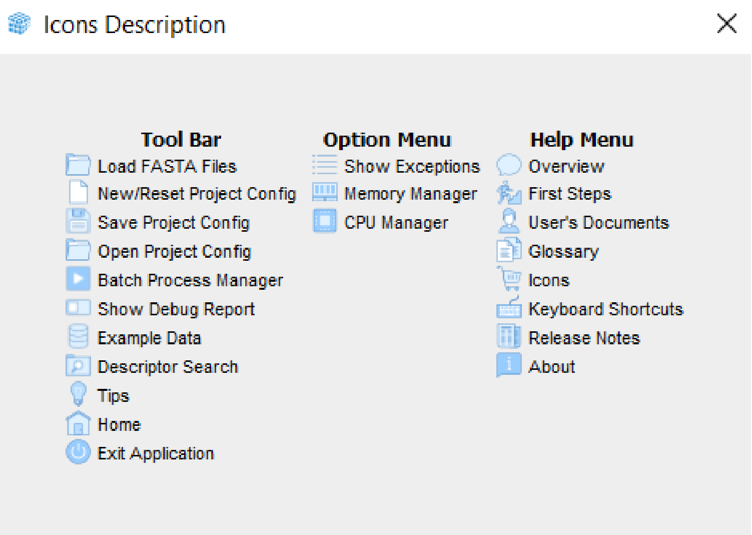
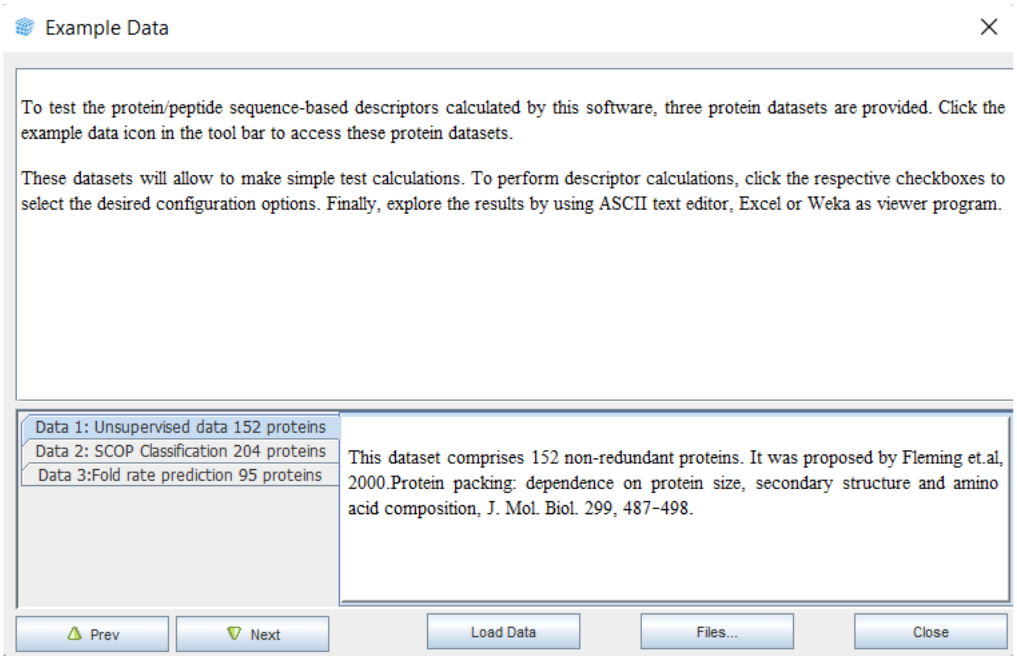


Figure 9. Icons description.

Example Data

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



To test the protein/peptide sequence-based descriptors calculated by this software, three protein datasets are provided. Click the example data icon in the tool bar to access these protein datasets.

These datasets will allow to make simple test calculations. To perform descriptor calculations, click the respective checkboxes to select the desired configuration options. Finally, explore the results by using ASCII text editor, Excel or Weka as viewer program.

Data 1: Unsupervised data 152 proteins
Data 2: SCOP Classification 204 proteins
Data 3:Fold rate prediction 95 proteins

This dataset comprises 152 non-redundant proteins. It was proposed by Fleming et.al, 2000.Protein packing: dependence on protein size, secondary structure and amino acid composition, J. Mol. Biol. 299, 487-498.

Figure 10. Example Data Tool.

Release Notes

Show the track changes since **MD-LAIs** was a whiteboard idea.

Home

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

Thanks

Recognition to different contributions to this project.

About

Several information about **MD-LAIs** 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 **MD-LAIs** menu commands. Clicking on toolbar icons enables the user to perform the following commands:

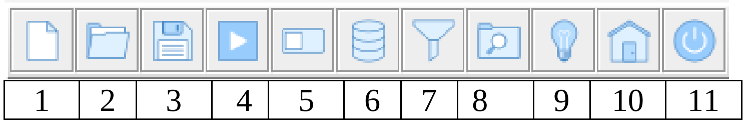


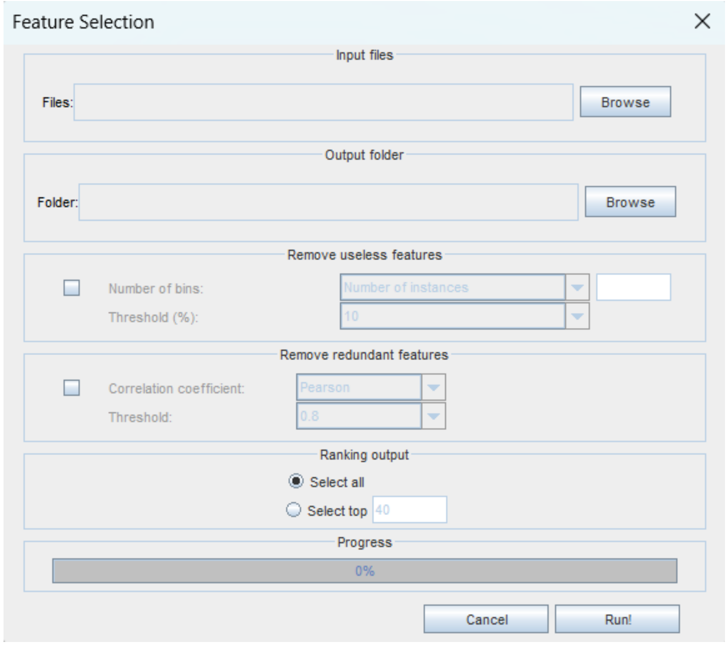
Figure 11. Tool Bar elements.

1. New Project
2. Save

3. Open
4. Batch Manager
5. On/Off Generate Debug Report
6. Launch Example data
7. Feature Selection
8. Descriptor Search Tool
9. Tips
10. Home
11. Exit Program

Feature Selection

MD-LAIs includes a module to feature selection. This module includes two unsupervised filters to retain relevant and low redundant MDs based on Shannon-Entropy and Pearson/ Spearman correlation coefficients, respectively. *See the picture below*.



Feature Selection

Input files

Files:

Output folder

Folder:

Remove useless features

□ Number of bins: Number of instances

Threshold (%): 10

Remove redundant features

□ Correlation coefficient: Pearson

Threshold: 0.6

Ranking output

● Select all

○ Select top 40

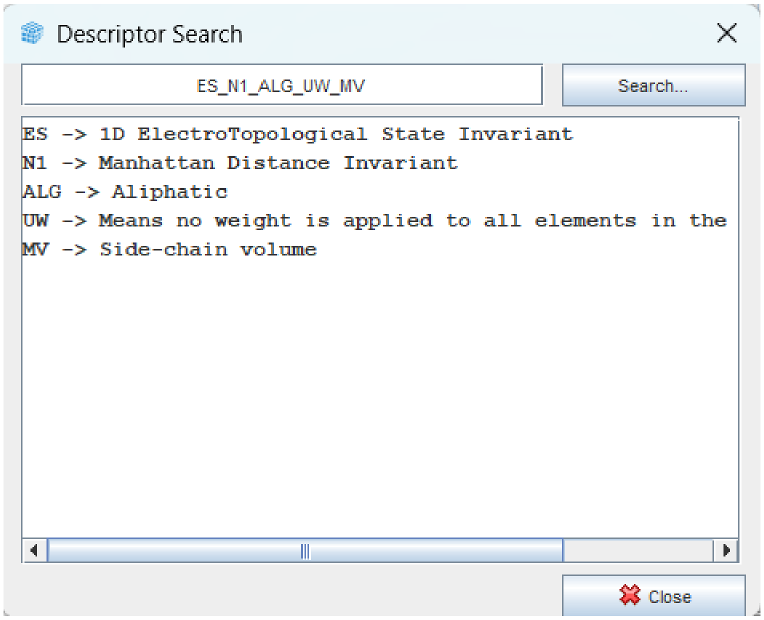
Progress

0%

Figure 12. Feature Selection Window.

Descriptor Search

MD-LAIs includes an optional use tool developed to automatically decode headers assigned to each one of the MDs. *See the picture below*.



```
ES -> 1D ElectroTopological State Invariant
N1 -> Manhattan Distance Invariant
ALG -> Aliphatic
UW -> Means no weight is applied to all elements in the
MV -> Side-chain volume
```

Figure 13. Dialog Search Window.

Status Bars

The status bars located at the bottom of the main window show the protein file name, the descriptor that is being calculated and the percentage of completion, also display the name and number for proteins and descriptors.

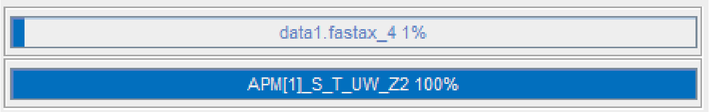
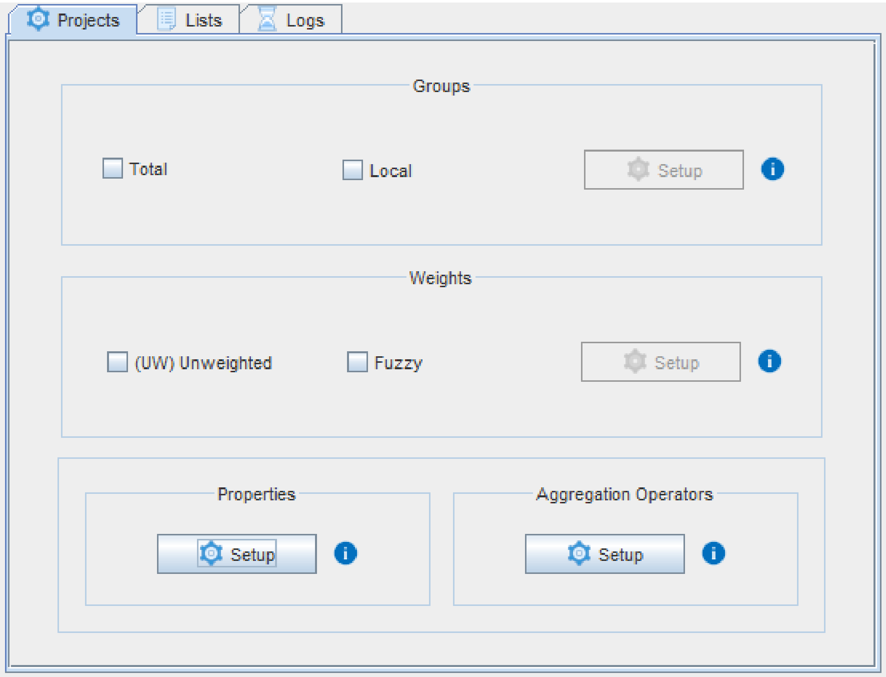


Figure 13. Status Bars.

Projects Tab

This area has seven sub-areas (panes) for MDs configuration. In each pane appears an info button (blue circle) that contains a short description about the theory associated to each part (for more details see **Starting MD-LAIs** and **Configuring a project** sections).



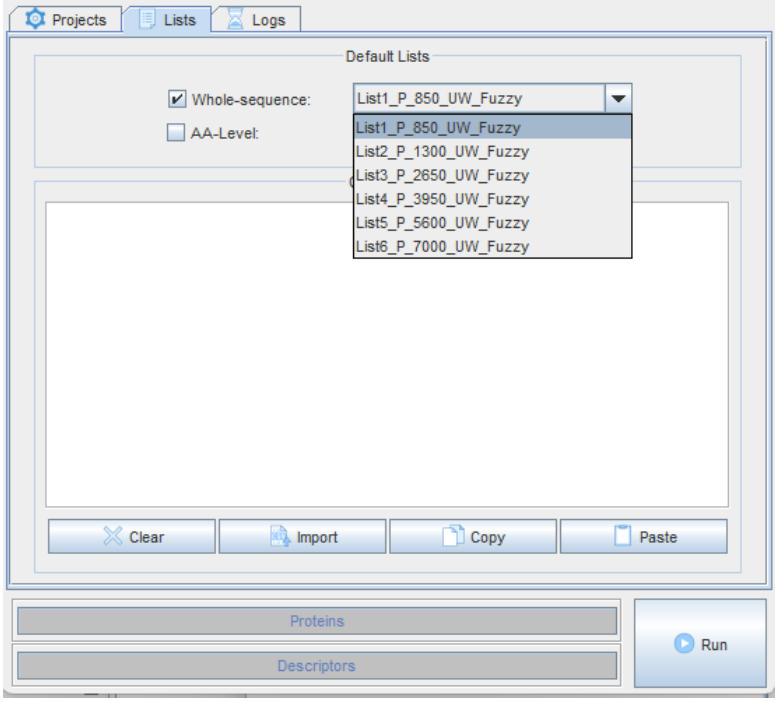
□ Total
□ Local

□ (UW) Unweighted
□ Fuzzy

Figure 14. Descriptors Configuration Area.

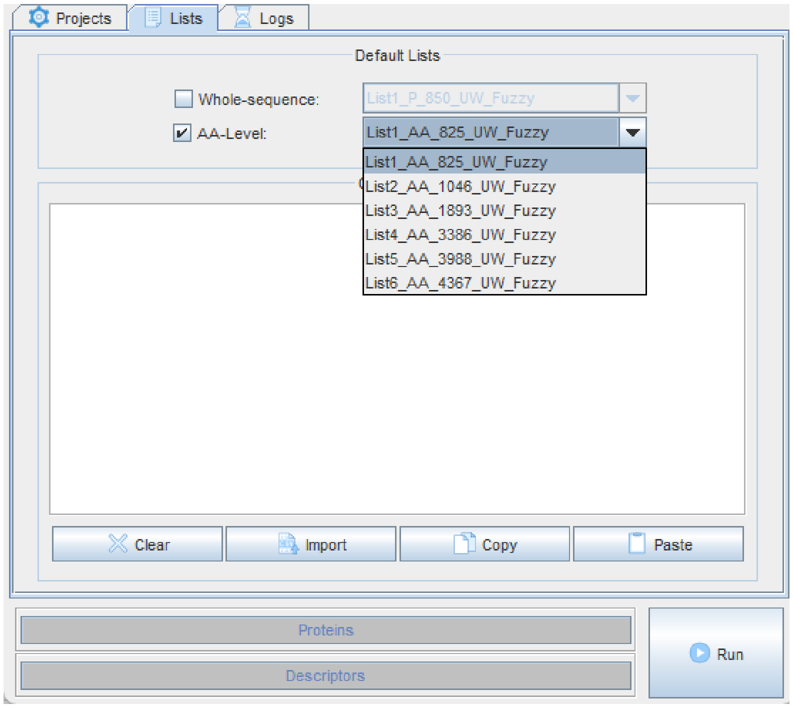
Lists Tabs

This panel allows for MDs computation from their headings. Just paste (or load from a file or use default) the headings of the descriptors that one want to compute in the textfield and press the Run button.



☑ Whole-sequence: List1_P_850_UW_Fuzzy
□ AA-Level:

Figure 14. A) Selecting Default Whole-Sequence MDs List.

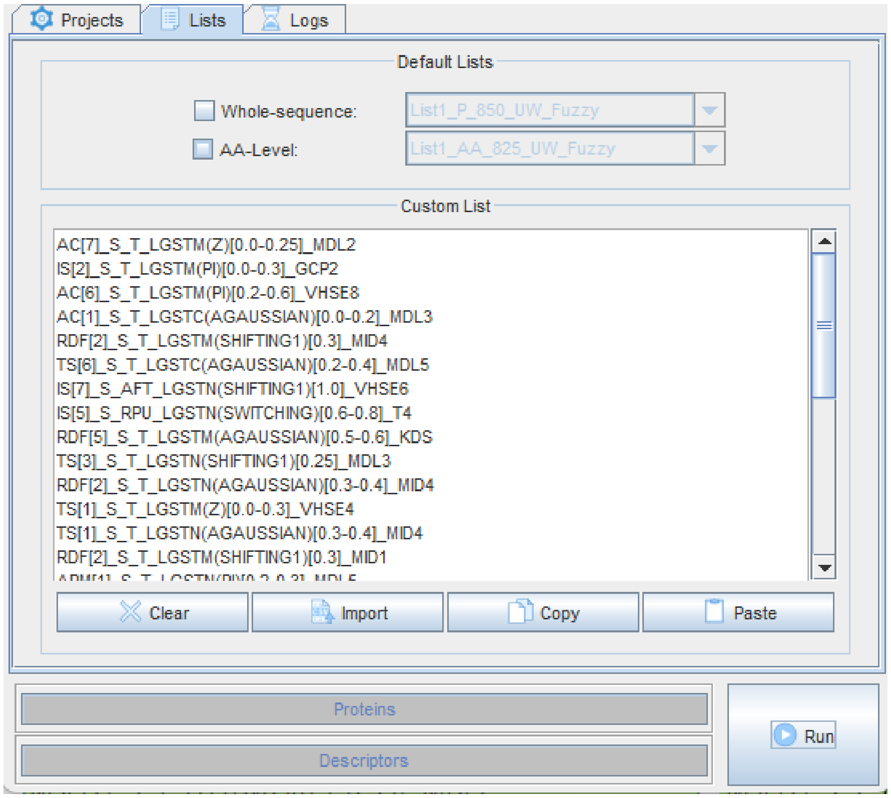


☐ Whole-sequence: List1_P_850_UW_Fuzzy

☑ AA-Level: List1_AA_825_UW_Fuzzy

```
List1_AA_825_UW_Fuzzy
List2_AA_1046_UW_Fuzzy
List3_AA_1893_UW_Fuzzy
List4_AA_3386_UW_Fuzzy
List5_AA_3988_UW_Fuzzy
List6_AA_4367_UW_Fuzzy
```

Figure 14. B) Selecting Default Amino Acid-Level MDs List.



☐ Whole-sequence: List1_P_850_UW_Fuzzy

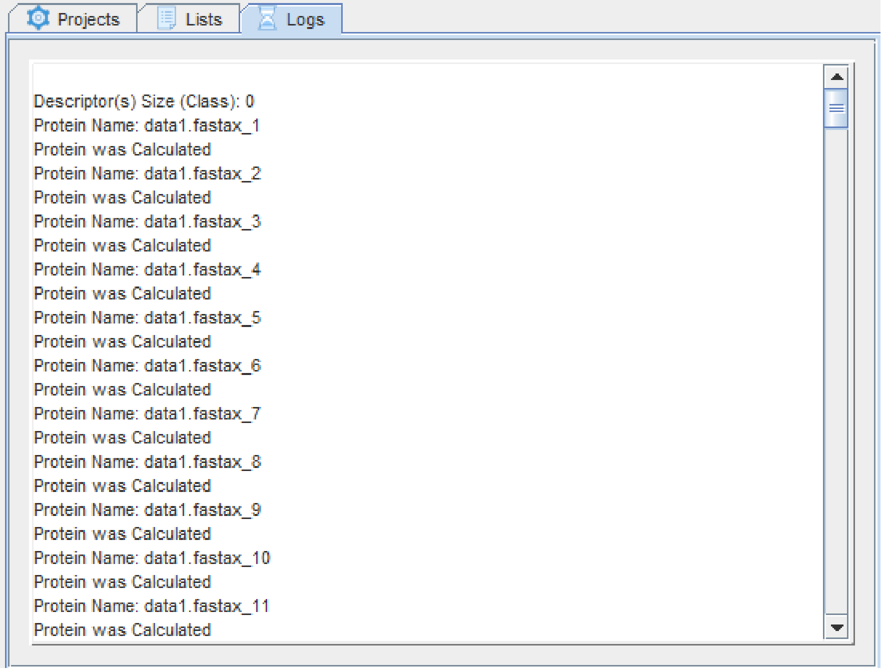
☐ AA-Level: List1_AA_825_UW_Fuzzy

```
AC[7]_S_T_LGSTM(Z)[0.0-0.25]_MDL2
IS[2]_S_T_LGSTM(PI)[0.0-0.3]_GCP2
AC[6]_S_T_LGSTM(PI)[0.2-0.6]_VHSE8
AC[1]_S_T_LGSTC(AGAUSSIAN)[0.0-0.2]_MDL3
RDF[2]_S_T_LGSTM(SHIFTING1)[0.3]_MID4
TS[6]_S_T_LGSTC(AGAUSSIAN)[0.2-0.4]_MDL5
IS[7]_S_AFT_LGSTN(SHIFTING1)[1.0]_VHSE6
IS[5]_S_RPU_LGSTN(SWITCHING)[0.6-0.8]_T4
RDF[5]_S_T_LGSTM(AGAUSSIAN)[0.5-0.6]_KDS
TS[3]_S_T_LGSTN(SHIFTING1)[0.25]_MDL3
RDF[2]_S_T_LGSTN(AGAUSSIAN)[0.3-0.4]_MID4
TS[1]_S_T_LGSTM(Z)[0.0-0.3]_VHSE4
TS[1]_S_T_LGSTN(AGAUSSIAN)[0.3-0.4]_MID4
RDF[2]_S_T_LGSTM(SHIFTING1)[0.3]_MID1
```

Figure 14. C) Selecting Custom MDs List.

History Window

Logging windows for all operations and task. Besides, after the calculation is finished, the **tab Logs** shows some details and statistics of the calculation process.



```
Descriptor(s) Size (Class): 0
Protein Name: data1.fastax_1
Protein was Calculated
Protein Name: data1.fastax_2
Protein was Calculated
Protein Name: data1.fastax_3
Protein was Calculated
Protein Name: data1.fastax_4
Protein was Calculated
Protein Name: data1.fastax_5
Protein was Calculated
Protein Name: data1.fastax_6
Protein was Calculated
Protein Name: data1.fastax_7
Protein was Calculated
Protein Name: data1.fastax_8
Protein was Calculated
Protein Name: data1.fastax_9
Protein was Calculated
Protein Name: data1.fastax_10
Protein was Calculated
Protein Name: data1.fastax_11
Protein was Calculated
```

Figure 26. Log (History) tab.

Exit System

Describe the actions to properly exit the system.

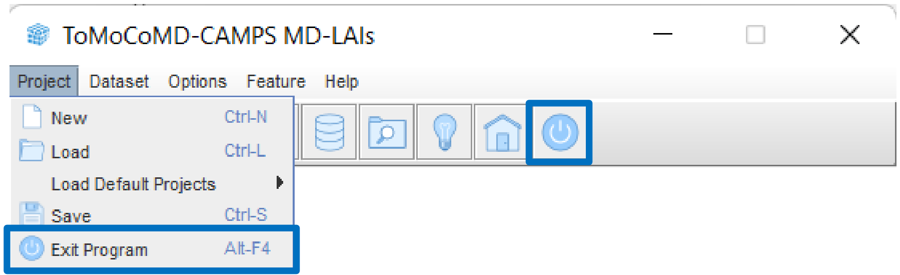
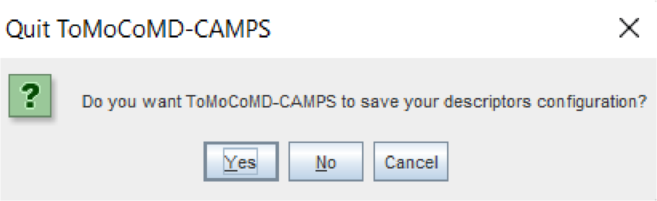


Figure 15. Program Exit Options.



Quit ToMoCoMD-CAMPS

Do you want ToMoCoMD-CAMPS to save your descriptors configuration?

Figure 16. Safe Exit Action prompt.

4.0 USING THE SYSTEM

4.0 USING THE SYSTEM

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

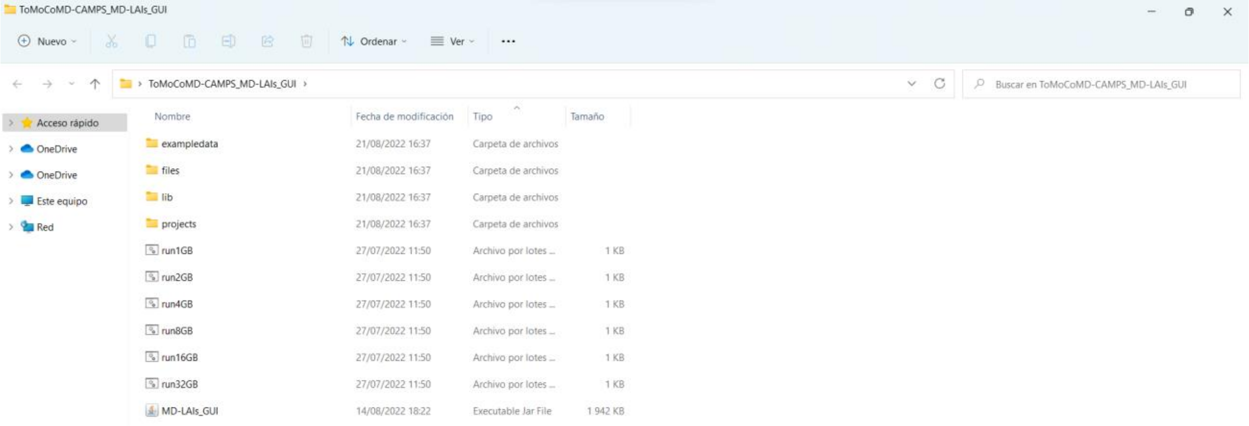
What you need to know before using MD-LAIs

To make use of MD-LAIs calculations, you must provide the FASTA files as input. **MD-LAIs** is not planned as QSAM software; it provides only molecular descriptors and does not perform QSAM analysis. However, by **MD-LAIs** it is possible to select a subset of relevant and non-redundant molecular descriptors for a set of molecules, providing a complete output file which is easily loaded by any correlation analysis application.

Starting MD-LAIs

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

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


ToMoCoMD-CAMPS_MD-LAIs_GUI
Nuevo
Ordenar
Ver
ToMoCoMD-CAMPS_MD-LAIs_GUI
Buscar en ToMoCoMD-CAMPS_MD-LAIs_GUI
Acceso rápido
OneDrive
OneDrive
Este equipo
Red

Nombre	Fecha de modificación	Tipo	Tamaño
exampledata	21/08/2022 16:37	Carpeta de archivos	
files	21/08/2022 16:37	Carpeta de archivos	
lib	21/08/2022 16:37	Carpeta de archivos	
projects	21/08/2022 16:37	Carpeta de archivos	
run1GB	27/07/2022 11:50	Archivo por lotes ...	1 KB
run2GB	27/07/2022 11:50	Archivo por lotes ...	1 KB
run4GB	27/07/2022 11:50	Archivo por lotes ...	1 KB
run8GB	27/07/2022 11:50	Archivo por lotes ...	1 KB
run16GB	27/07/2022 11:50	Archivo por lotes ...	1 KB
run32GB	27/07/2022 11:50	Archivo por lotes ...	1 KB
MD-LAIs_GUI	14/08/2022 18:22	Executable Jar File	1 942 KB

Figure 17. MD-LAIs-MAS Windows Batch Files (.bat).

Indeed, for each heap memory limit, a command line scripts were targeted for two different kinds 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 MD-LAIs_GUI.jar
```

After splash, the main window (GUI) will be displayed on the screen. The follow step is selecting the ***input*** proteins 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 "**Run!**" begins the calculation of the selected descriptors for the structures in the input file. When the calculation is finished, an external dialog window appears showing 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 shows the list of proteins with structural errors. In addition, the **Log 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 **MD-LAIs** session, *e.g.*, Groups, weights, properties, aggregation operators, and so on.

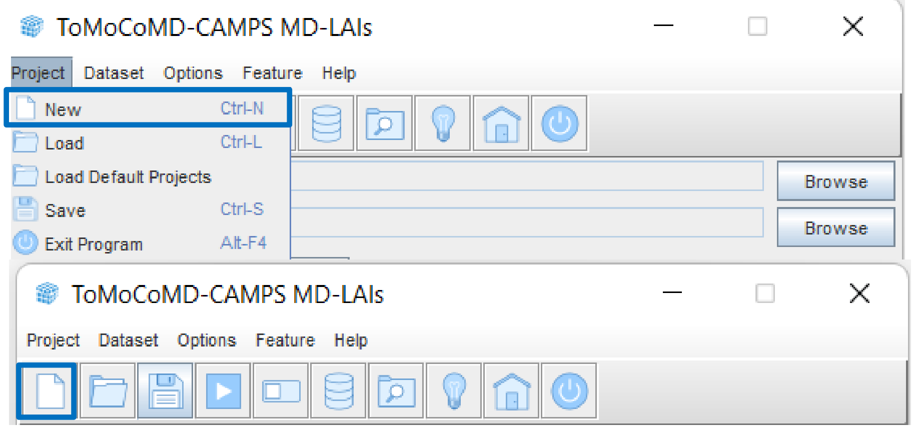


Figure 18. Creating a new project.

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.

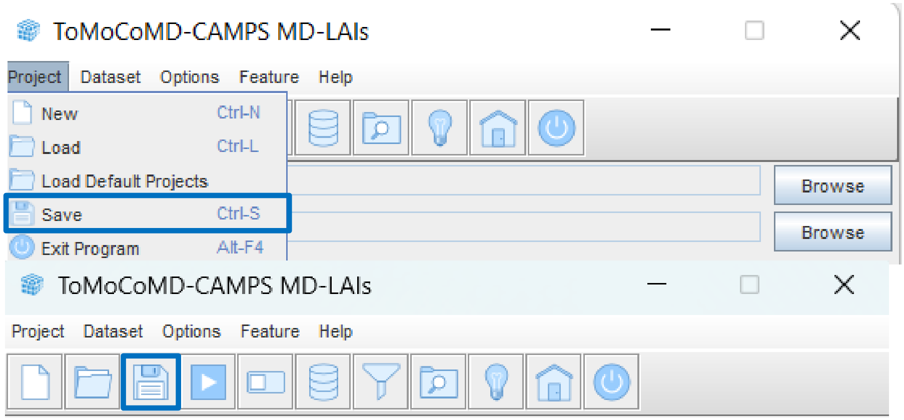


Figure 19. Saving project configuration.

Loading a Project File

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

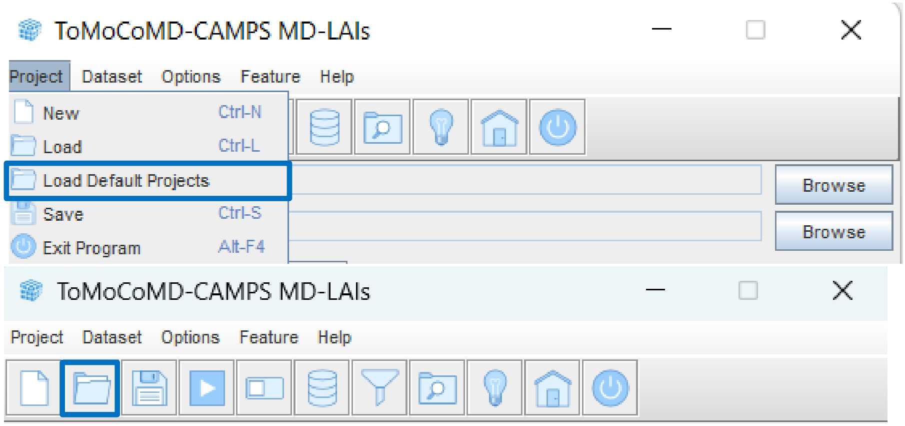


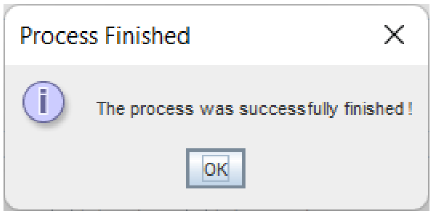
Figure 20. Loading projects button.

Running a saved project

The previously saved project can be reloaded in order to restore all parameters of a previous session. Finally, the user can run the calculation by clicking the button **Run**. 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 **MD-LAIs** GUI. The Calculate section consists of the three buttons **Run/Cancel**, **Exception Window** and **Logs Tab**. The button **Run/ Cancel** begins the calculation of the selected descriptors. When the calculation is started the **Run** 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 protein and the current descriptor that is being calculated and the percentage of completion, also the name and number for proteins and descriptors are displayed. When the calculation is finished a message appears on the screen that displays "***The Process was successfully finished.***"



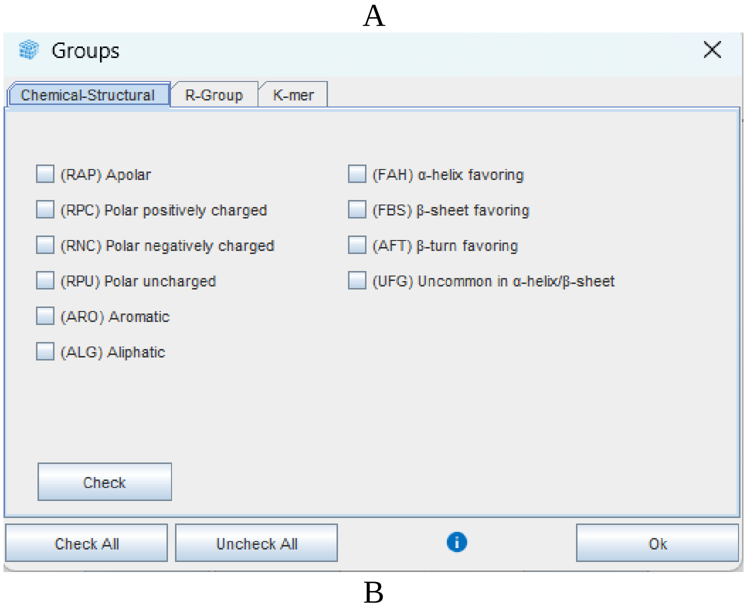
Process Finished

The process was successfully finished!

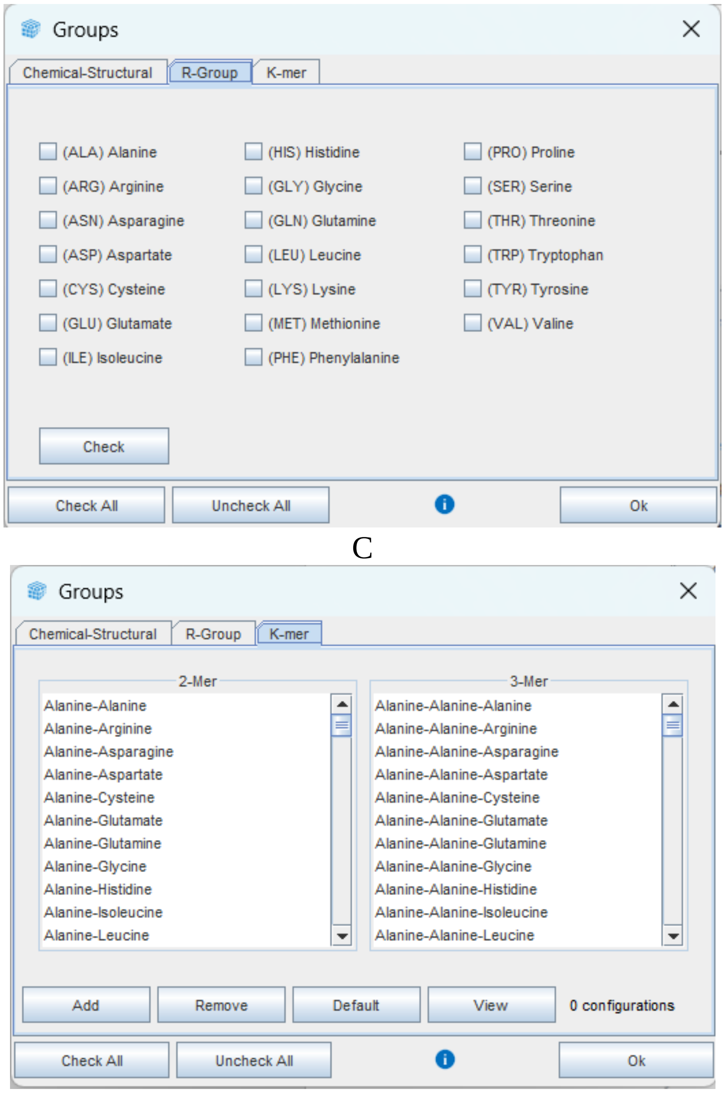
Figure 21. Task finished confirmation window.

Configuring a project

The descriptors section *(Configuration Area)* in the middle part of the **MD-LAIs** GUI consists of four different sub-areas. In each sub-area the user can select the possible parameters by clicking the corresponding button. The sub-areas are: Groups (Figure 22), Weight (Figure 23), Properties (Figure 24) and Aggregation Operators (Figure 25).



☐ (RAP) Apolar
☐ (RPC) Polar positively charged
☐ (RNC) Polar negatively charged
☐ (RPU) Polar uncharged
☐ (ARO) Aromatic
☐ (ALG) Aliphatic
☐ (FAH) α-helix favoring
☐ (FBS) β-sheet favoring
☐ (AFT) β-turn favoring
☐ (UFG) Uncommon in α-helix/β-sheet



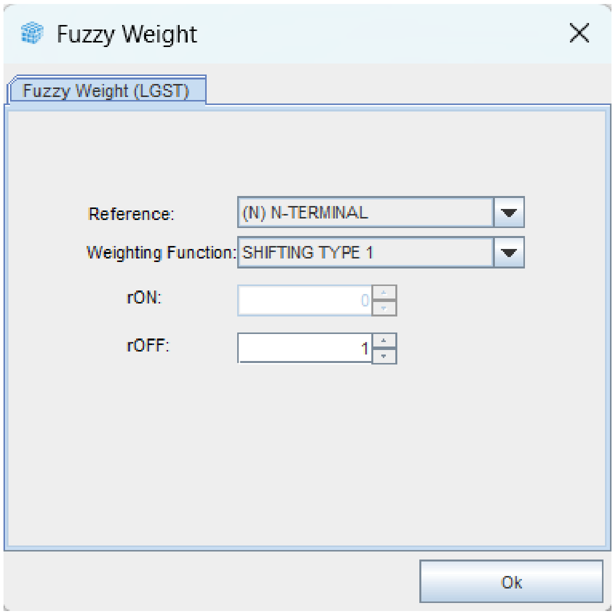
□ (ALA) Alanine
□ (ARG) Arginine
□ (ASN) Asparagine
□ (ASP) Aspartate
□ (CYS) Cysteine
□ (GLU) Glutamate
□ (ILE) Isoleucine
□ (HIS) Histidine
□ (GLY) Glycine
□ (GLN) Glutamine
□ (LEU) Leucine
□ (LYS) Lysine
□ (MET) Methionine
□ (PHE) Phenylalanine
□ (PRO) Proline
□ (SER) Serine
□ (THR) Threonine
□ (TRP) Tryptophan
□ (TYR) Tyrosine
□ (VAL) Valine

C

2-Mer	3-Mer
Alanine-Alanine	Alanine-Alanine-Alanine
Alanine-Arginine	Alanine-Alanine-Arginine
Alanine-Asparagine	Alanine-Alanine-Asparagine
Alanine-Aspartate	Alanine-Alanine-Aspartate
Alanine-Cysteine	Alanine-Alanine-Cysteine
Alanine-Glutamate	Alanine-Alanine-Glutamate
Alanine-Glutamine	Alanine-Alanine-Glutamine
Alanine-Glycine	Alanine-Alanine-Glycine
Alanine-Histidine	Alanine-Alanine-Histidine
Alanine-Isoleucine	Alanine-Alanine-Isoleucine
Alanine-Leucine	Alanine-Alanine-Leucine

0 configurations

Figure 22. Groups. A) Chemical-structural, B) R-group-based and C) K-Mer



Fuzzy Weight

Reference: (N) N-TERMINAL

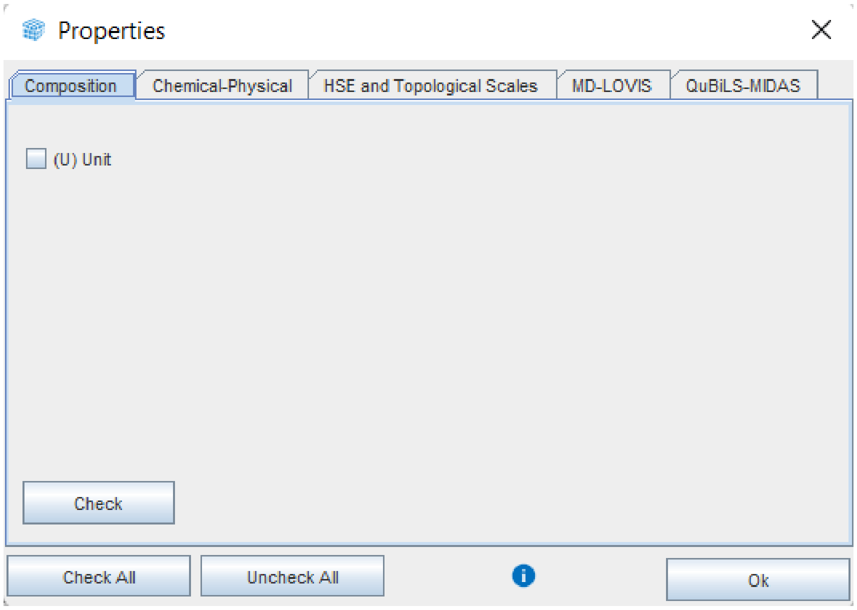
Weighting Function: SHIFTING TYPE 1

rON: 0

rOFF: 1

Figure 23. Dialog window to set the fuzzy membership weight

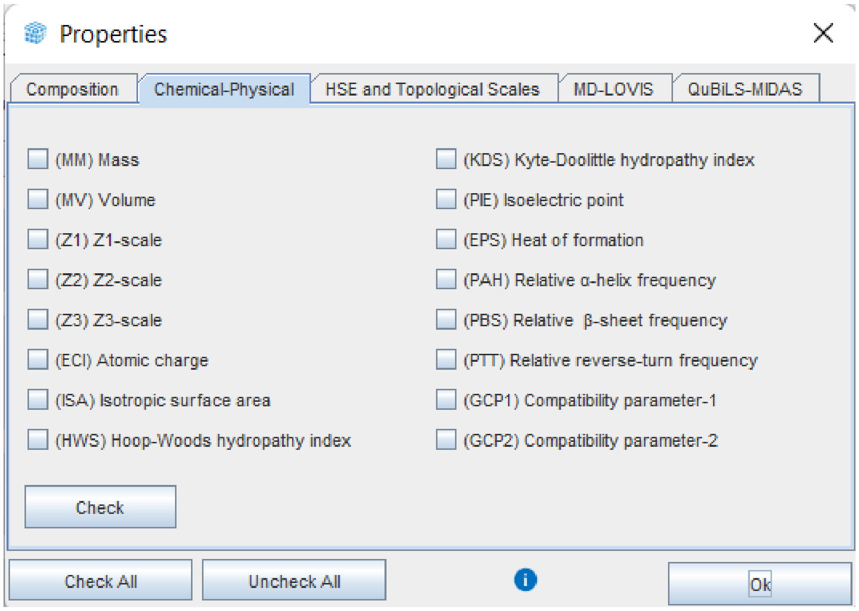
A



Properties

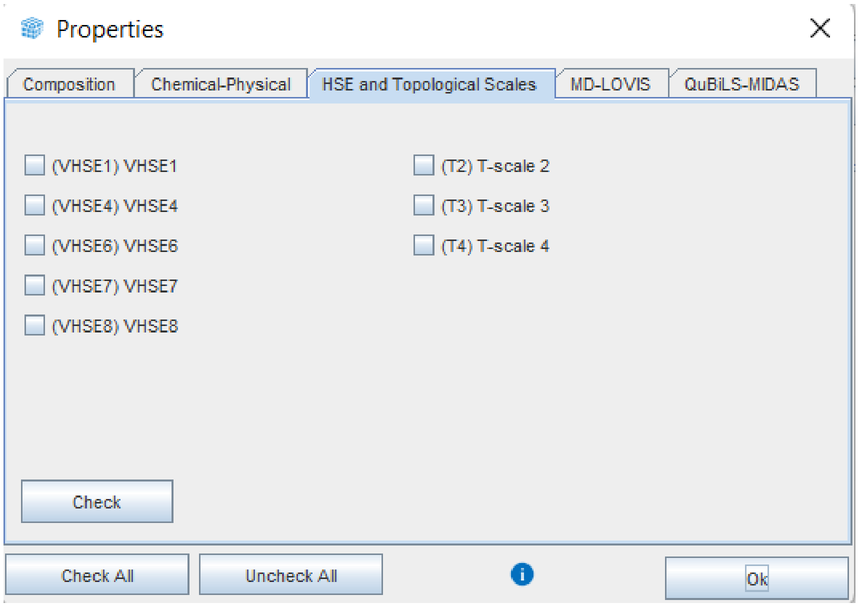
□ (U) Unit

B



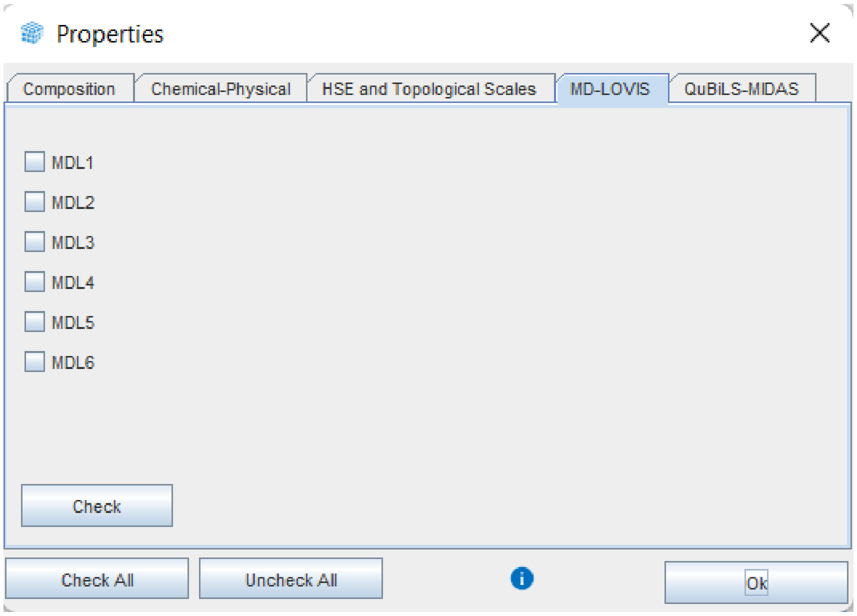
□ (MM) Mass
□ (MV) Volume
□ (Z1) Z1-scale
□ (Z2) Z2-scale
□ (Z3) Z3-scale
□ (ECI) Atomic charge
□ (ISA) Isotropic surface area
□ (HWS) Hoop-Woods hydropathy index
□ (KDS) Kyte-Doolittle hydropathy index
□ (PIE) Isoelectric point
□ (EPS) Heat of formation
□ (PAH) Relative α-helix frequency
□ (PBS) Relative β-sheet frequency
□ (PTT) Relative reverse-turn frequency
□ (GCP1) Compatibility parameter-1
□ (GCP2) Compatibility parameter-2

C



□ (VHSE1) VHSE1
□ (VHSE4) VHSE4
□ (VHSE6) VHSE6
□ (VHSE7) VHSE7
□ (VHSE8) VHSE8
□ (T2) T-scale 2
□ (T3) T-scale 3
□ (T4) T-scale 4

D



E

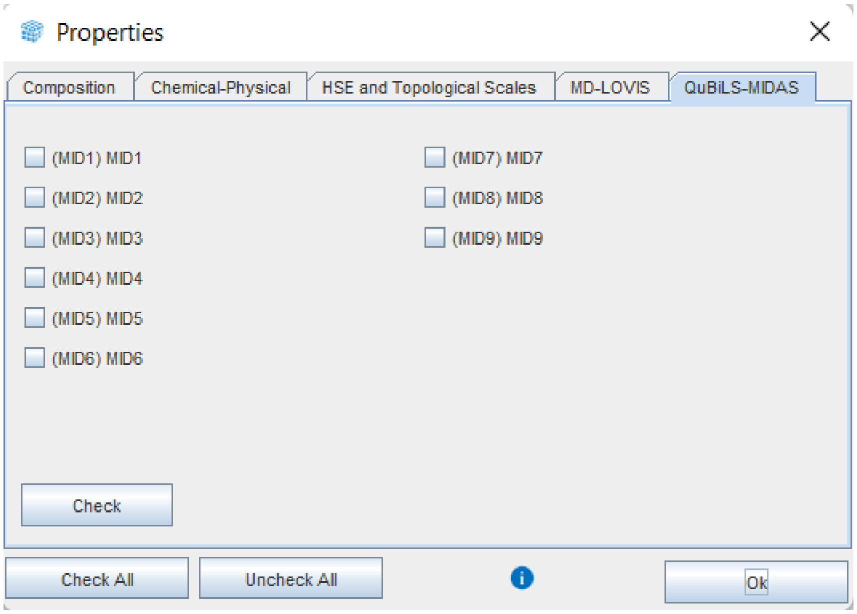
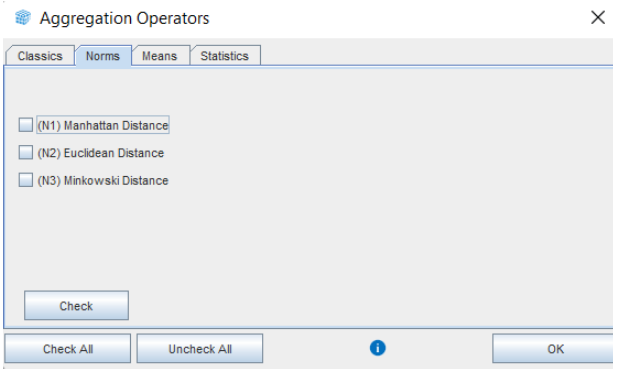
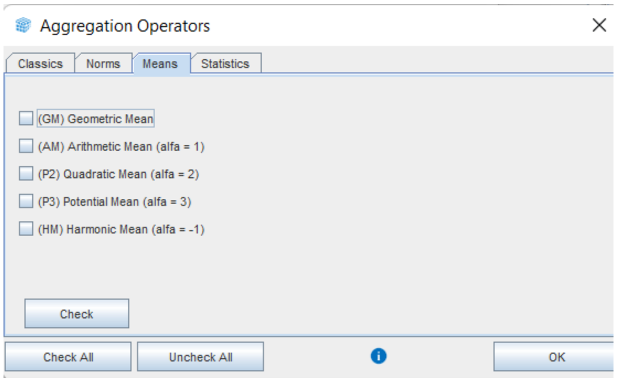


Figure 24. A) Unitary property (composition) , B) Chemical-physical properties, C) Hydrophobic, steric and electronic scales, D) MD-LOVIS scales , E) QuBiLS-MIDAS scales.

A B

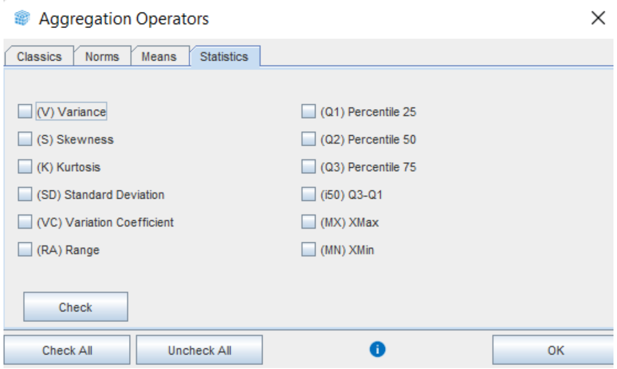


□ (N1) Manhattan Distance
□ (N2) Euclidean Distance
□ (N3) Minkowski Distance



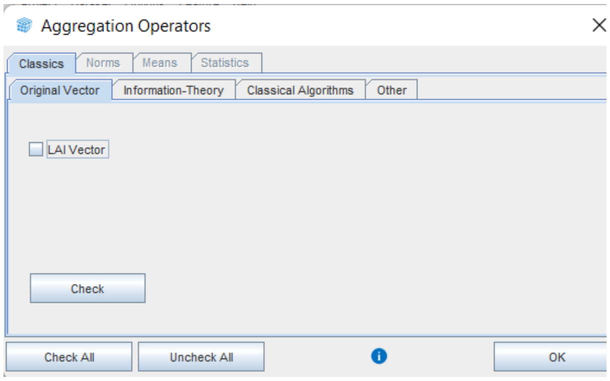
□ (GM) Geometric Mean
□ (AM) Arithmetic Mean (alfa = 1)
□ (P2) Quadratic Mean (alfa = 2)
□ (P3) Potential Mean (alfa = 3)
□ (HM) Harmonic Mean (alfa = -1)

C



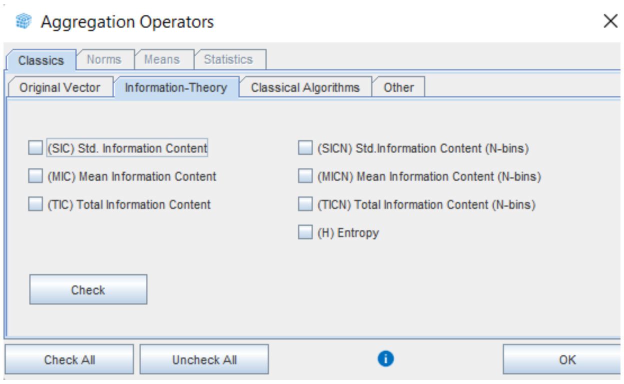
□ (V) Variance
□ (S) Skewness
□ (K) Kurtosis
□ (SD) Standard Deviation
□ (VC) Variation Coefficient
□ (RA) Range
□ (Q1) Percentile 25
□ (Q2) Percentile 50
□ (Q3) Percentile 75
□ (i50) Q3-Q1
□ (MX) XMax
□ (MN) XMin

D



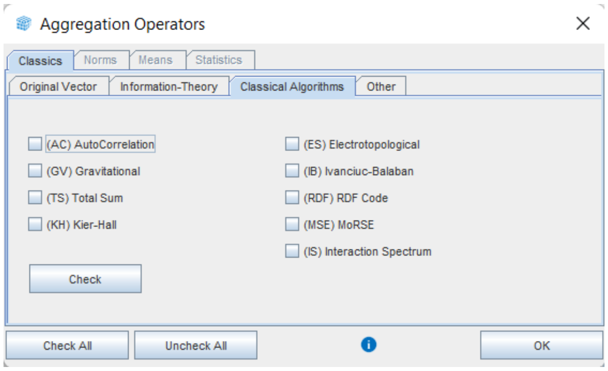
□ LAI Vector

E



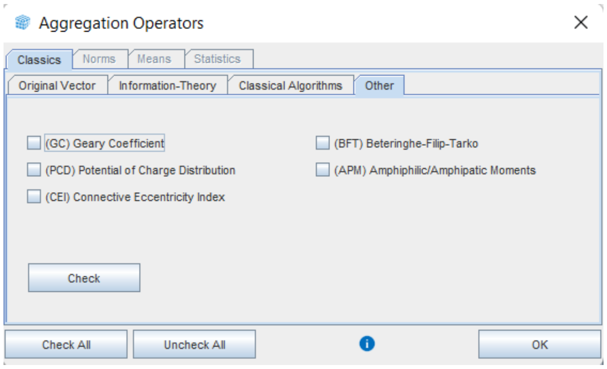
□ (SIC) Std. Information Content
□ (MIC) Mean Information Content
□ (TIC) Total Information Content
□ (SICN) Std.Information Content (N-bins)
□ (MICN) Mean Information Content (N-bins)
□ (TICN) Total Information Content (N-bins)
□ (H) Entropy

F



□ (AC) AutoCorrelation
□ (GV) Gravitational
□ (TS) Total Sum
□ (KH) Kier-Hall
□ (ES) Electrotopological
□ (IB) Ivanciuc-Balaban
□ (RDF) RDF Code
□ (MSE) MoRSE
□ (IS) Interaction Spectrum

G



□ (GC) Geary Coefficient
□ (PCD) Potential of Charge Distribution
□ (CEI) Connective Eccentricity Index
□ (BFT) Beteringhe-Filip-Tarko
□ (APM) Amphiphilic/Amphipatic Moments

Figure 25. Dialog windows to select the aggregation operator (s) to use to compute the MD-LAIs MDs. A) Norms. B) Means. C) Statistics. D,E,F and G) Classics.

Groups

In addition to ***total** indices* computed for the whole protein molecule, a **group-based** formalism can be developed. In this way, the macromolecular vectors are transformed to account for information related with certain subsets of amino acids. So, the group-based macromolecular vectors are used as basis to compute the *group-based indices*. The chemical groups employed in this software are:

- Apolar (RAP)
- Polar positively charged (RPC)
- Polar negatively charged (RNC)
- Polar uncharged (RPU)
- Aromatic (ARO)
- Aliphatic (ALG)

Also we defined groups that include the amino acids that do not favor the folding and/or cannot be commonly found in proteins as part of α-helices or β-sheets (UFG), α-helices favoring amino acids (FAH), β-sheets favoring amino acids (FBS) and β-turns favoring amino acids (AFT). Additionally, groups composed of amino acids of the same type (R-group) in the protein were defined, that is, 20 groups, one per each α-amino acid, (e.g. F=Ala, F=Arg,…, F=Val), and dipeptides (2-mers), that is, 400 groups (e.g. F=Ala-Ala, Ala-Arg,…, Val-Val), as well as tripeptides (3-mers), that is 8000 groups (e.g. F=Ala-Ala-Ala, Ala-Ala-Arg,…, Val-Val-Val).

Weights

- **UW** (Unweighted). All elements in vectors kept their original value.
- **LGST[TF(r_{on}-r_{off})]** (lag ST). Defines a weighting method based on fuzzy membership functions and functions traditionally used in molecular dynamics, is applied over the elements of the vectors. The parameters r_{on} and r_{off} define the lower and upper limits for the interval of the fuzzy set employed by the fuzzy membership functions (FMFs). The FMFs determine a weight value for each amino acid according to its distance either to the N-terminal (N), the center (middle) (M) or the C-Terminal (C) of a protein sequence. The weight for each amino acid represents the membership value of the amino acid to a zone of interest.

Properties

Thirty-nine **properties** (labels) are employed in **MD-LAIs software** as amino acid weights. These properties are grouped as follows:

- Chemical-physical (16 properties)

1. Mass (MM)
2. Volume (MV)
3. Z1-scale (Z1)
4. Z2-scale (Z2)

5. Z3-scale (Z3)
6. Atomic charge (ECI)
7. Isotropic surface area (ISA)
8. Hoop-Woods hydropathy index (HWS)
9. Kyte-Dolittle hydropathy index (KDS)
10. Isoelectric point (PIE)
11. Heat of formation (EPS)
12. Relative Alpha helix frequency (PAH)
13. Relative Beta-sheet frequency (PBS)
14. Relative Reverse turn frequency (PTT)
15. Geometric compatibility parameter 1(GCP1)
16. Geometric compatibility parameter 2(GCP2)

- HSE and Topological Scales (8 properties)

17. Vector of Hydrophobic, Steric and Electronic Properties (VHSE1)
18. Vector of Hydrophobic, Steric and Electronic Properties (VHSE4)
19. Vector of Hydrophobic, Steric and Electronic Properties (VHSE6)
20. Vector of Hydrophobic, Steric and Electronic Properties (VHSE7)
21. Vector of Hydrophobic, Steric and Electronic Properties (VHSE8)
22. Topological indices-based scale (T2)
23. Topological indices-based scale (T3)
24. Topological indices-based scale (T4)

- MD-LOVIS (6 properties)

25. MD-LOVIS PCA scale (MDL1)
26. MD-LOVIS PCA scale (MDL2)
27. MD-LOVIS PCA scale (MDL3)
28. MD-LOVIS PCA scale (MDL4)
29. MD-LOVIS PCA scale (MDL5)
30. MD-LOVIS PCA scale (MDL6)

- QuBILS-MIDAS (4 scales, 5 indices)

31. QuBILS-MIDAS scale (MID1)
32. QuBILS-MIDAS scale (MID2)
33. QuBILS-MIDAS scale (MID3)
34. QuBILS-MIDAS scale (MID4)
35. AC[3]_S_F_AB_nCi_2_M12_MP7_T_LGP[1]_c_MID (MID5)
36. K_B_AB_nCi_2_M16_SS1_X_LGL[1-2]_c-p_MID (MID6)
37. GV[1]_S_Q_AB_nCi_2_M13_NS3_o_C_KA_r_MID (MID7)
38. ES_S_Q_AB_nCi_2_M15_SS4_T_LGL[2-3]_c_MID (MID8)
39. S_F_AB_nCi_2_M8_NS2_T_LGP[1]_c_MID (MID9)

Aggregation Operators

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

LAIs of a protein are usually collected into an *n*-dimensional vector (*n* = number of amino acids). LAIs are used to calculate several molecular indices by applying different aggregation operators. *L* is adopted here as the general symbol for local invariants.

Over the years, it has been generally accepted that the definition of global (or local) indices from LAIs (L_i) implies the summation of the contributions of the elements that constitute a given protein. However, summation (Minkowski's first norm (N1) in our specific case) is just one of the many invariants capable of globally characterizing a given LAIs. In this software, a series of *aggregation operators* (AOs) that generalize the traditional method of obtaining global (or local) indices by summation of the LAIs are employed. These AOs are classified in four major groups:

1) **Norms (or Metrics)**:
   - **a)** Minkowski's norms (N1, N2, N3)
2) **Mean Invariants (first statistical moment)**:
   - **a)** Geometric Mean (GM),
   - **b)** Arithmetic Mean (AM),
   - **c)** Quadratic Mean (P2),
   - **d)** Potential Mean (P3) and
   - **e)** Harmonic Mean (HM)
3) **Statistical Invariants (highest statistical moments):**
   - **a)** Variance (V),
   - **b)** Skewness (S),
   - **c)** Kurtosis (K),
   - **d)** Standard Deviation (SD),

e) Variation Coefficient (VC),
f) Range (R),
g) Percentile 25 (Q1),
h) Percentile 50 (Q2),
i) Percentile 75 (Q3),
j) Inter-quartile Range (I50),
k) Xmax (MX) and
l) Xmin (MN)

4) "Classical algorithms" Invariants (functions to derive MDs from LAIs):

a) Autocorrelations AC(i),
b) Gravitational (GV(i)),
c) Total sum at *k* lags (TSk(i)),
d) Mean information content (MIC(i)),
e) Total information content (TIC),
f) Standardized information content (SIC),
g) Mean information content *N*-binned (MICN(i)),
h) Total information content *N*-binned (TICN),
i) Standardized information content *N*-binned (SICN),
j) Entropy (H),
k) Electro-topological state (ES(i)),
l) Kier-Hall,
m) Ivanciuc-Balaban,
n) Geary Coefficient (GC(i)),
o) Potential Charge Distribution (PCD(i)),
p) Connective Eccentricity Index (CEI),
q) Radial Distribution Function (RDF),
r) MoRSE (MSE),
s) Interaction spectrum (IS),
t) Beteringhe-Filip-Tarko (BTF), and
u) Amphiphilic/amphipatic moments (APM)

1. Norms (or Metrics): Mathematical definition

Name	ID	Formula
Minkowski norm (p = 1) Manhattan norm	N1	$N1 = \sum_{a=1}^{n} \lvert L_a \rvert$
Minkowski norm (p = 2) Euclidean norm	N2	$N2 = \sqrt{\sum_{a=1}^{n} \lvert L_a \rvert^2}$
Minkowski norm (p = 3)	N3	$N3 = \sqrt[3]{\sum_{a=1}^{n} \lvert L_a \rvert^3}$

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

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

But given that $\overline{y}$ = (0, 0, 0), this formula reduces to $\left\| \overline{x} \right\|_2 = \sqrt{\sum_{i=1}^{n} \lvert x_i \rvert^2}$.

2. Mean Invariants: Mathematical definition

Name	ID	Formula
Geometric Mean	GM	$G = \sqrt[n]{\prod_{a=1}^{n} L_a}$
Arithmetic Mean (Power mean of degree β = 1)	AM	$M_\beta = \left(\frac{L_1^\beta + L_2^\beta + \ldots + L_n^\beta}{n} \right)^{\frac{1}{\beta}}$
Quadratic Mean (Power mean of degree β = 2)	P2	
Power mean of degree β = 3	P3	
Harmonic Mean (Power mean of degree β = -1)	HM	

3. Statistical Invariants: Mathematical definition

Name	ID	Formula
Variance	V	$V = \frac{\sum_{a=1}^{n} \left(L_a - M \right)^2}{n - 1}$
Skewness	S	$S = \frac{n * (X_3)}{(n-1)(n-2)(DE)^3}$ $X_3 = \sum_{a=1}^{n} (L_a - M)^3$ M, arithmetic mean DE, standard deviation

Kurtosis	K	$k = \frac{n(n+1)X_4 - 3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(DE)^4}$ M, $X_j = \sum_{a=1}^{n} (L_a - M)^j$ arithmetic mean SD, standard deviation
Standard Deviation	SD	$SD = \sqrt{\frac{\left(\sum L_a - M\right)^2}{n-1}}$
Variation Coefficient	VC	CV = SD/M
Range	R	R = $L_{\max} - L_{\min}$
Percentile 25	Q1	$\text{Q1} = \left[\frac{N}{4} + \frac{1}{2}\right]$ N, L_a number
Percentile 50	Q2	$\text{Q2} = \left[\frac{N}{2} + \frac{1}{2}\right]$ N, L_a number
Percentile 75	Q3	$\text{Q3} = \left[\frac{3N}{4} + \frac{1}{2}\right]$ N, L_a number
Inter-quartile Range	I50	I50 = $Q3 - Q1$
Maximum value	MX	MX = L_a max
Minimum value	MN	MN = L_a min

4. Classical Invariants: Mathematical definition

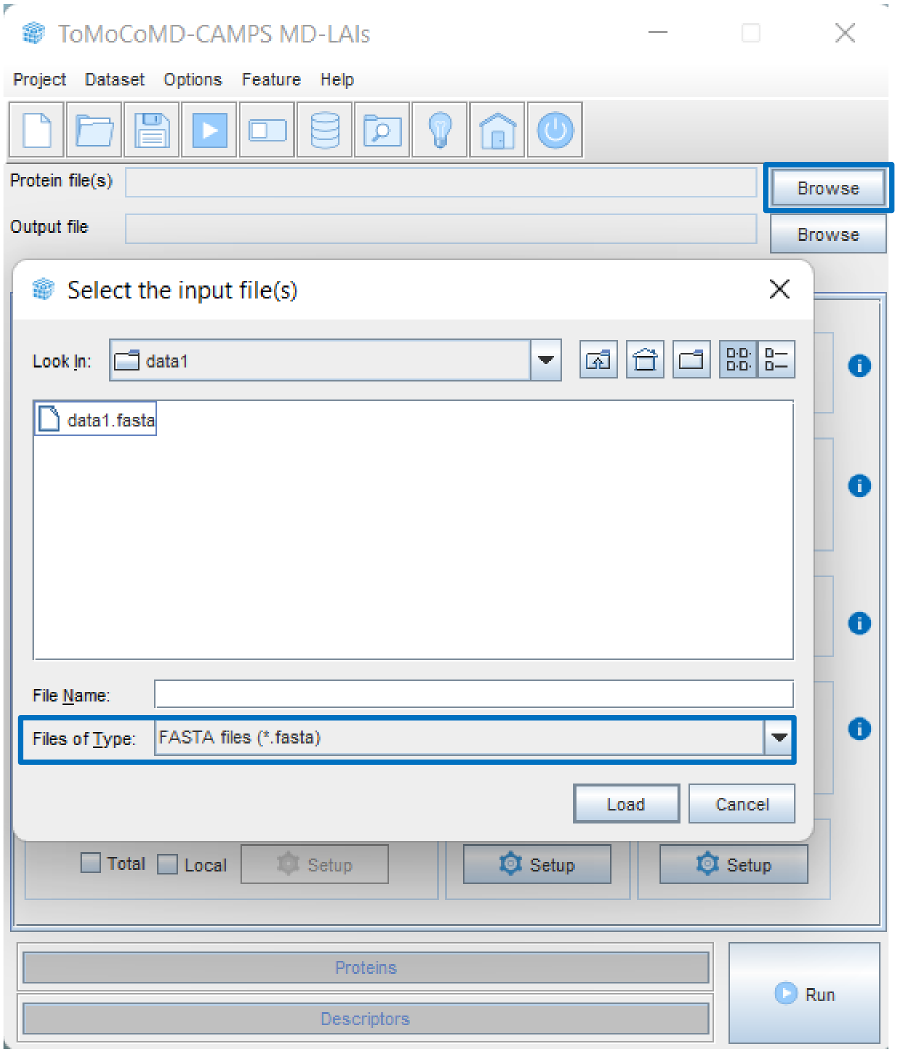
Name	ID	Formula
Autocorrelation	AC^k	$AC_k = \sum_{i=1}^{n}\sum_{j\geq 1}^{n} L_i \times L_j \bullet (\delta(d_{ij}, k))$ $k = 1,2,..7$ *where, d_{ij} is the topological distance between amino acid i and j and k is the cutoff distance.*
Gravitational	GV^k	$GV_k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{L_i L_j}{{}^k d_{ij}} \bullet \delta(d_{ij}, k))$ $k = 1,2,..7$
Total sum (lag k)	TS^k	$TS_k = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij} \bullet \delta(d_{ij}, k))$ $k = 1,2,\ldots,7$
Kier-Hall Connectivity	KH_k	$CN_k = \sum_{k=1}^{K}\left(\prod_{i=1}^{nk} L_i\right)_k^{\lambda}$ *where, K is the number of path sub-graphs, nk is the number of amino acids in a fragment, λ is equal to ½, and*

		k is the order of path subgraph (number of amino acids in the path considered). Here, k =1, 1,…,7
Mean Information Content	MIC	$MIC = -\sum_{i=1}^{A} \frac{N_g}{N_o} \cdot \log_2 \frac{N_g}{N_o}$ *where, Ng is the number of amino acids with the same LAI value. N_o is the number of amino acids in a protein.*
Total Information Content	TIC	$TIC = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g$
Standardized Information Content	SIC	$SIC = \frac{IT}{N_0 \cdot \log_2 N_0}$
Mean Information Content (N-binned)	MICN	$MIC = -\sum_{i=1}^{A} \frac{N_g}{N_o} \cdot \log_2 \frac{N_g}{N_o}$ *where, Ng is the number of amino acids that fall into different intervals (bins), N_o is the number of amino acids in a protein*
Total Information Content (N-binned)	TICN	$TIC = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g$
Standardized Information Content (N-binned)	SICN	$SIC = \frac{IT}{N_0 \cdot \log_2 N_0}$
Entropy	H	$H = \sum_{i=1}^{n} -p_i log p_i, p_i = \frac{L_i{'}}{\sum_{i=1}^{n} L_i{'}}, L_i{'} = \frac{L_i - \bar{L}}{s_L}$ *where, $\bar{L}$ and s_L are the arithmetic mean and standard deviation of LAIs vector (L), respectively.*
Ivanciuc-Balaban	IB	$IB = \frac{n^2 \cdot B}{n + C + 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \left[L_i \times L_j\right]^{\frac{-1}{2}}$ *where, the summation goes over all pairs of amino acids, but only pairs of adjacent amino acids are accounted for by means of the elements a_{ij} of the adjacency matrix. The n, B, and C are the number of amino acids, amide bonds, and rings (cyclomatic number), respectively.*
Geary Coefficient	GC^k	$GC_k = \frac{\frac{1}{2\Delta k} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial(d_{ij}, k) \left[L_i - L_j\right]^2}{\frac{1}{n-1} \sum_{i=1}^{n} (L_i - \bar{L})}$ *where L_i is the LAI value for amino acid i, $\bar{L}$ is its average value on the protein, n is the number of amino acids, k is the lag considered, d_{ij} is the topological distance between amino acid i and j, and $\partial(d_{ij}, k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. Δk is the number of vertex pairs at distance equal to k.*
Potential of Charge Distribution	PCD	$PCD = \sum_{i=1}^{n} \frac{L_i}{d_i}$ *where, L_i is the LAI value for amino acid i, d_i is the distance from each amino acid to the center (half of sequence) of sequence.*
Connective Eccentricity index	CEI	$CEI = \sum_{i=1}^{n} \frac{L_i}{\partial_i}$

		where, $\partial_i$is the topological eccentricity of amino acid i, that is, the largest topological distance from amino acid i to n-1 amino acids.
RDF Code	RDF (R)	$\mathrm{RDF}(R) = f \times \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} L_i \times L_j \times e^{-\beta\cdot(R-d_{ij})^2}$ *where, f is a scaling factor (here 1/n), L_i characteristic amino acidic LAIs of the amino acids i and j, d_{ij} the topological distance between amino acid i and j, and n the number of amino acids. β is a smoothing parameter (here 100 that defines the probability distribution of the individual interatomic distances; β can be interpreted as a temperature factor that defines the movement of amino acids. Here, seven radiuses are employed R=(n/8,n/7,n/6,n/5,n/4,n/3 and n/2).*
Beteringhe–Filip–Tarko	BFT	$BFT = \frac{B}{n+B} \times \sum_{i=1}^{n} \log{(L_i)}^2$ *where, n is the number of amino acids and B the number of amide bonds.*
MoRSE	MSE	$MSE = \sum_{i=2}^{n} \sum_{j=1}^{i-1} L_i L_j \frac{\sin r_{ij}}{s d_{ij}}$ *where, s is the scattering parameter, d_{ij} is the topological distance between amino acid i and j.*
Inter-amino acid Interaction Spectrum	IS(R)	$IS(R) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{L_i L_j}{1 + \sqrt{\frac{R - d_{ij}}{0.1}}}$ *d_{ij} is the topological distance between amino acid i and j. R is the radius. Here, seven radiuses are employed R=(n/8,n/7,n/6,n/5,n/4,n/3 and n/2).*
Amphiplilic/Amphiphatic moments	APM	$APM = \sum_{i=1}^{n} d_i \, L_i$ *where, $d_i$is the topological distance of amino acid i to the farthest hydrophobic/hydrophilic/amphipatic amino acid.*

Input and Output Files

The following is a description of the **Input** and **Output** section of the **MD-LAIs** GUI. In these sections, input sequences can be loaded and the output descriptor files can be chosen. The FASTA input files are selected and loaded by clicking the *Browse* button in the **Input Protein (s)** 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 by MD-LAIs, so you can easy locate your files.



Look In: data1

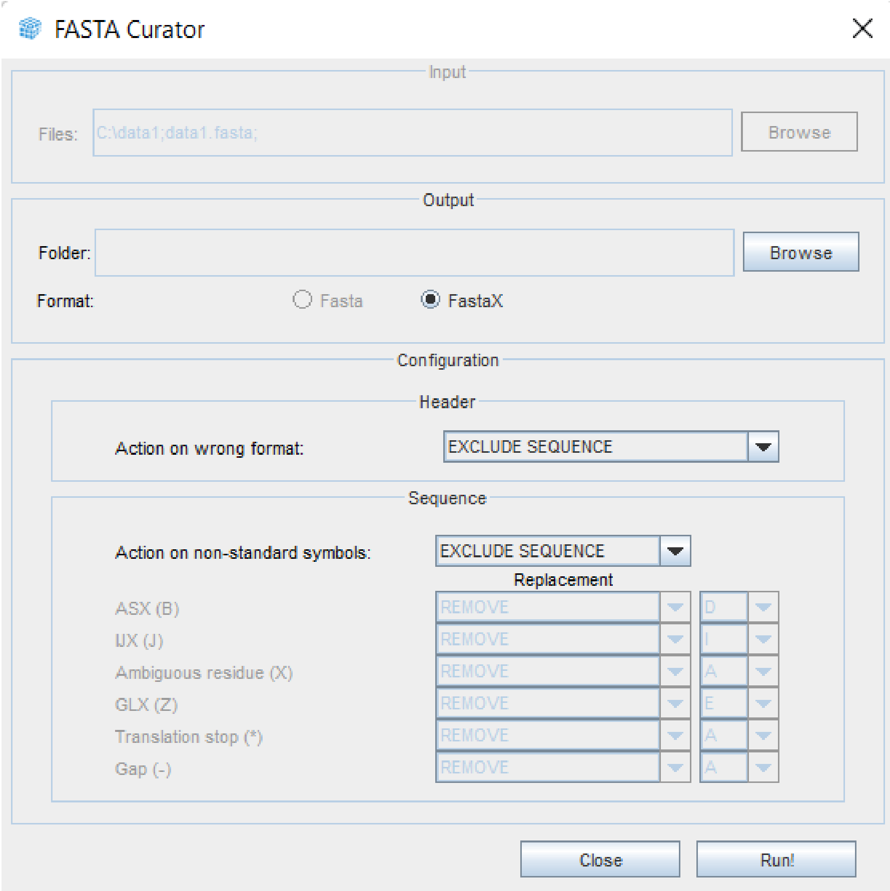
File Name:

Files of Type: FASTA files (*.fasta)

Figure 26. Browsing Input files (*. fasta).

When you are browsing for the input files there are two possibilities:
1) The selected file(s) has (have) the extension (s) (*.fasta).
2) The selected file(s) has (have) the extension (s) (*.fastax).

In the first case, a window for data curation is prompted. This window contains options for fixing the headers and sequences of the input files. The core objective of this module is to fix (or discard) the FASTA entries that contain format errors or non-standard amino acids and to output standardized error-free files (denoted here as an internal format *.fastax), which in turn can be employed in subsequent calculations.



Files: C:\data1;data1.fasta;

Folder:

Format: ○ Fasta ● FastaX

Action on wrong format: EXCLUDE SEQUENCE

Action on non-standard symbols: EXCLUDE SEQUENCE

	Replacement	
ASX (B)	REMOVE	D
IJX (J)	REMOVE	I
Ambiguous residue (X)	REMOVE	A
GLX (Z)	REMOVE	E
Translation stop (*)	REMOVE	A
Gap (-)	REMOVE	A

Figure 27. FASTA Curator Window

In the second case (*.fastax files), no additional action is required since it is assumed that these files are already standardized and free of errors.

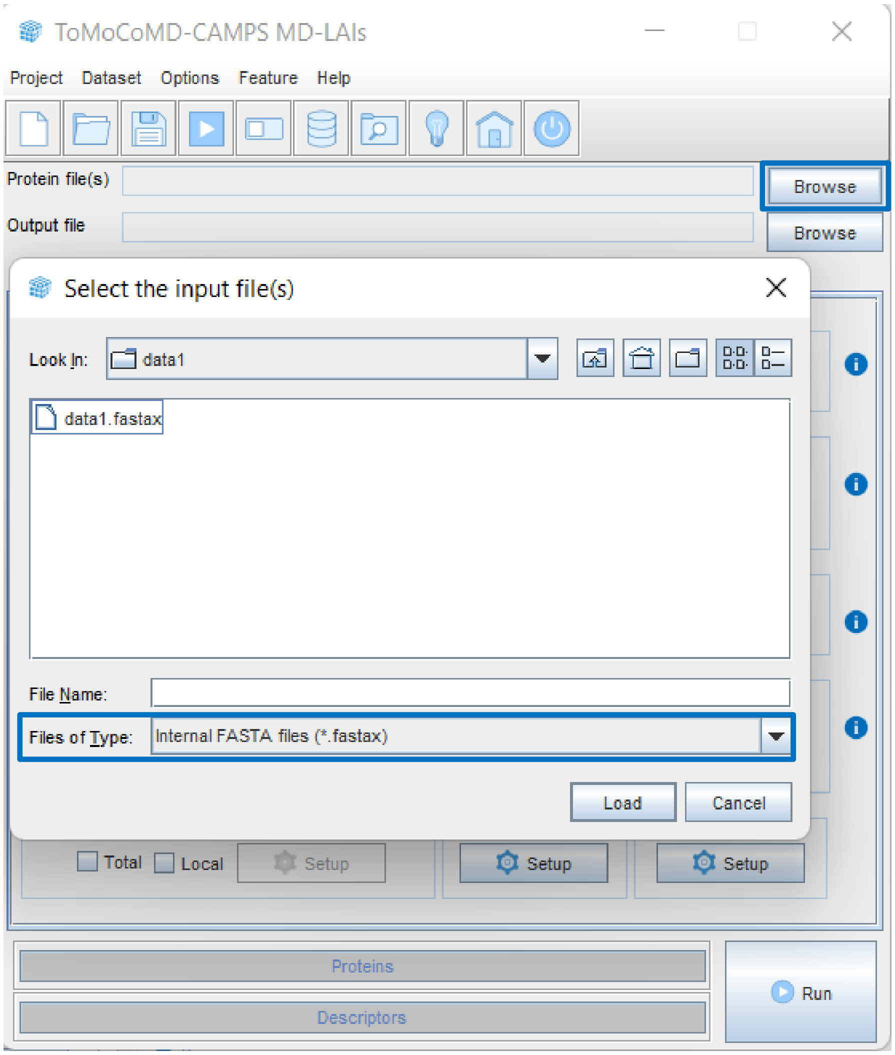
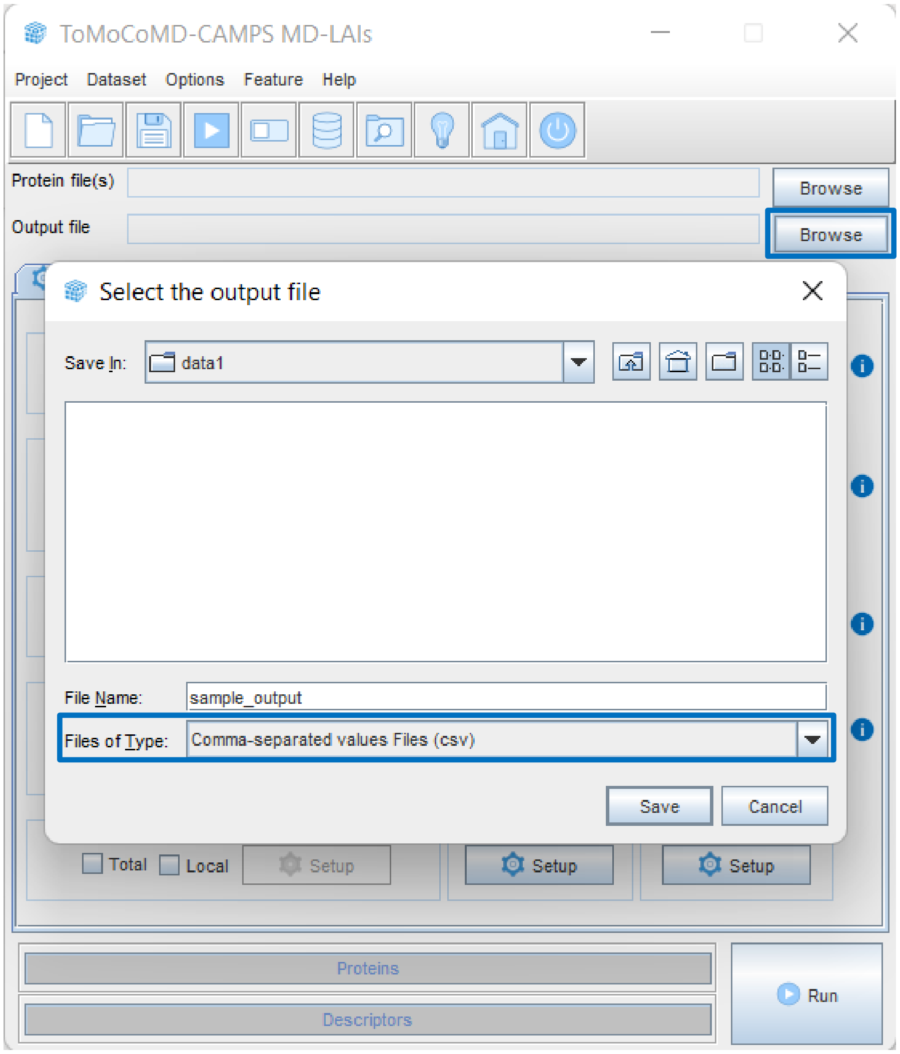


Figure 28. Browsing Input standardized files (*.fastax)

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. MD-LAIs supports, CSV format (Comma Separated Value), TSV format (Tab-separated Values File) and ARFF (Attribute-Relation File Format) Weka file.



Save In: data1

File Name: sample_output

Files of Type: Comma-separated values Files (csv)

Figure 29. Browsing Output file.

Supported File Formats

INPUT

FASTA

FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

OUTPUT

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 MD-LAIs

MD-LAIs produces an output file containing the values of the calculated and selected MDs, together with the additional information imported by the user. The Output File can be selected by clicking on the *Browse* button in the *Output* section and MD-LAIs 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 calculating the descriptors.
2. Unexpected Error in calculating a descriptor.

The **standard MD-LAIs 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 the proteins, *that is*: the “name of each file” plus “.fastax” plus “_” plus the number of protein according to the order in which it appears in the FASTA file.
- The following records (columns) contain the **variable labels** *(descriptor headings)*, *i.e.* **N1_ T_UW_Z1** (see Descriptor Search Tool in order to automatic decodify each header).

Table 1. An example Protein-level Output file.

proteins	N1_T_UW_Z1	K_T_UW_Z1
data1.fastax_1	39.71	-1.261523006
data1.fastax_2	106.42	-0.757320979
data1.fastax_3	7.76	-1.169913512
data1.fastax_4	-8.84	-1.421946488
data1.fastax_5	206.15	-0.779043813

Table 6. An example Amino acid-level Output file.

aminoacids	N1[0]_T_UW_Z1	N1[5]_T_UW_Z1
data1.fastax_1_MET_1	-2.49	-8.04
data1.fastax_1_ASN_2	3.22	-12.23
data1.fastax_1_ILE_3	-4.44	-9.35
data1.fastax_1_PHE_4	-4.92	-13.79
data1.fastax_1_GLU_5	3.08	-10.15

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.

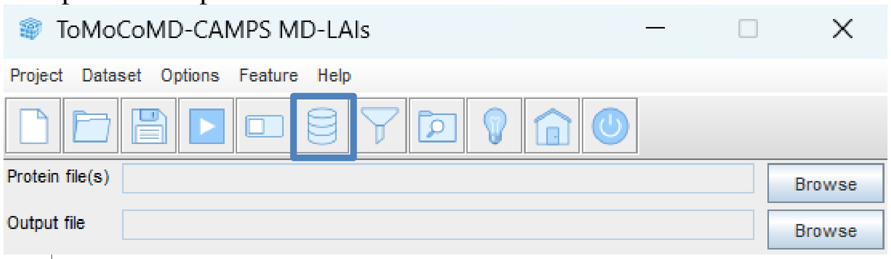
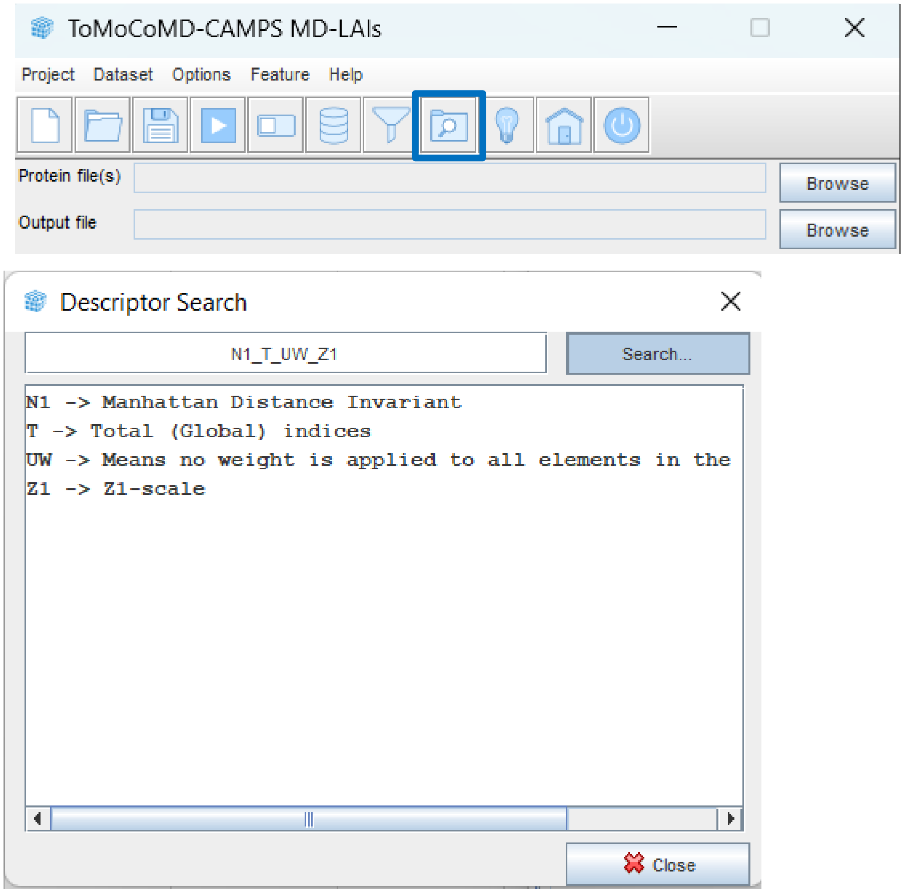


Figure 30. Quick access shortcuts for example data tool.

Searching for Descriptors Headers

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



N1_T_UW_Z1

```
N1 -> Manhattan Distance Invariant
T -> Total (Global) indices
UW -> Means no weight is applied to all elements in the
Z1 -> Z1-scale
```

Figure 31. Descriptor Search Tool windows.

Debug Report Capability

This option permits, for each protein in the dataset, the user to save a txt file with the property and LAIs vectors.

Property Vector
LAIs Vector

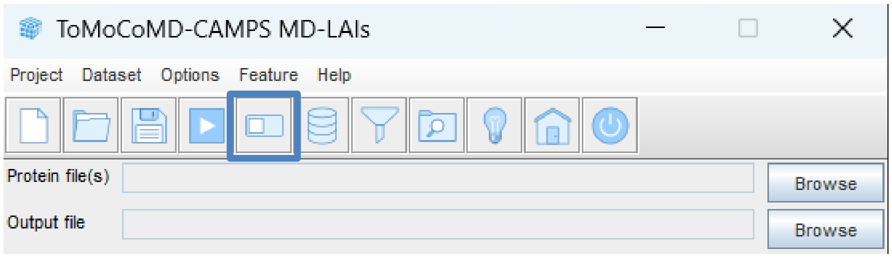


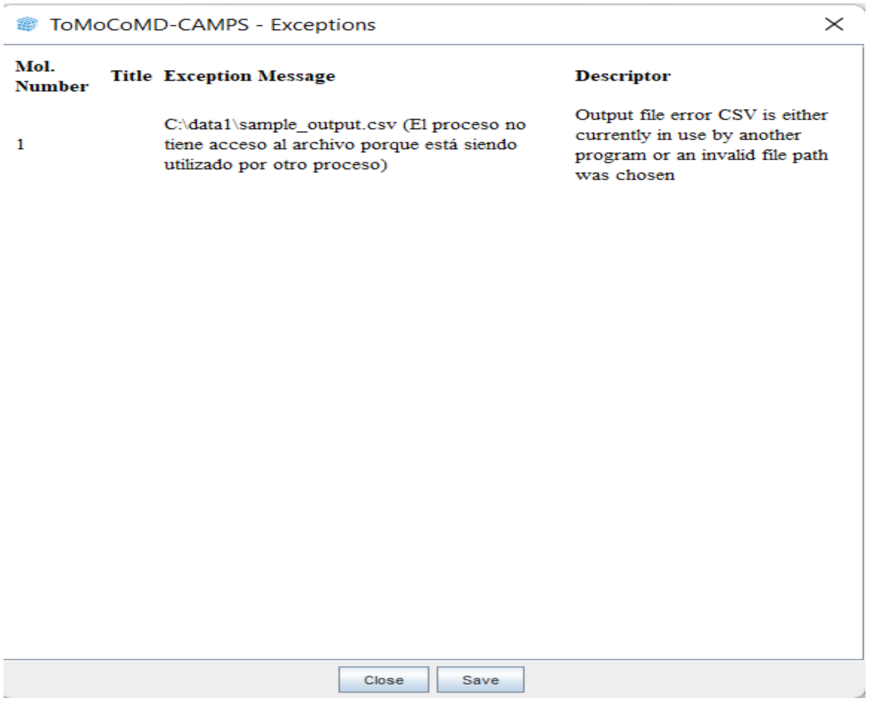
Figure 32. On/Off Generate Debug Report button.

Special Instructions and Exceptions

During the descriptor calculation **MD-LAIs** writes out a log file that shows some statistics on the program run (Log 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 ‘Logs 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*). Two errors or exceptions types are possible:

1. Errors calculating the MD-LAIs descriptors. The missing values label *NaN* is placed as descriptor value for each invalid entry.
2. 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 proteins with problems due to error in calculations is shown in the ‘Exception file’ window together with the error type. The error and warning messages given below are printed in a log file that can be saved by clicking the **save** button.



Mol. Number	Title	Exception Message	Descriptor
1		C:\data1\sample_output.csv (El proceso no tiene acceso al archivo porque está siendo utilizado por otro proceso)	Output file error CSV is either currently in use by another program or an invalid file path was chosen

Figure 33. Exception window.