starPep toolbox v0.8

User Guide

By Developer Team

Contents

Chapter 1

General Information

1.1 System Overview

StarPep toolbox has been designed in a user-friendly way for supporting visual network analysis of bioactive peptides of pharmaceutical interest. The bioactive peptides are collected from a large variety of biological data sources to be organized into an integrated graph database called starPepDB, where peptide nodes are at the center surrounded by their neighbor nodes representing metadata (resembling an asterisk or star schema). This integrated graph database is embedded in starPep toolbox to enable end-user querying, filtering, visualizing, and analyzing the bioactive peptides taking advantage of network-based representations.

1.1.1 License

1.2 System requirements

1.2.1 Hardware

- Memory (RAM): A minimum of 4 GB is required, but we recommend 8GB or more.
- Processors: We recommend a multi-core processor due to the fact that the software has been implemented to enable parallel processing of computationally intensive tasks.
- Hard Disk: a minimum of 500 MB of free space is required.

1.2.2 Software

- Java SE Runtime Environment 8.

Note: It does not work (yet) with versions of Java greater than 8.

1.3 Downloading and installing the program

The binary executable files for Windows, Mac, and Linux are located in http://mobiosd-hub.com/starpep/. You can download the zip distribution and extract it to a folder or use an installer for the application.

1.4 Issue with java versions

StarPep toolbox **does not yet support** any version of Java $>$ 8. **The requirement is java 8**. If you have multiple Java versions installed on your system, please configure starPep toolbox to run on the supported one (Java 8). Find the etc/starPep.conf file in the installed directory and configure the jdkhome="/path/to/jdk" accordingly. The symbol "#" at the beginint of the line means that it is commented out, please remove it.

1.5 Increasing the memory heap size

You may increase the memory heap further if there is enough RAM available in your system (**recommended**). First, you have to switch to the directory where the application has been installed or extracted. Open the text file "starpep.conf" located under the *etc* folder. Once the file has been open, locate the default_options setting and change the min/max heapsize values (-J-Xms or -J-Xmx). For instance, to increase the memory heap size from 4G to 8G, enter the value:

Before (4G):

```
default_options="--branding starpep -J-Xms24m -J-Xmx4G"
```

After (8G):

```
default_options="--branding starpep -J-Xms24m -J-Xmx8G"
```

Then save the text file *etc/starpep.conf* and run the application.

1.6 Running starPep toolbox

StarPep toolbox can be initiated by running the *bin* executable files located in the installed directory, or by clicking the application icon (if installed).

Figure 1.1: Loading screen of StartPep toolbox

Chapter 2

Getting Started

2.1 Main view

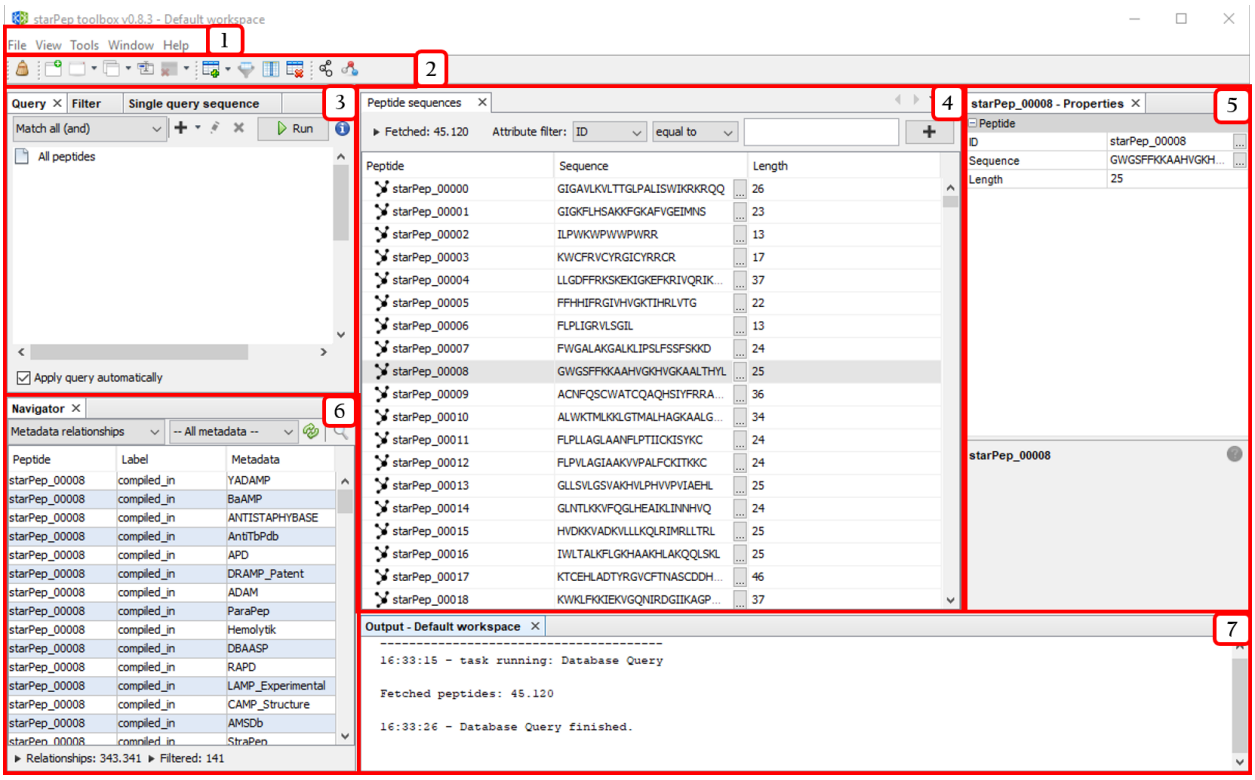


Figure 2.1: StartPep toolbox main window

1. Menu bar
2. Quick access bar
3. Tools panel
4. Central panel
5. Properties panel

6. Navigator panel
7. Output panel

Note: The above windows panels may be opened from the Window option in the menu bar.

2.2 Menu bar

2.2.1 File

The following options are accessible from the File option:

Note: Workspaces may be used to work with different data models: one per workspace.

File > New workspace Creates a new workspace
File > Select workspace > [workspace_name] Switches to a new workspace
File > Copy data to > New workspace Duplicates data model to a new workspace
File > Rename current workspace Renames the current workspace
File > Remove workspace
- File > Remove workspace > Remove current workspace Removes the current workspace
- File > Remove workspace > Remove other workspaces Removes the other workspaces, and only remains the current workspace

File > Clean project Removes all workspaces and sets the default workspace with the default data model
File > Import > Peptide sequences (FASTA format) Imports peptide sequences into a new workspace.
File > Export Exports the following data
- File > Export > Peptide sequences (**FASTA format**)
- File > Export > Molecular descriptors (**CSV format**)
- File > Export > Networks (**GraphML format**)
- File > Export > Metadata relationships (**CSV format**)

File > Exit Shutdowns the program

2.2.2 View

The following options are accessible from the View option:

View > Toolbars . Shows/hides a quick access bar.
- View > Toolbars > File
- View > Toolbars > Workspace
- View > Toolbars > Network
- View > Toolbars > Molecular Descriptors

View > Full Screen Switches to full screen.

2.2.3 Tools

The following options are accessible from the Tools option:

- Tools > Peptide querying Opens/selects the query tab in the Tools panel.
- Tools > Peptide search by
  - Tools > Peptide search by > Single Query sequence Opens/selects the single query tab in the Tools panel.
  - Tools > Peptide search by > Multiple Query sequences Opens/selects the multiple query tab in the Tools panel.
  - Tools > Peptide search by > Non-redundant set Opens/selects the non-redundant set tab in the Tools panel.
- Tools > Peptide filtering Opens/selects the filter tab in the Tools panel.
- Tools > Molecular features
  - Tools > Molecular features > Extraction > [molecular descriptor option] Opens/selects the molecular descriptor tab in the Tools panel.
  - Tools > Molecular features > Selection > [unsupervised feature selection] Opens/selects the unsupervised feature selection tab in the Tools panel.
  - Tools > Molecular features > Explorer Opens the feature explorer window.
  - Tools > Molecular features > Removing Opens the feature removing window.
- Tools > Network
  - Tools > Network > Metadata Metadata Opens the window to generate a metadata network.
  - Tools > Network > Similarity Network Opens/selects the chemical space tab in the Tools panel.
  - Tools > Network > Appearance Opens/selects the appearance tab in the Tools panel.
  - Tools > Network > Layout > [layout algorithm] Opens/selects the layout algorithm tab in the Tools panel.
  - Tools > Network > Clustering > [clustering algorithm] Opens/selects the clustering algorithm tab in the Tools panel.
  - Tools > Network > Centrality > [measure] Opens/selects the centrality measure tab in the Tools panel.
  - Tools > Network > Subnetwork mining > [graph-based algorithm] Opens/selects the graph-based algorithm tab in the Tools panel.
- Tools > Options Displays the software configuration window.

2.2.4 Window

The following options are accessible from the Window option:

- Window > Peptide sequences Opens/selects the peptide sequences window in the center panel.
- Window > Network visualization Opens/selects the network visualization window in the center panel.
- Window > Properties Opens/selects the properties panel.
- Window > Navigator Opens/selects the navigator panel.

Window > Output Opens/selects the output panel.
Window > Configure Window > [options] Window settings
Window > Reset Windows
Window > Close Window
Window > Close All Documents
Window > Close Other Documents
Window > Documents... Opens the Document management window.

2.2.5 Help

The Help > About Opens the About information window.

2.3 Quick access bar

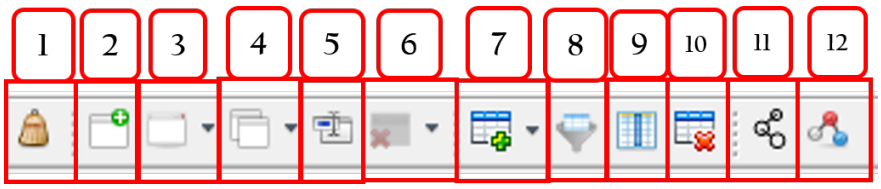


Figure 2.2: Quick access bar.

Note: These options may be shown/hidden from the menu entry: View > Toolbars > [option]

1. Shortcut to File > Clean project.
2. Shortcut to File > New workspace.
3. Shortcut to File > Select workspace.
4. Shortcut to File > Copy data to > New workspace.
5. Shortcut to File > Rename current workspace.
6. Shortcut to File > Remove workspace.
7. Shortcut to Tools > Molecular features > Extraction.
8. Shortcut to Tools > Molecular features > Selection.
9. Shortcut to Tools > Molecular features > Explorer.
10. Shortcut to Tools > Molecular features > Removing.
11. Shortcut to Tools > Network > Metadata Network.
12. Shortcut to Tools > Network > Similarity Network.

2.4 Tool panels: an overview

2.4.1 Query panel

This panel may be opened from Tools > Peptide querying.

Note: The recovered peptides are those linked to the specified metadata nodes.

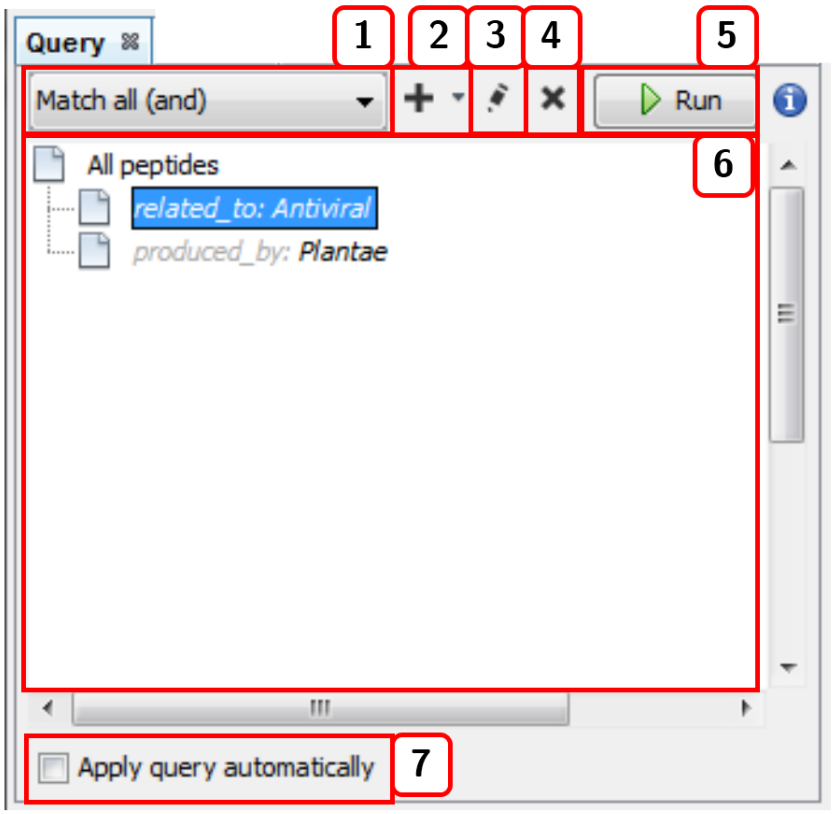


Figure 2.3: Query panel

1. Selects the joining condition for the query criteria: Match all (and) or Match all (or).
2. Adds a new term (linked metadata) to the query.
3. Edits the query term selected.
4. Deletes the query term selected.
5. Runs the query.
6. List of current query terms.
7. Applies the query automatically with each change.

2.4.2 Filter panel

This panel may be opened from Tools > Peptide filtering.

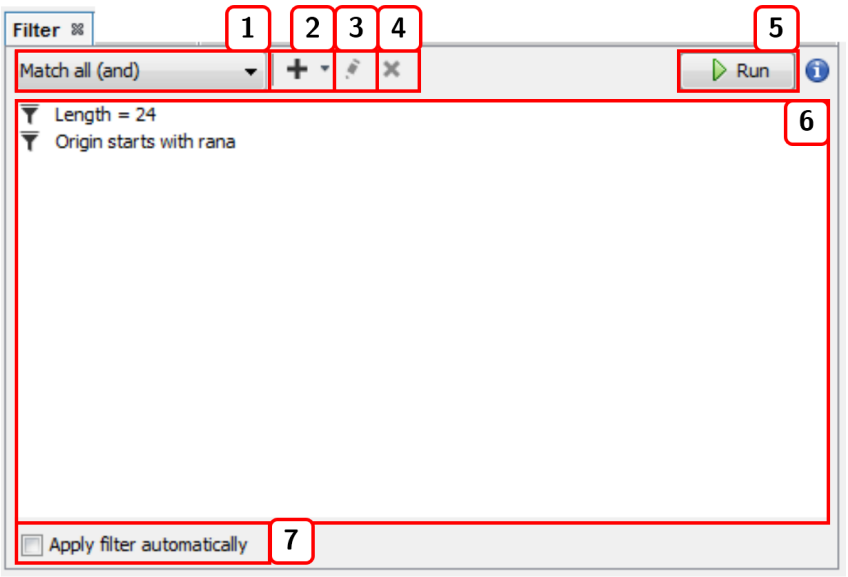


Figure 2.4: Filter panel

1. Selects the joining condition for the filter criteria: Match all (and) or Match all (or).
2. Adds a new filter.
3. Edits the selected filter.
4. Deletes the selected filter.
5. Runs the filter.
6. List of current filters.
7. Applies the filter automatically with each change.

2.4.3 Sequence search

This panel can be opened from Tools > Peptide search by > [sequence search option]. For instance, Single query sequence:

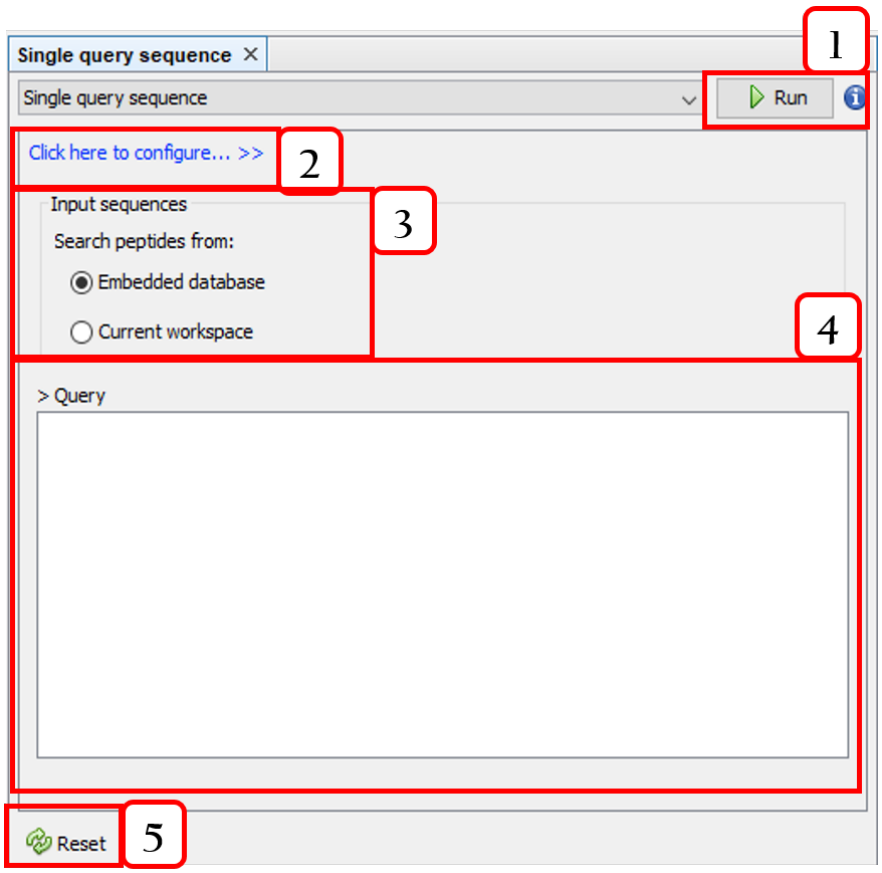


Figure 2.5: Single query sequence.

1. Runs the query.
2. Configures the sequence alignment (Fig. 2.6).
3. Selects the target sequences.
4. Input sequence.
5. Resets the query.

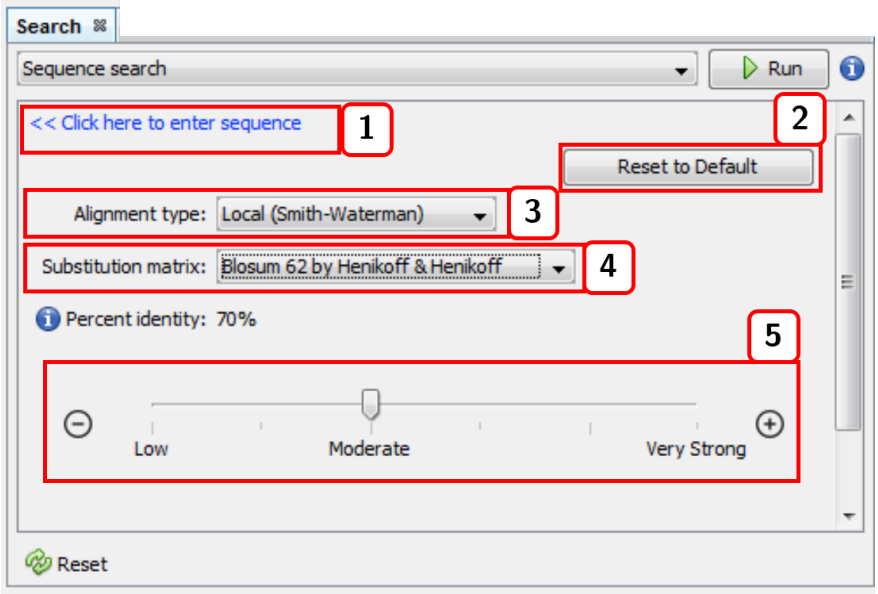


Figure 2.6: Sequence alignment settings.

1. Returns to the input sequence view.
2. Resets the alignment configuration.
3. Alignment type (local or global).
4. Substitution matrix.
5. Percent identity (default: 98%).

2.4.4 Molecular feature extraction

This option is accessible from the menu option: Tools > Molecular features > Extraction > [molecular descriptor option]. For instance, All descriptors:

Note: The calculated molecular descriptors can be removed by accessing the menu options Tools > Molecular features > Removing.

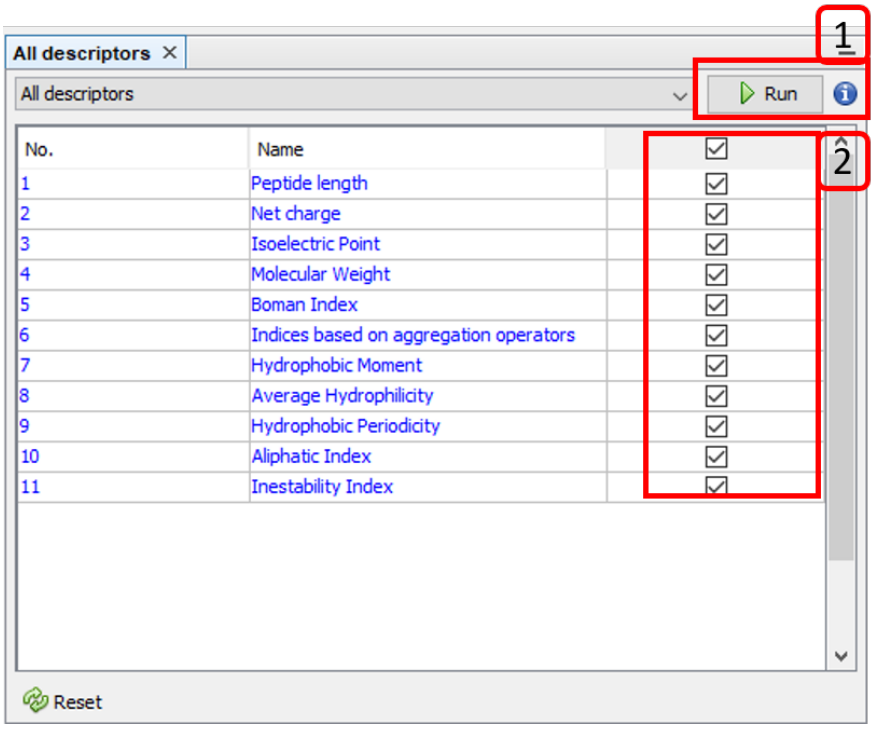


Figure 2.7: Calculating all molecular descriptors.

1. Runs the selected molecular descriptor algorithms.
2. Selects/Unselect molecular descriptor algorithms.

Besides, calculated molecular features can be displayed in the columns list at the center panel (enabling molecular feature filtering). This option is accessible from the menu option: Tools > Molecular features > Explorer:

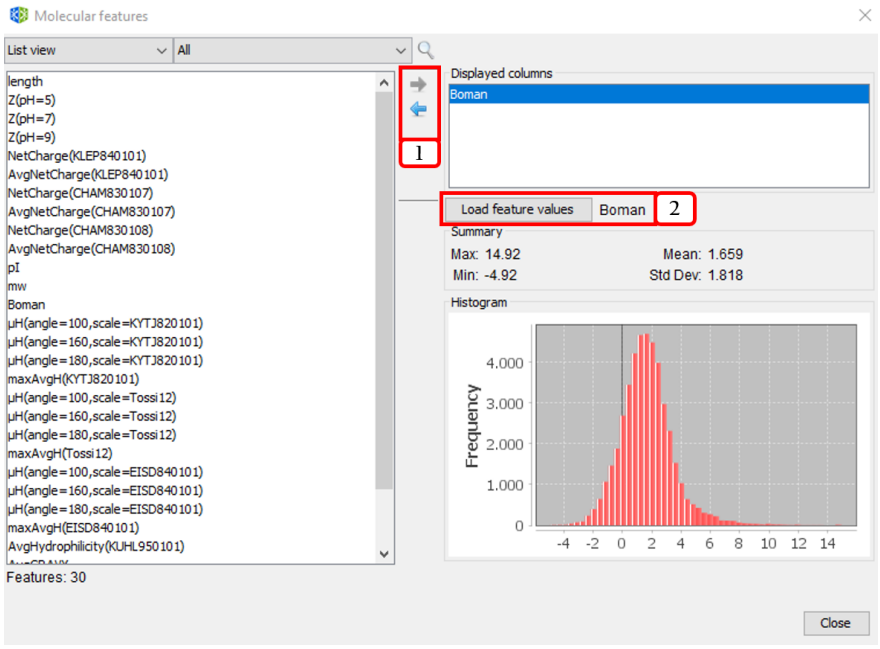


Figure 2.8: Adding molecular features (*Boman*) to the displayed columns list.

1. Adds/Removes molecular descriptors to/from the displayed columns list.
2. Visualizes the histogram and data summary (max, min, mean, and standard deviation) of molecular feature values.

2.4.5 Molecular feature selection

This option is accessible from the menu option: Tools » Molecular features » Selection » [unsupervised feature selection]. For instance, Filtering & subset optimization:

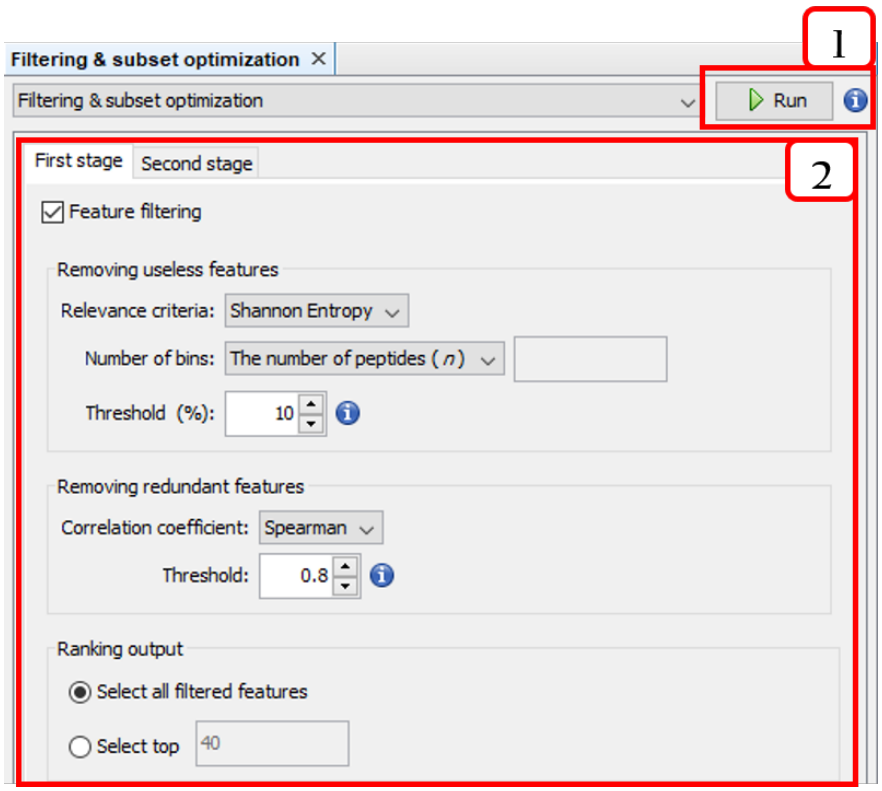


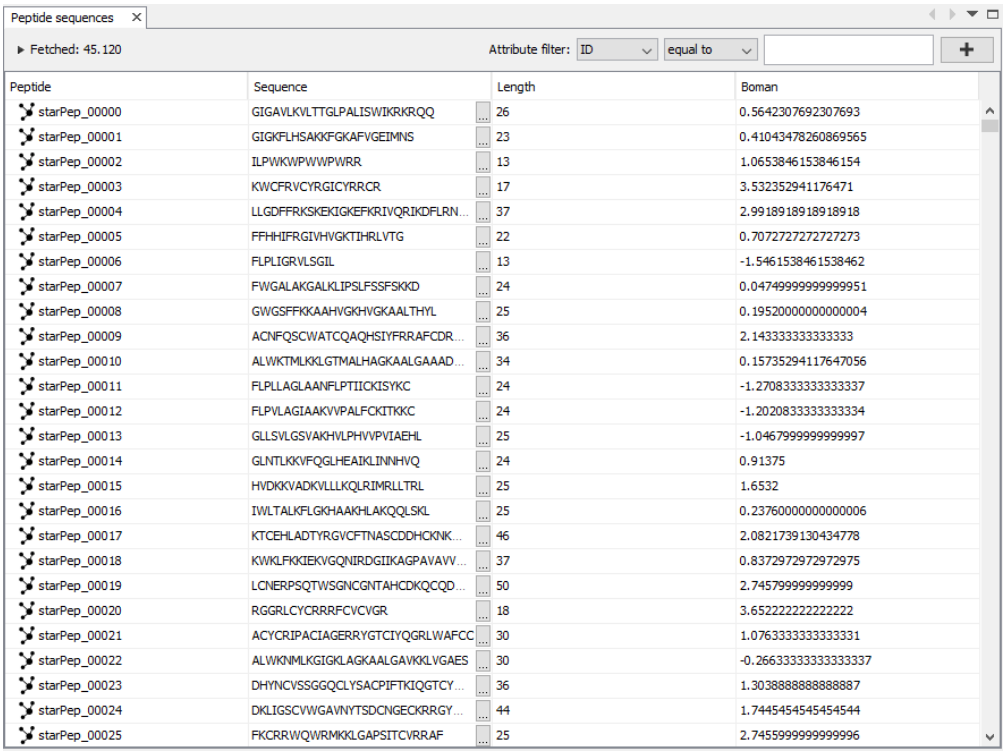
Figure 2.9: The two-stage unsupervised feature selection

1. Runs the two-stage unsupervised feature selection.
2. Configures the two-stage unsupervised feature selection.

2.5 Center panels

2.5.1 Peptide sequences window

This window is opened from Window > Peptide sequences. The Peptide sequences window shows the result of applying a query, filter, or search. The rows showed can also be filtered by attributes such as ID, Sequence, Length, or calculated features.



Peptide	Sequence	Length	Boman
starPep_00000	GIGAVLKVLTTGLPALISWIKRKRQQ	26	0.5642307692307693
starPep_00001	GIGKFLHSAKKFGKAFVGEIMNS	23	0.41043478260869565
starPep_00002	ILPWKWPWWPWRR	13	1.0653846153846154
starPep_00003	KWCFRVCYRGICYRRCR	17	3.532352941176471
starPep_00004	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRN...	37	2.9918918918918918
starPep_00005	FFHHIFRGIVHVGKTIHRLVTG	22	0.7072727272727273
starPep_00006	FLPLIGRVLSGIL	13	-1.5461538461538462
starPep_00007	FWGALAKGALKLIPSLFSSFSKKD	24	0.04749999999999951
starPep_00008	GWGSFFKKAAHVGKHVGKAALTHYL	25	0.19520000000000004
starPep_00009	ACNFQSCWATCQAQHSIYFRRAFCDR...	36	2.143333333333333
starPep_00010	ALWKTMLKKLGTMALHAGKAALGAAAD...	34	0.15735294117647056
starPep_00011	FLPLLAGLAANFLPTIICKISYKC	24	-1.2708333333333337
starPep_00012	FLPVLAGIAAKVVPALFCKITKKC	24	-1.2020833333333334
starPep_00013	GLLSVLGSVAKHVLPHVVPVIAEHL	25	-1.0467999999999997
starPep_00014	GLNTLKKVFQGLHEAIKLINNHVQ	24	0.91375
starPep_00015	HVDKKVADKVLLLKQLRIMRLLTRL	25	1.6532
starPep_00016	IWLTALKFLGKHAAKHLAKQQLSKL	25	0.23760000000000006
starPep_00017	KTCEHLADTYRGVCFTNASCDDHCKNK...	46	2.0821739130434778
starPep_00018	KWKLFKKIEKVGQNIRDGIIKAGPAVAVV...	37	0.8372972972972975
starPep_00019	LCNERPSQTWSGNCGNTAHCDKQCQD...	50	2.745799999999999
starPep_00020	RGGRLCYCRRRFCVCVGR	18	3.652222222222222
starPep_00021	ACYCRIPACIAGERRYGTCIYQGRLWAFCC...	30	1.0763333333333331
starPep_00022	ALWKNMLKGIGKLAGKAALGAVKKLVGAES	30	-0.26633333333333337
starPep_00023	DHYNCVSSGGQCLYSACPIFTKIQGTCY...	36	1.3038888888888887
starPep_00024	DKLIGSCVWGAVNYTSDCNGECKRRGY...	44	1.7445454545454544
starPep_00025	FKCRRWQWRMKKLGAPSITCVRRAF	25	2.7455999999999996

Figure 2.10: Peptide sequences window

2.5.2 Network visualization window

This window is opened from Window > Network Visualization. It consists of two views: Scene and Preview. The Scene view allows to customize some visual properties of the network such as background color, zoom, position, and individual colors for edges and nodes. The options highlighted in Fig. 2.11 are the following:

1. Switch background.
2. Zoom options.
3. Selector. It allows to change the node diameter of the cursor while selecting nodes.
4. Additional options. It allows to enable or disable the options Autoselect neighbors and Show peptide labels. **Note:** we recommend disabling the latter in order to render clearer graphs in metadata network analysis.
5. More advanced sizing and coloring options for nodes. By pressing More.../Less..., the options are shown/hidden.
6. Network rendering area.
7. Node label options. The first one allows to show/hide the node labels. The second one brings three options to modify node label size: Fixed, Scale size, and Node size. **Note:** The option Node size is handy for adjusting the label size proportionally to the node size. The third one modify the label color options. There are three choices: Unique, Object, and Text.
8. Node label font properties.
9. Node label size.

10. Two edges options. The first one shows/hides edges. The second one enables edges to have the attached node color.
11. Edge thickness.
12. Shows/Hides edge labels.
13. Edge label font properties.
14. Edge label size.

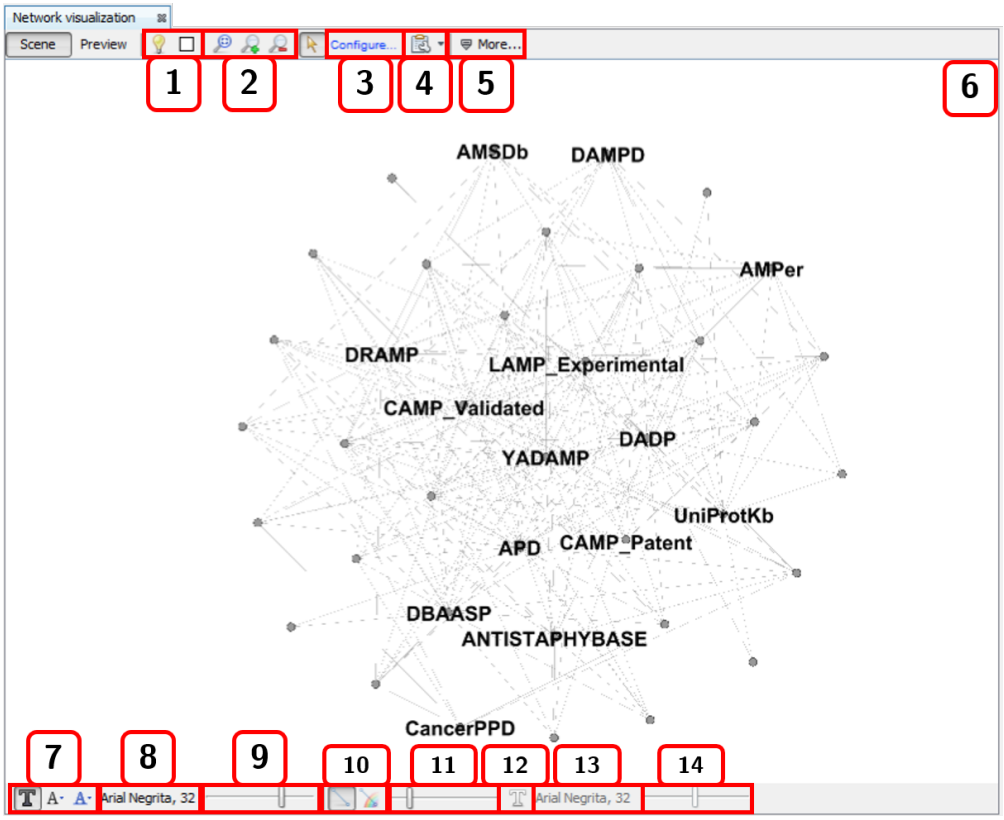


Figure 2.11: Network visualization window: Scene view.

Note: When you right-click the mouse on the scene view, a context menu is displayed.

The Preview view shows the rendered the graph according to the calculated layout and all the configurations. Attractive networks may be rendered in this other view. To update the drawing, press the Refresh button.

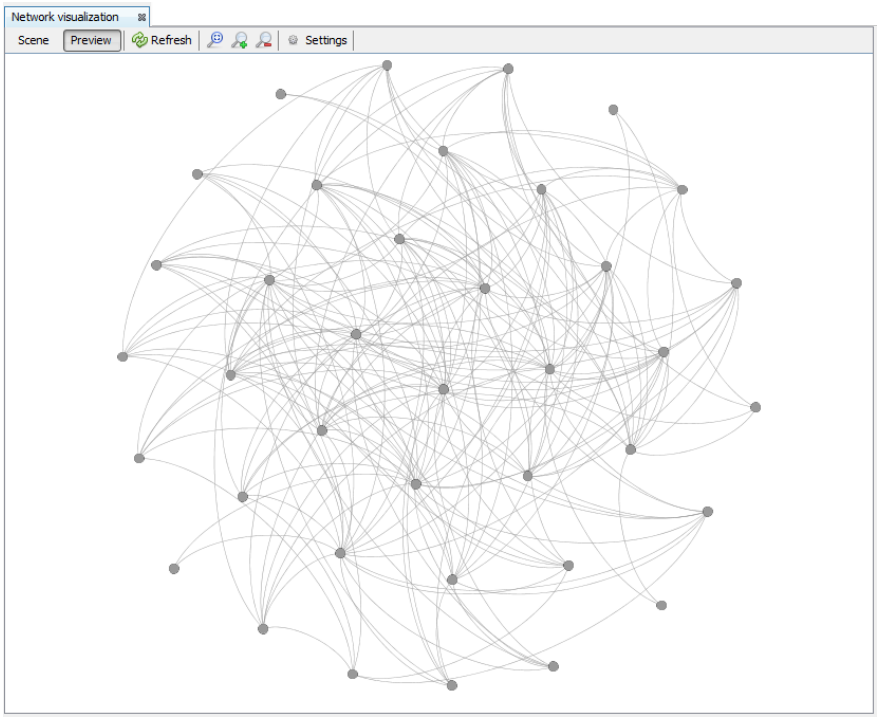


Figure 2.12: Network visualization window: Preview view.

2.6 Navigator panel

This panel is opened from Window > Navigator. The navigator changes between the Metadata relationships and Graph table options according to whether the Peptide sequences or Network visualization window is active.

On the one hand, in the Metadata relationships view, the user can seek metadata nodes. Right-click on a row will show a context menu to select or center nodes on the graph, as well as the Properties window for the relationship.

Note: If a peptide sequence is selected in the center panel, only the related metadata are shown in this navigator panel. Click on the Refresh button to show all.

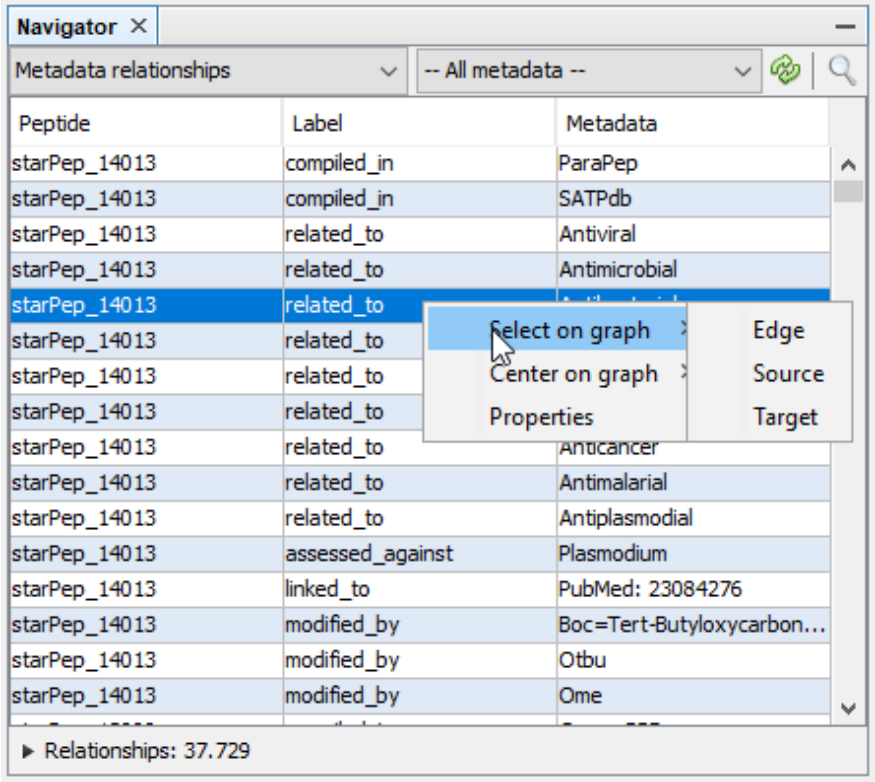


Figure 2.13: Navigator for the Peptide sequences window

On the other hand, in the Graph table view, the user can switch the view from nodes table to edges table, and also customize the columns (such as network measures) shown in the data grid. These data tables can be exported to an external text file (CSV format). There is also a context menu that is accessed via right-click on any row.

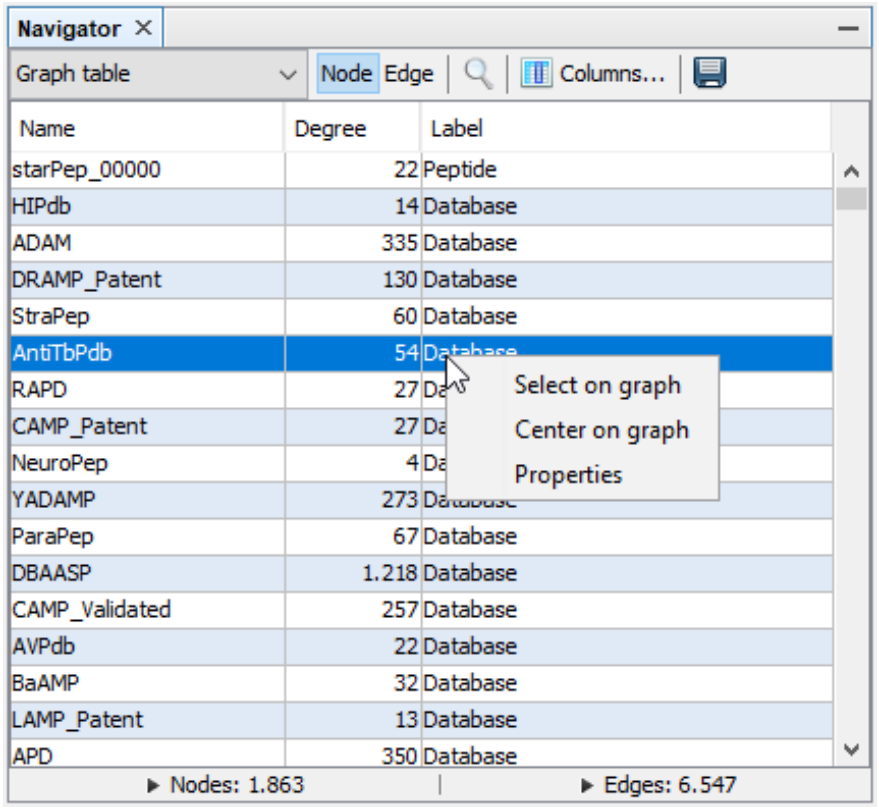


Figure 2.14: Navigator for the Network visualization window

Chapter 3

Working with networks

3.1 Metadata network

The construction of metadata network is accessible from the menu option: Tools > Network > Metadata Network

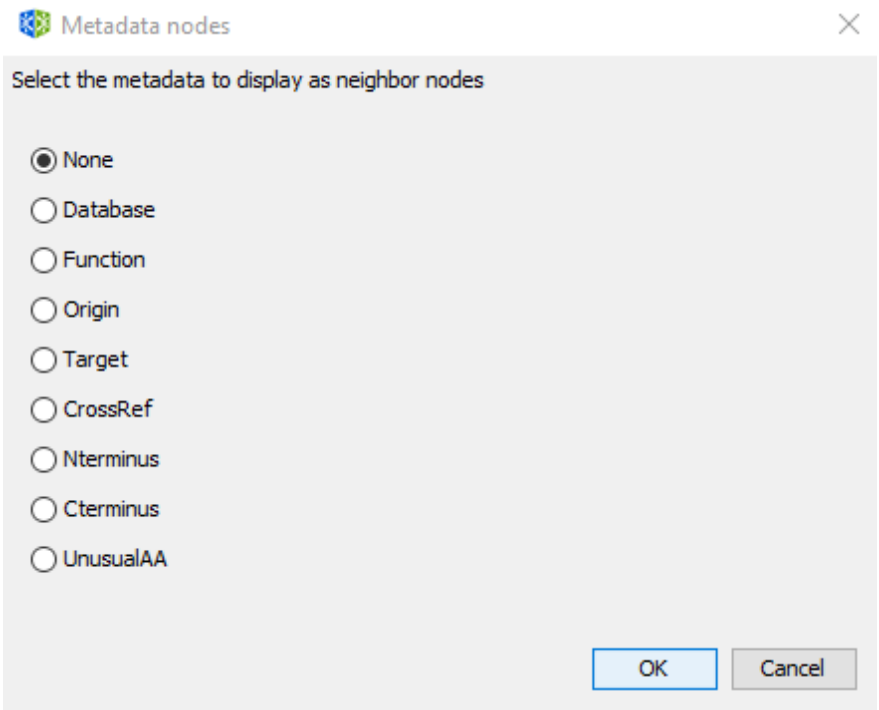


Figure 3.1: Options for metadata network

In metadata networks, nodes representing metadata are connected to nodes representing peptides by the following relationships.

Metadata node	Relationship
Origin	*produced_by*
Target	*assessed_against*
Function	*related_to*
Database	*compiled_in*
Crossref	*linked_to*
Nterminus	*modified_by*
UnusualAA	*constituted_by*
Cterminus	*modified_by*
Subcategory of another node	*is_a*

Table 3.1: Metadata node names and relationships in starPepDB.

3.2 Similarity network

The construction of similarity network is accessible from the menu option: Tools > Network > Similarity Network. To create a similarity network, we first recommended to configure the workflow using the Configuration Wizard and then press the button Run.

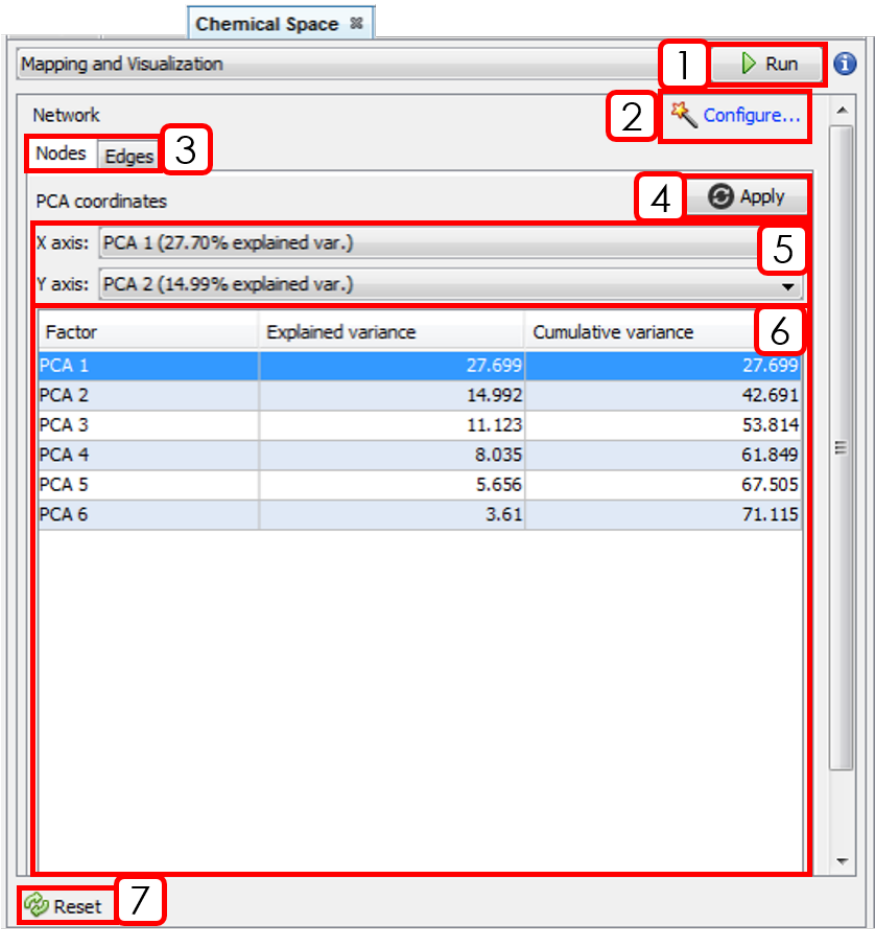


Figure 3.2: Chemical Space window (Nodes tab).

1. Runs the workflow for building the similarity network.
2. Opens Configuration Wizard (Sect. 3.2.1).
3. Changes between Nodes and Edges tabs.
4. Applies PCA coordinates changes.
5. Selects X and Y axis for PCA coordinates.
6. PCA results panel.

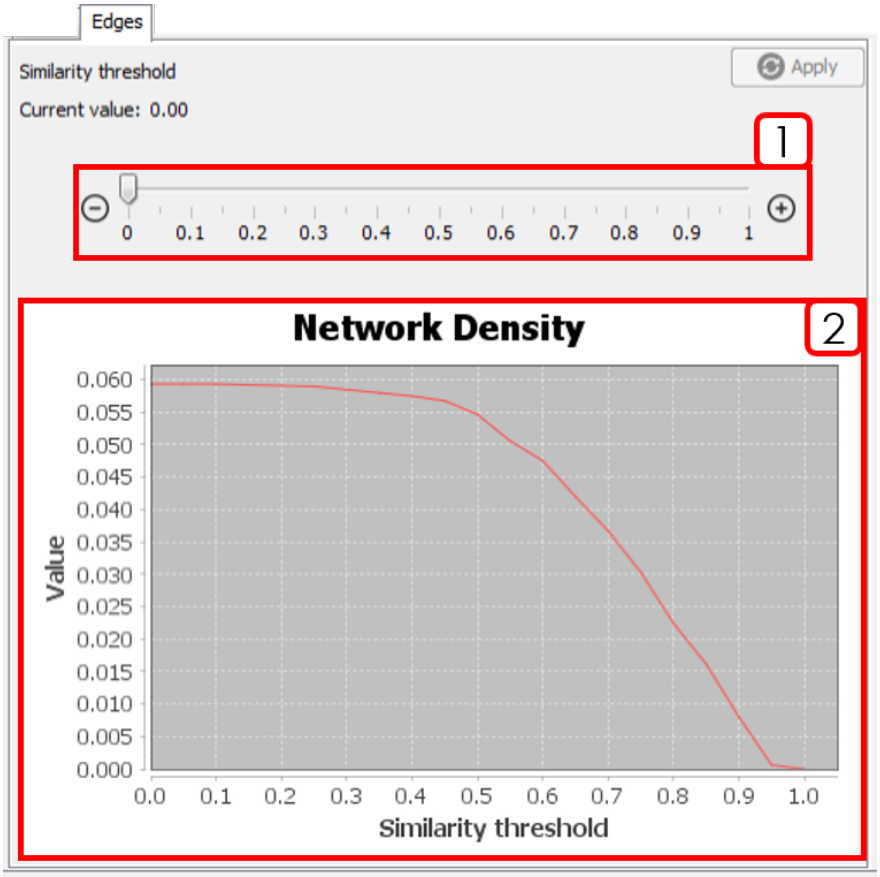


Figure 3.3: Chemical Space window (Edges tab).

1. Similarity threshold selector. After changing the value, it is necessary to press ”Apply”.
2. Network Density plot. Helps to decide a similarity threshold.

3.2.1 Configuration wizard

This section will show the configuration wizard for mapping and visualizing the Chemical Space.

Wizard Step 1: Input sequences

To remove redundant sequences, press Yes (**recommended**). Then, you can choose between local or global alignment, multiple substitution matrices, and a identity threshold.

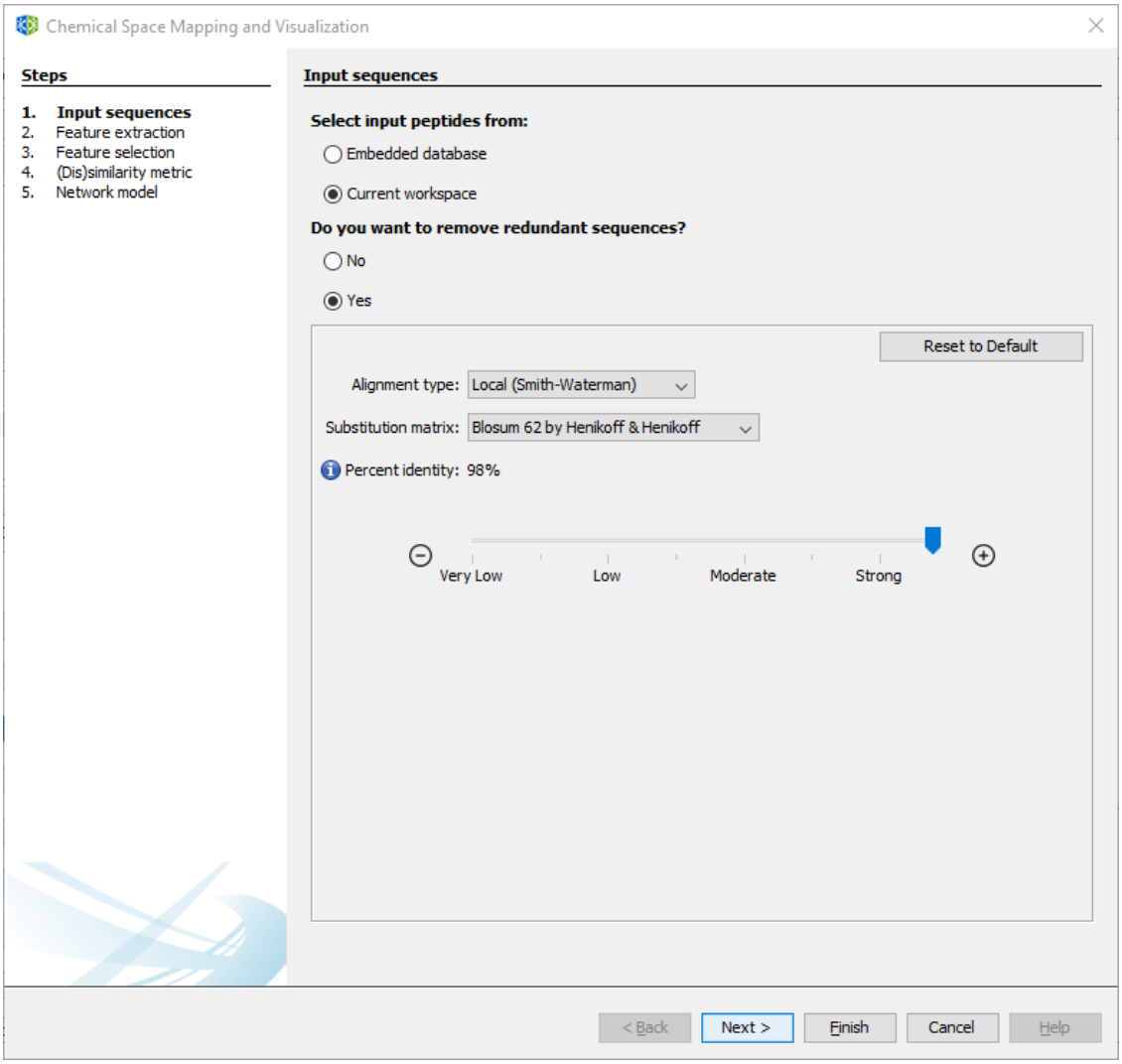


Figure 3.4: Wizard Step 1: Input sequences.

Wizard Step 2: Feature extraction

If you already calculated a set of molecular descriptors, you can select the first option and press Next. If not, select the new descriptors to be calculated.

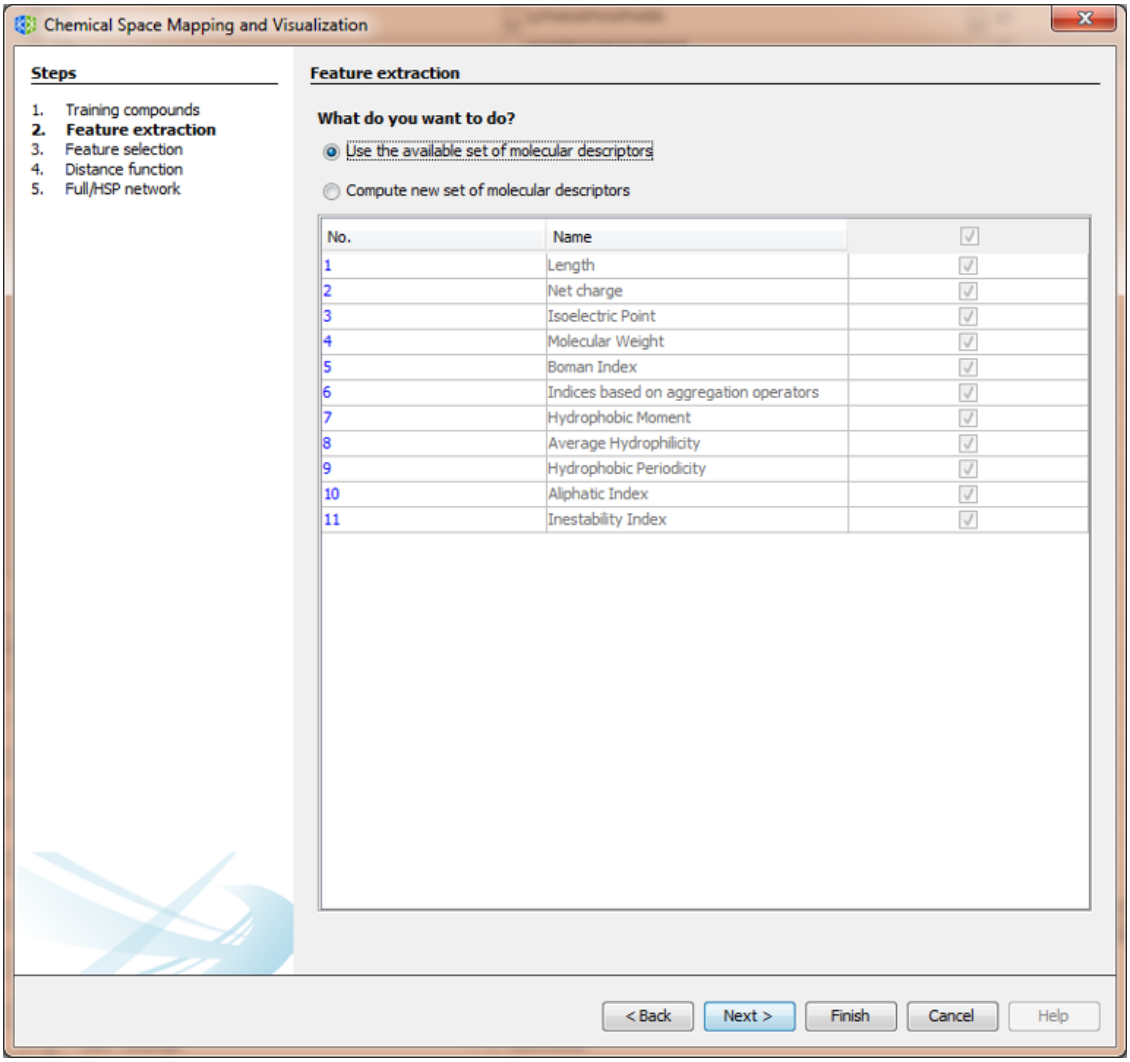


Figure 3.5: Wizard Step 2: Feature extraction

Wizard Step 3: Feature selection

If you plan to use all available descriptors, select the first option, and press Next. If not, select and configure the two-stage unsupervised feature selection method.

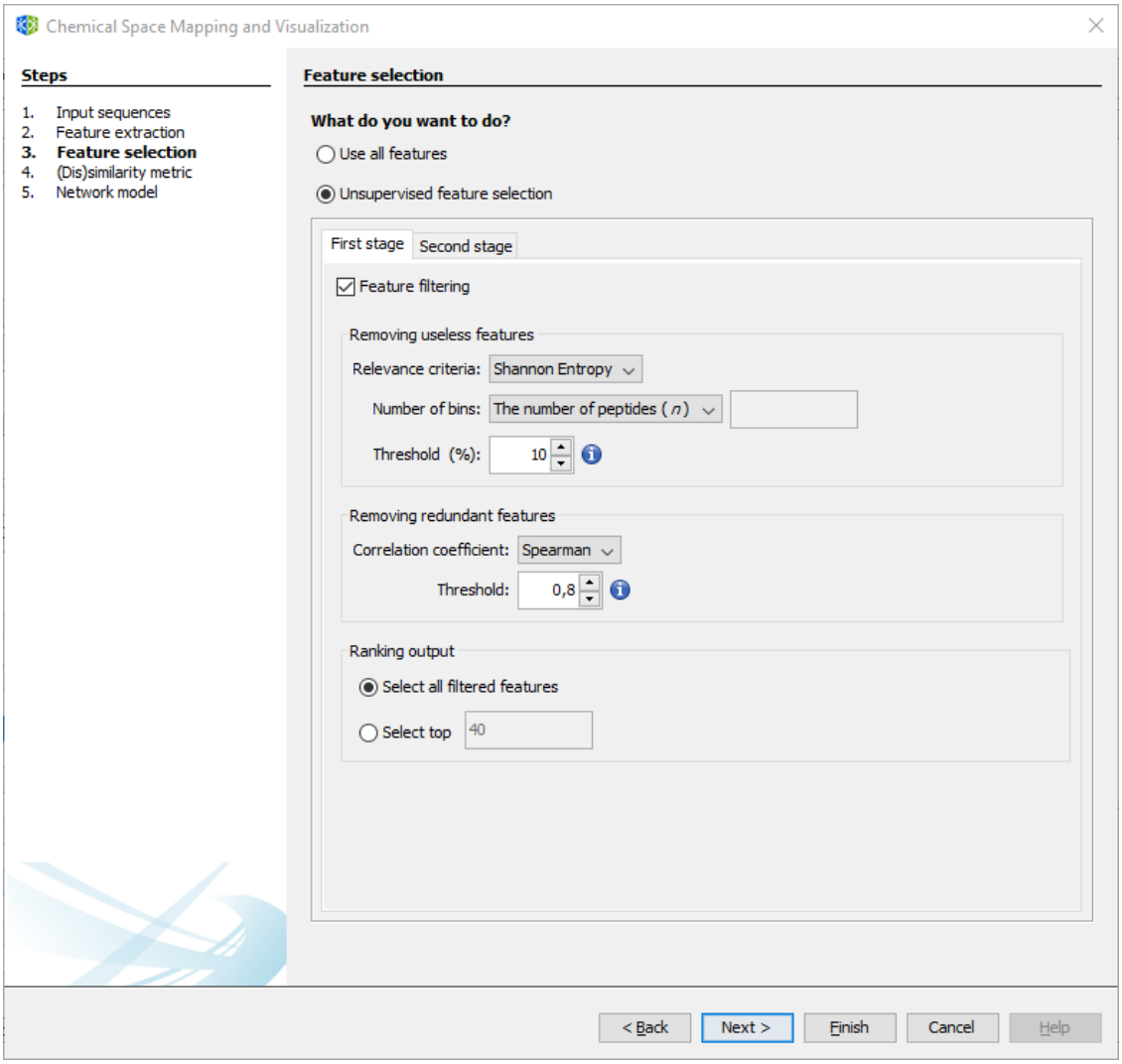


Figure 3.6: Wizard Step 3: Feature selection.

Wizard Step 4: Distance function.

Select the desired distance function and the standardization/normalization for the calculated descriptors.

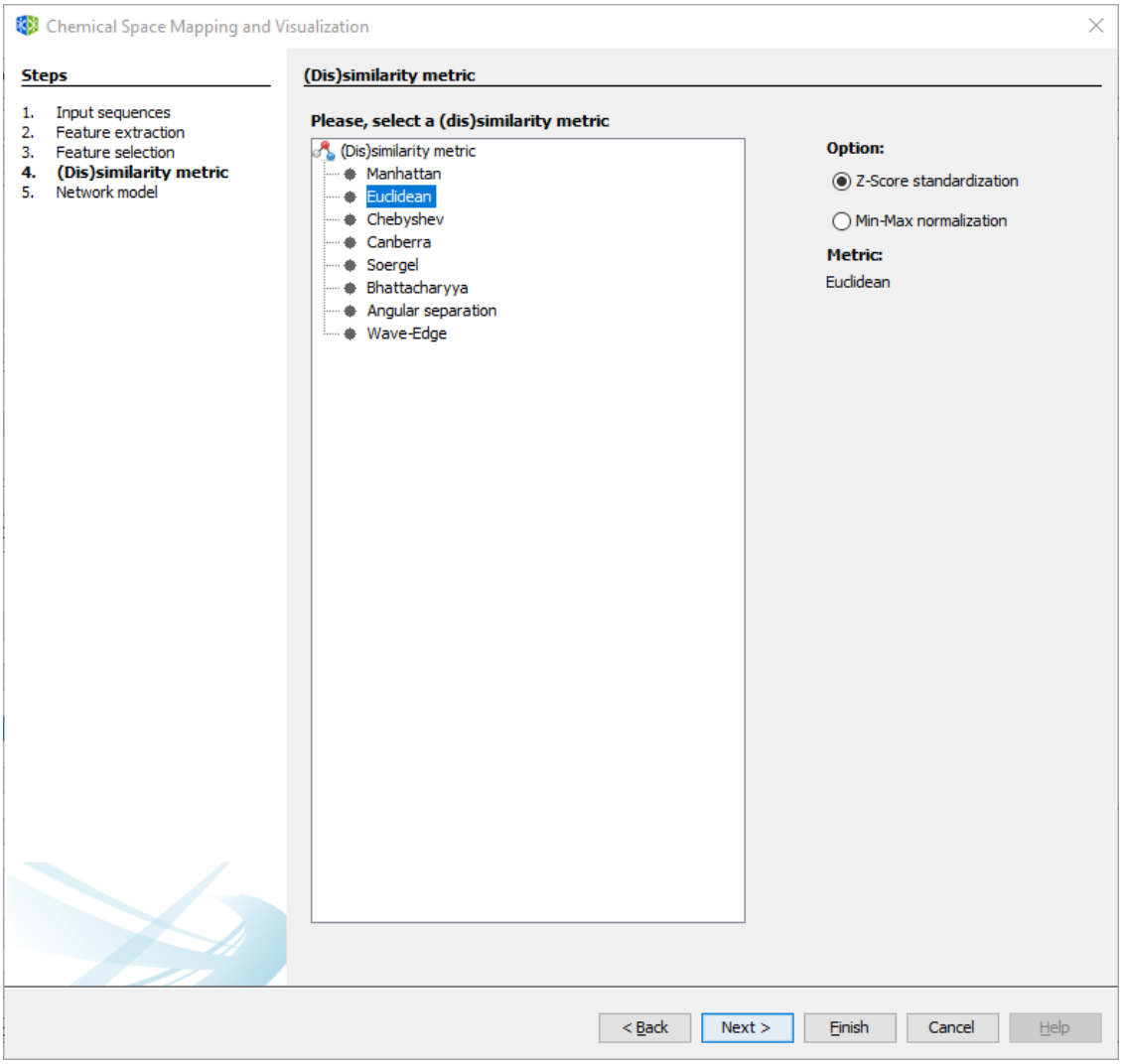


Figure 3.7: Wizard Step 4: Distance function.

Wizard Step 5: Network model

For generating a network model, select between the Half-Space Proximal Network or the traditional Similarity Network (**not recommended for large datasets**). For more details, please refer to the main paper.

Note: The position of nodes may be determined by the first two principal components of descriptor space. However, layout algorithms are recommended for a better rearrangement of nodes.

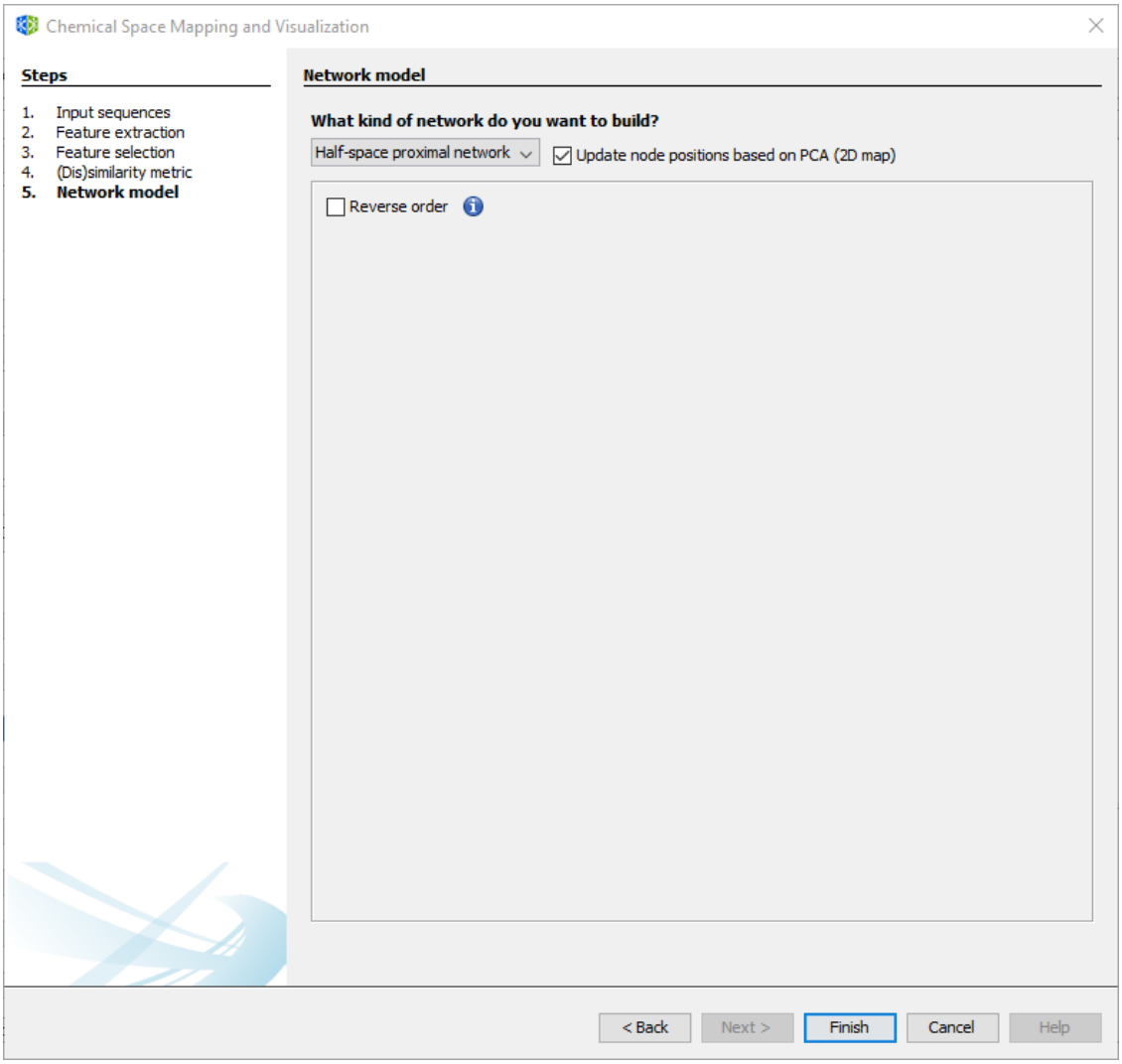


Figure 3.8: Wizard Step 5: Network model.

3.3 Network model options

After creating the network model, the following options are available.

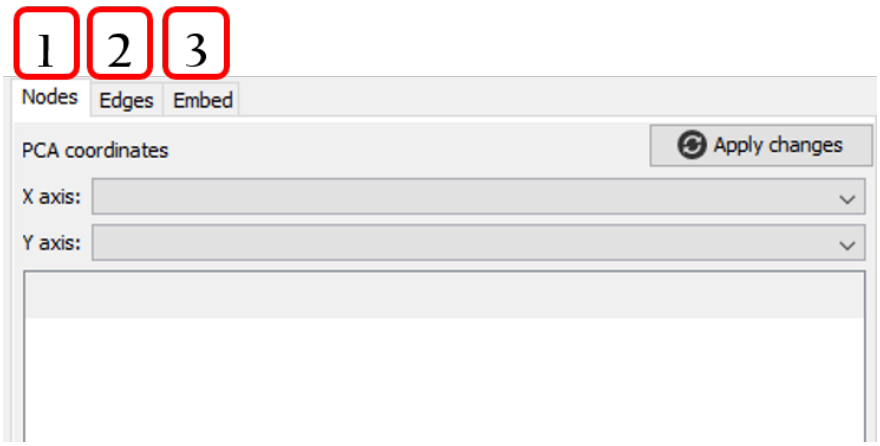


Figure 3.9: Network model options.

1. Positioning nodes.
2. Adding/removing similarity edges.
3. Embedding new peptides. When new peptides are projected, a network model will be opened into a new workspace.

3.4 Layout algorithms

A layout algorithm option may be opened from Tools > Network > Layout > [layout option]. The main graph layouts available are Fruchterman Reingold, ForceAtlas 2, Yifan Hu Proportional, and Random Layout. Any layout result could be adjusted using the options Rotate, Contraction, Expansion, Noverlap, and Label Adjust.

Note: ForceAtlas 2 is recommended to handle large networks while keeping a very good quality. Random Layout may be a good start point for applying new layouts. Fruchterman Reingold and Yifan Hu Proportional may be used for small or middle size networks.

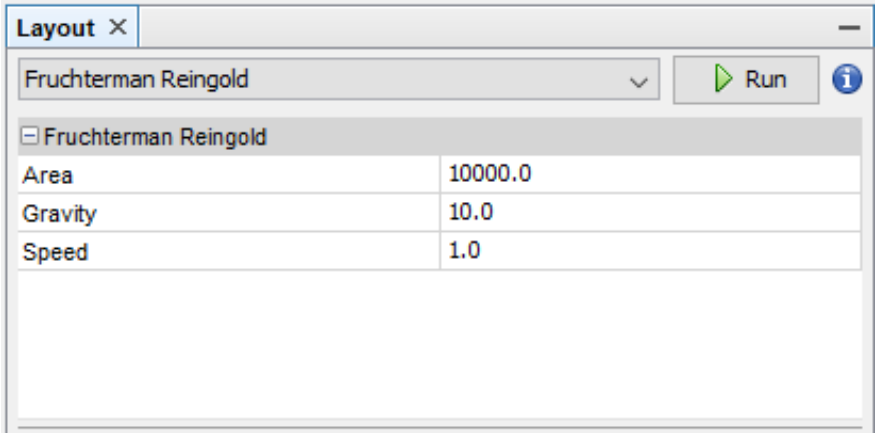


Figure 3.10: Fruchterman Reingold option.

3.5 Appearance

This panel is opened from Tools > Network > Appearance.

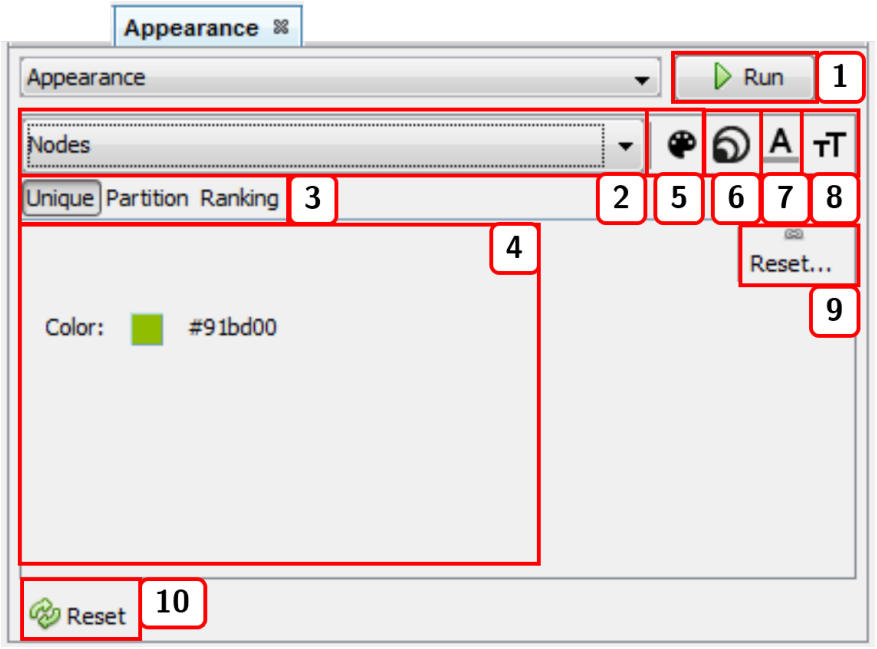


Figure 3.11

1. Runs the appearance customization of either nodes or edges. **Note:** if the Preview window is active in the Network visualization window, you need to press the button Refresh from said window to update the network view.
2. Selects the elements (either Nodes or Edges) whose appearance is to be changed.
3. Applies configuration via Unique, Partition, or Ranking functions. For nodes, the calculated measures are available as Partition or Ranking options.
4. Modifiable configurations panel. For color options, you need to press and drag the cursor to the desired color, or press right-click to open the color window.
5. Changes the color of either Nodes or Edges (if edges are not taking the color of attached nodes, see sect. 2.5.2).
6. Changes Nodes size (this option only applies to nodes).
7. Changes label color of either Nodes or Edges.
8. Changes label size of either Nodes or Edges.
9. Resets current options.
10. Resets customization to the default appearance.

3.6 Clustering

A clustering panel (Fig. 3.12) may be opened from Tools > Network > Clustering > [clustering algorithm]. For instance, k-means:

Note: After running the clustering algorithm, you may visualize the network structure in Tools > Network > Appearance > Nodes > Partition.

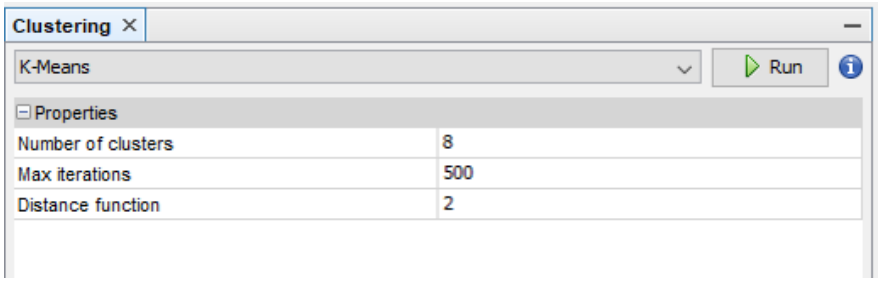


Figure 3.12: k-means clustering option.

3.7 Centrality

A centrality panel (Fig. 3.13) may be opened from Tools > Network > Centrality > [measure option]. For instance, Betweenness Centrality:

NOTE: After running the centrality measure, you may customize the appearance of nodes according to the centrality values in Tools > Network > Appearance > Nodes > Ranking.

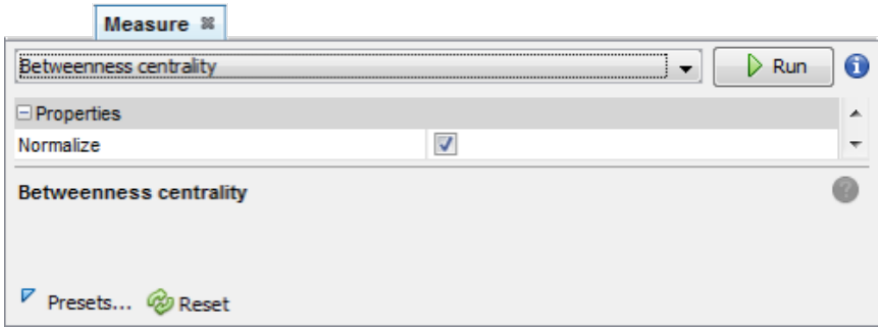


Figure 3.13: Betweenness Centrality option.

3.8 Case study

In this case study, we try to answer the following questions for a given sequence of interest.

```
>Example sequence
FLPAIVGAAGQFLPKIFCAISKKC
```

3.8.1 Which biological database holds peptides similar to the sequence of interest?

Step 1: Opens the Search panel with the commands Tools > Peptide search by > Single query sequence. Types the query sequence in the input field, configures the sequence alignment at 70% of sequence identity, and press Run. This search should return 25 peptide sequences and 595 metadata relationships.

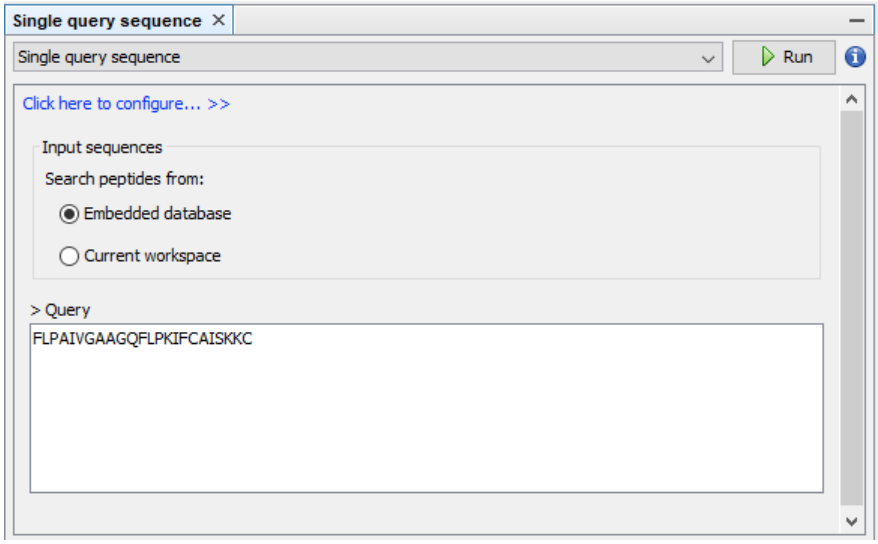


Figure 3.14: Sequence search panel

Step 2: Creates the metadata network by selecting the option Database.

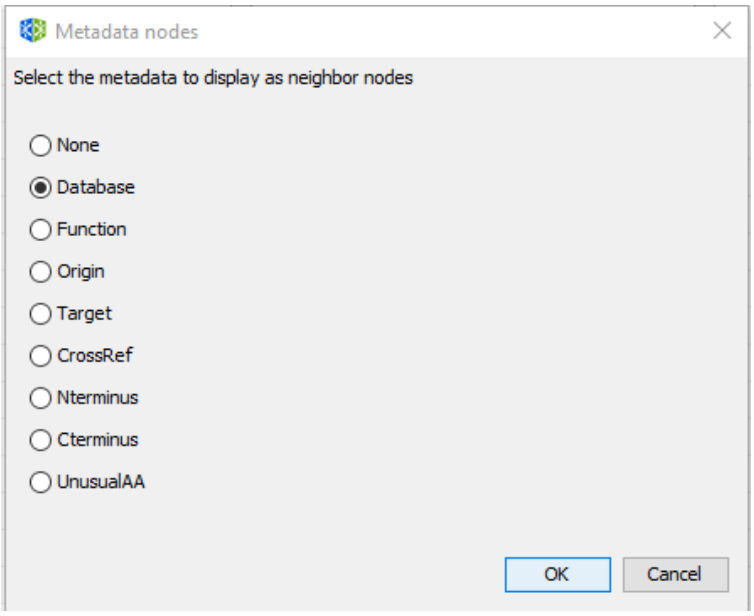


Figure 3.15: Options for metadata network

Step 3: In the graph table view of Navigator window, select the option Columns... and then mark Degree and click OK.

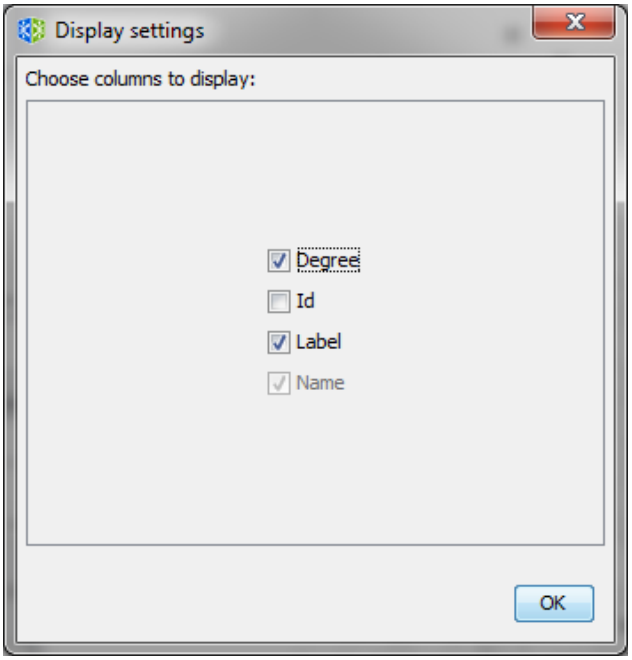


Figure 3.16: Display settings

We can sort the graph table by node Degree by clicking the Degree column 2 times, and now we can observe that the database **SATPdb** contains the most similar sequences to the query sequence.

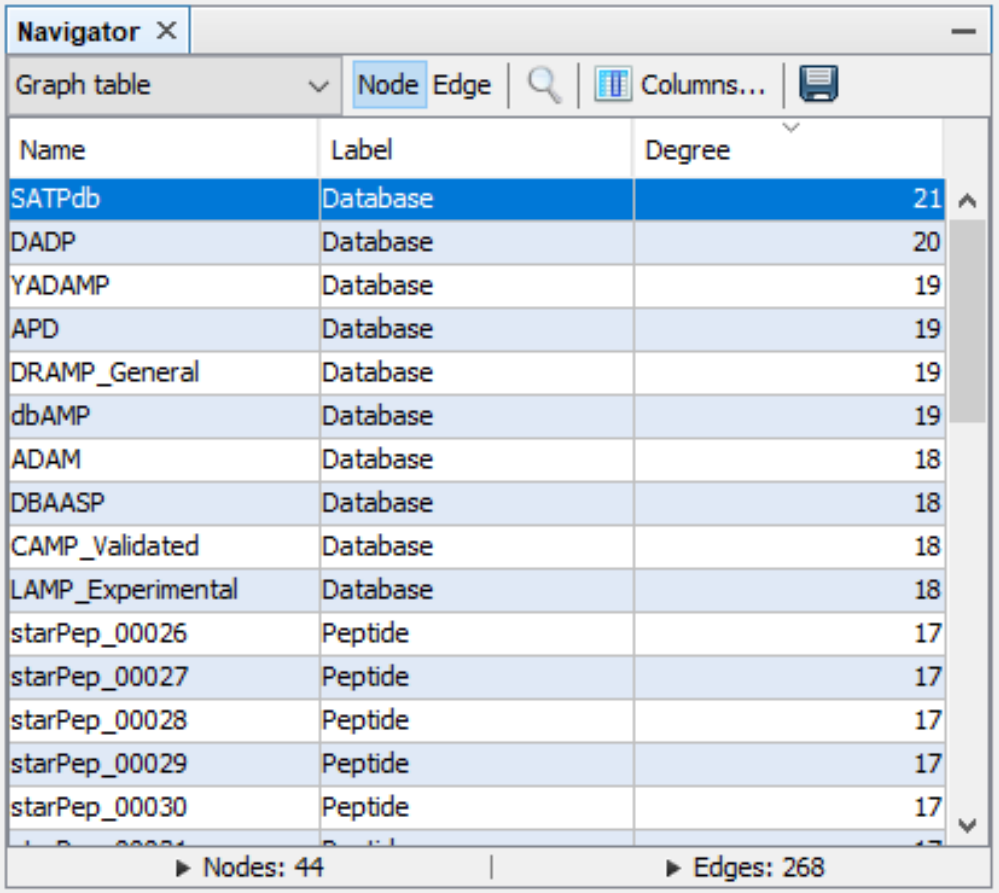


Figure 3.17: Graph table

3.8.2 What are the biological functions of peptides similar to the sequence of interest?

Follow the **Step 1** of the previous example.

Step 2: Creates the metadata network by selecting the option Function.

Step 3: In the graph table view of Navigator window, select the option Columns... and then mark Degree and click OK.

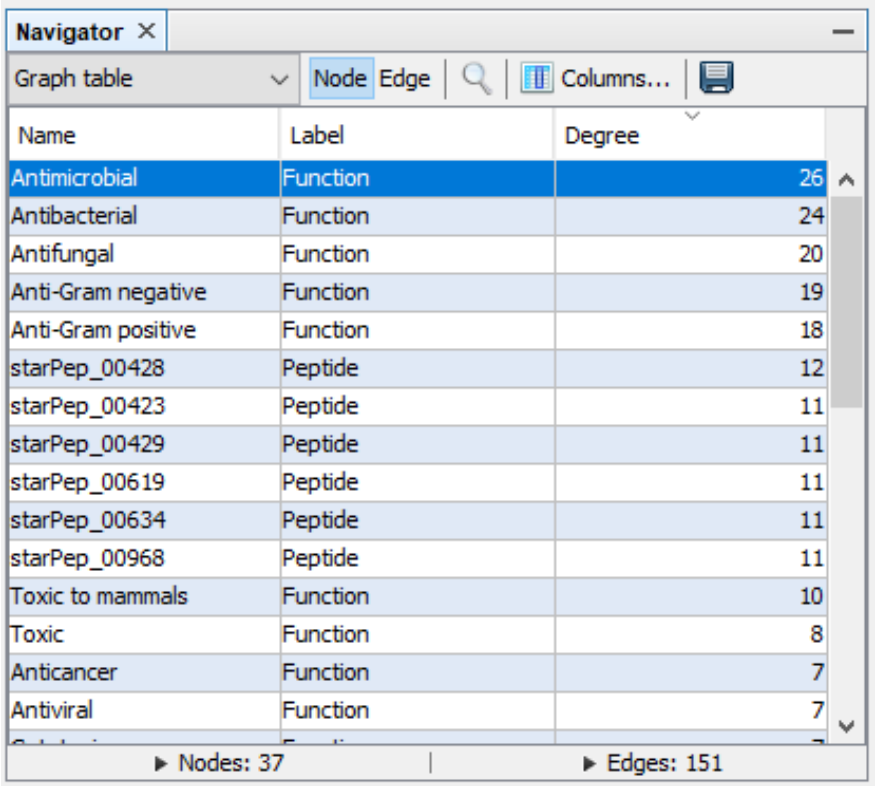


Figure 3.18: Graph table

Step 4: In the Network visualization window, select the following options.

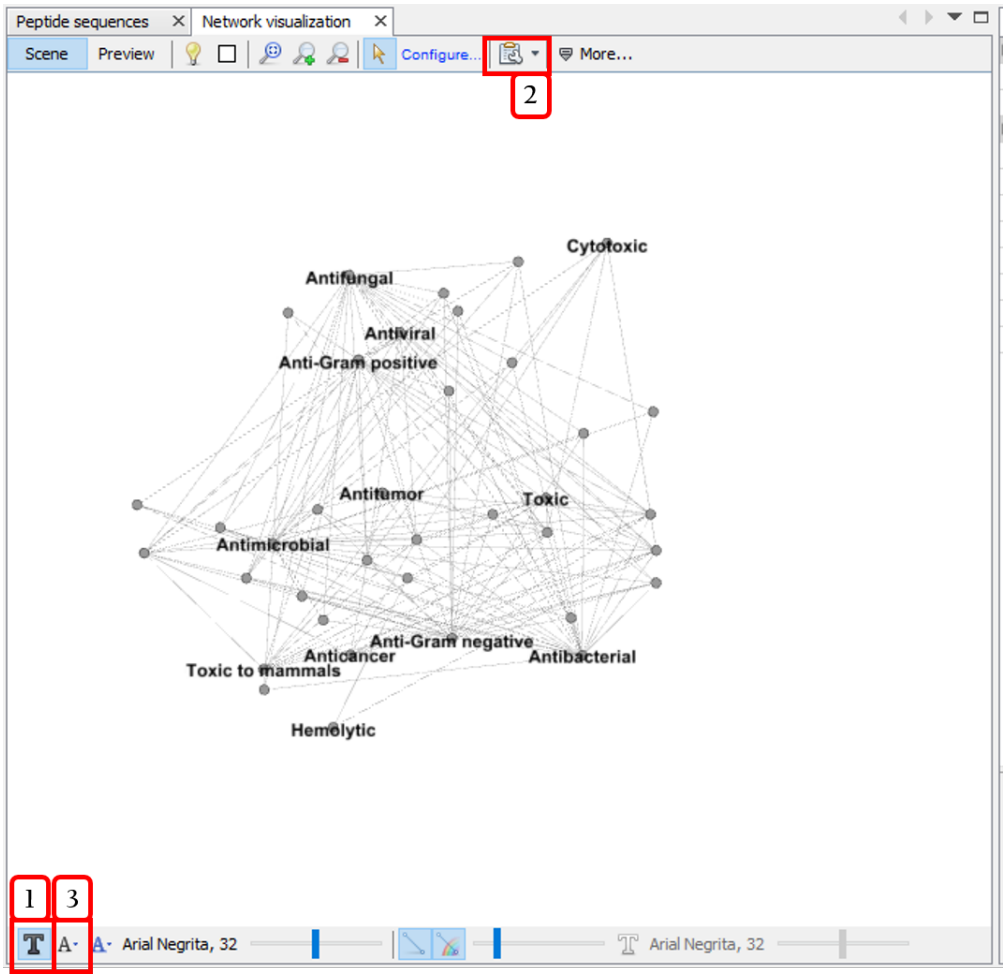


Figure 3.19: Network visualization

1. Shows node labels.
2. Disables the option Show peptide labels.
3. Modifies the label size to Node size.

In the Appearance panel (see sect. 3.5), customizes the appearance of nodes for sizing and coloring nodes according to the degree measure.

1. In the Nodes view, select Node size > Ranking > Degree. Set min and max sizes to 5 and 100 respectively, select the interpolator Bezier, select the second predefined spline and press Run.
2. In the Nodes view, customizes Node color > Ranking > Degree, and press Run.

Run the Tools > Network > Layout > Fruchterman Reingold about 10s and then press stop. The result may be similar to the following network:

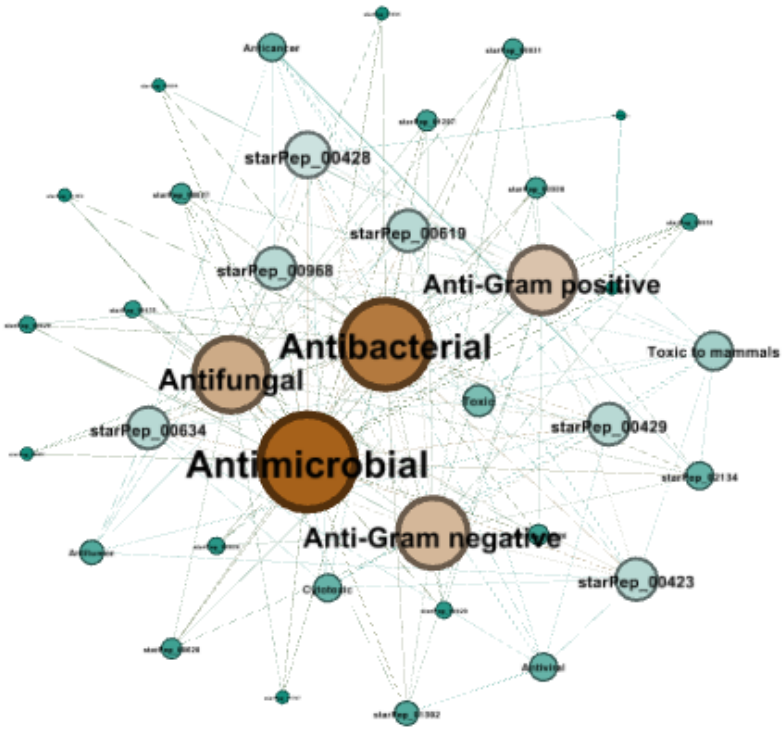


Figure 3.20: Metadata network