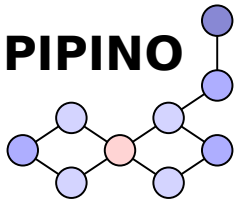
Short User Manual for PIPINO



provided for version v0.3b

Autor: Stefan Schildbach

Last update: 28th October 2015

Contents

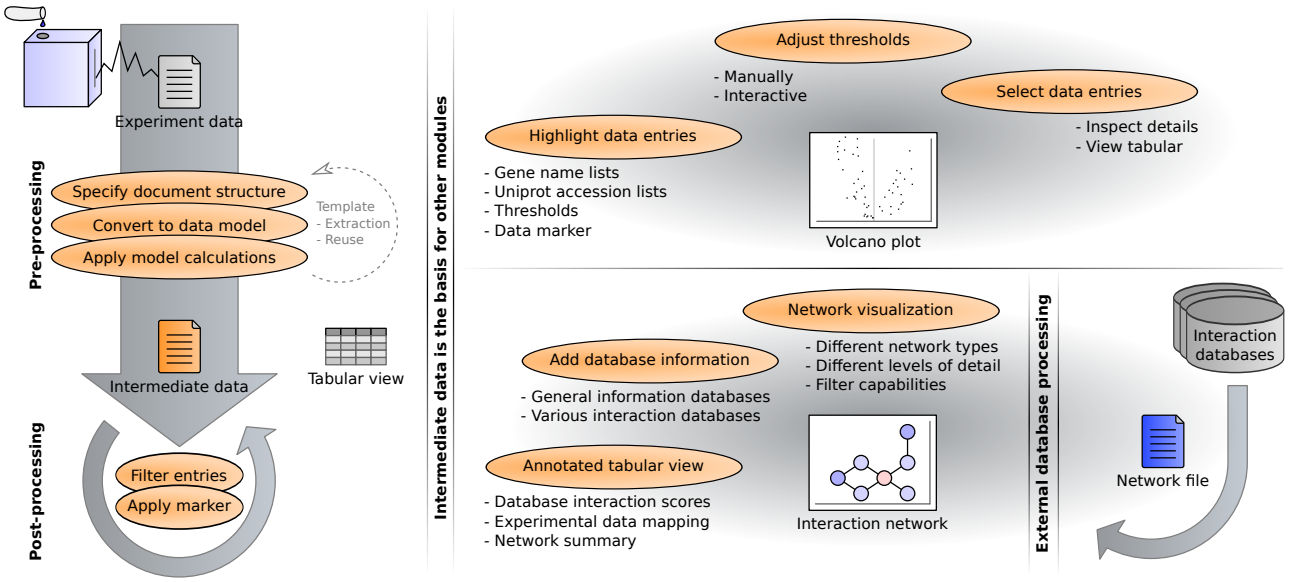
1 Requirements and Installation

PIPINO is written in Java and therefore not limited to a single operating system. It should run on every system with Java 7 (1.7) or higher installed. It is recommended to have at least 2 GB of RAM and a decent processor. The real hardware requirements depend on the analyzed data size and the used network size.

The software package is portable and therefore the extraction of the compressed archive to a user defined location is everything it takes to install PIPINO. The archive contains two startup scripts to launch PIPINO and a program folder. Choose the appropriate launch script depending on your operating system:

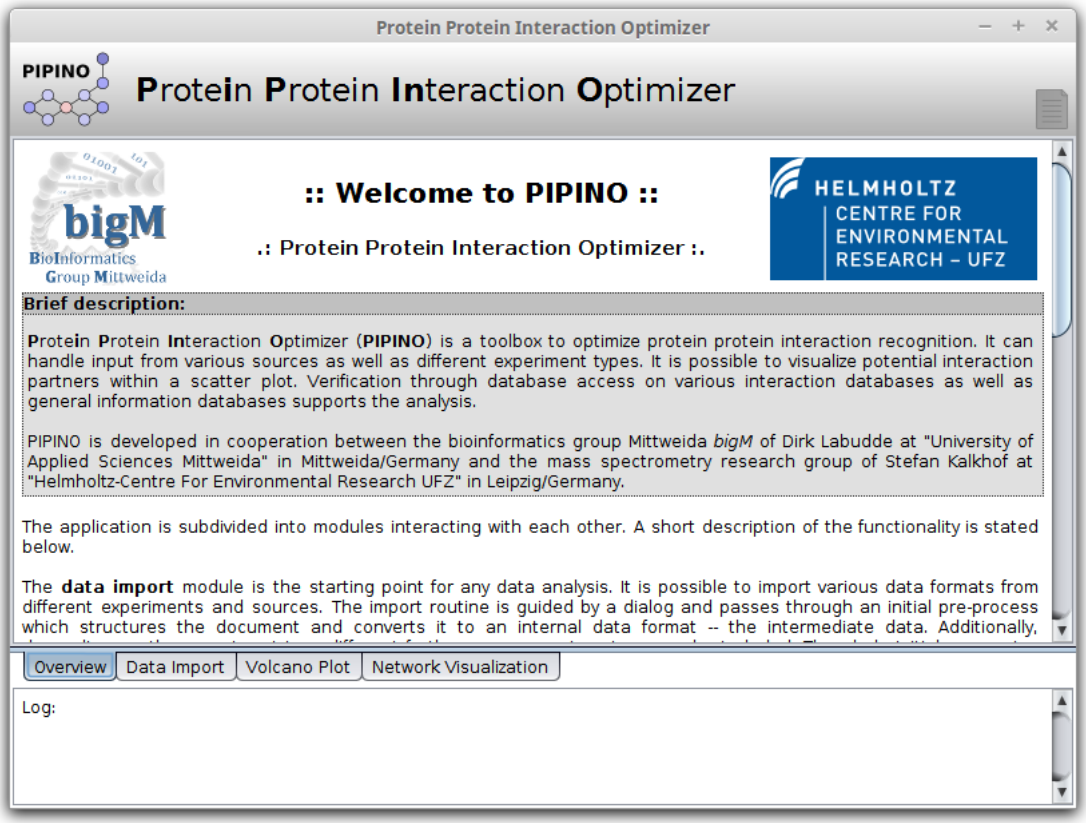
PIPINO.bat for Windows
PIPINO.sh for Linux

PIPINO offers different modules interacting with each other. A general workflow for orientation purposes is depicted below.



2 Overview and Window Layout

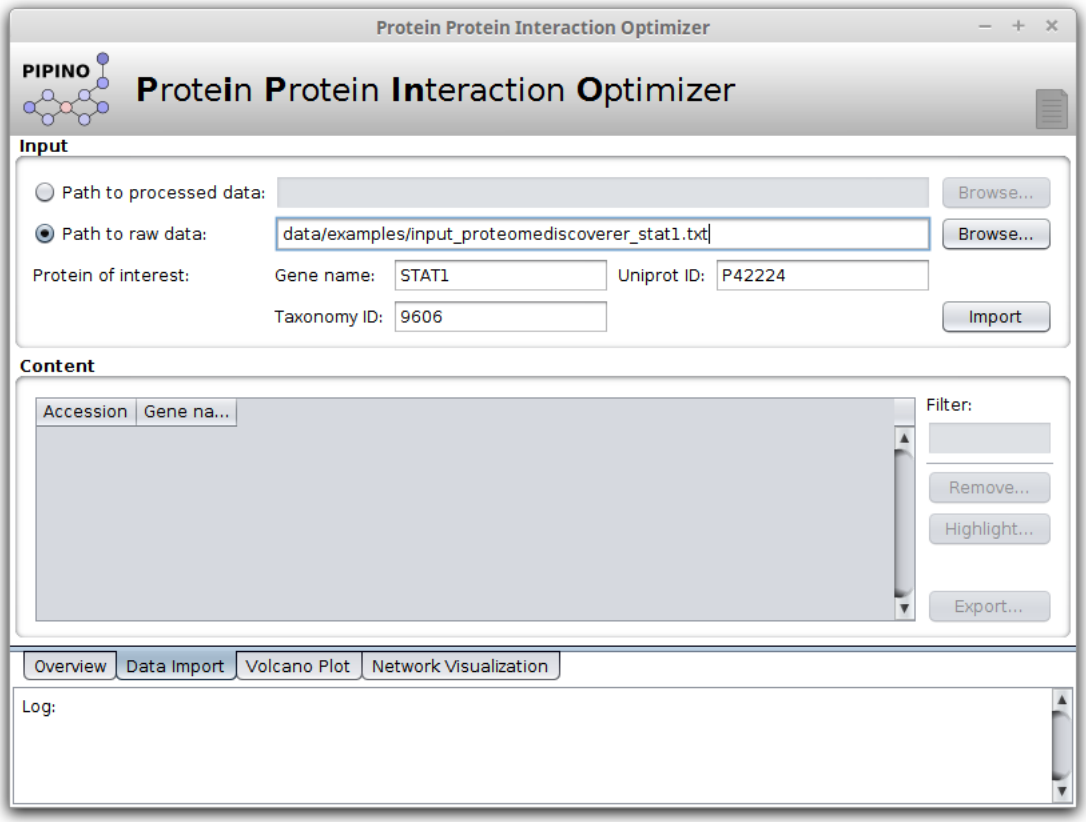
After starting up, the overview tab should be displayed with general information about PIPINO and a short description of the current modules. The main window is divided into four areas.



The program logo and name, and the data indicator are located in the upper part of the window. The primary working area is positioned below this header and displays the core functionalities of the modules. A module can be accessed via the tab bar following the center area. To trace the current program status a log area is present at the bottom of the window.

3 Getting started – Data Import

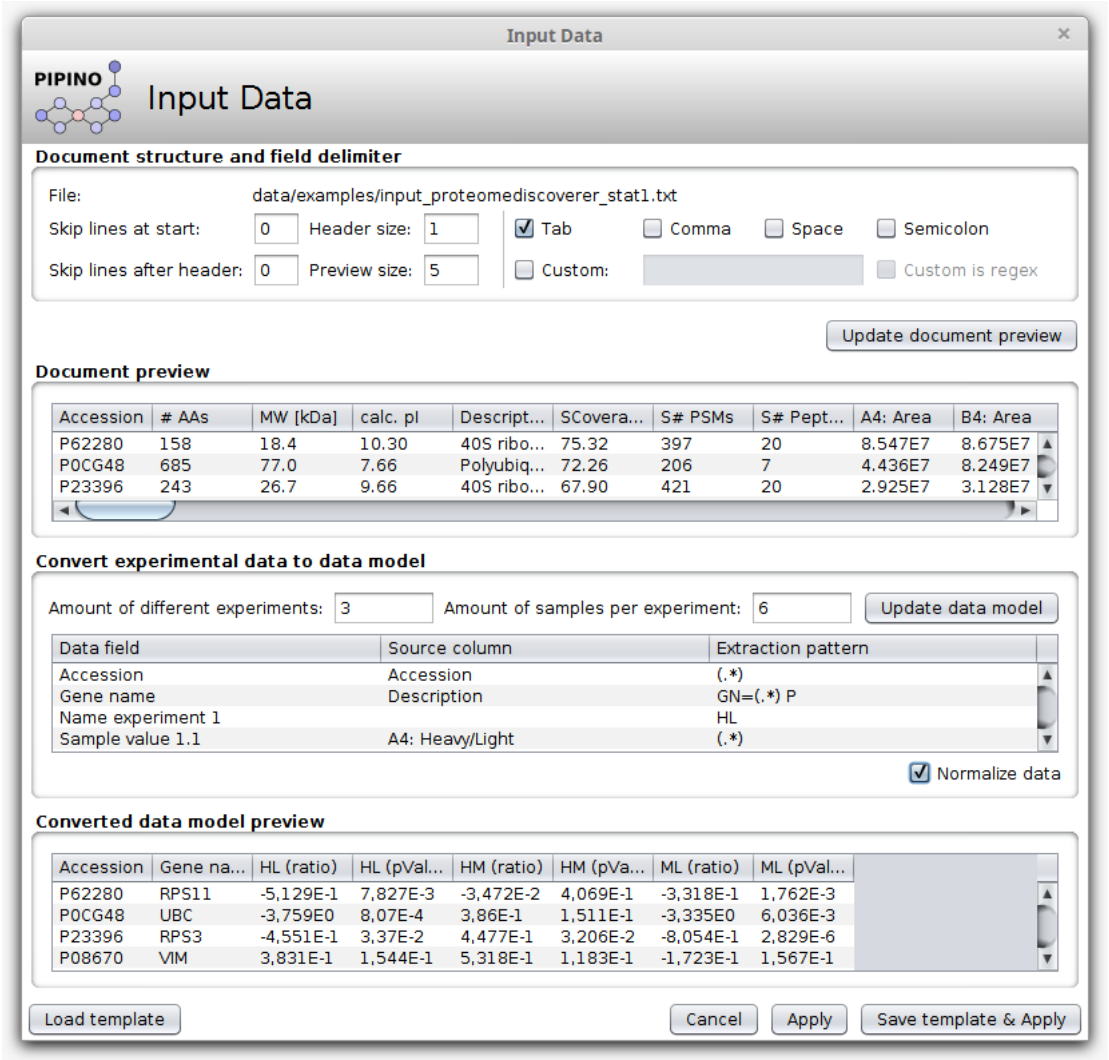
The first step towards a successful data analysis is the conversion of your experiment data set to an intermediate format understandable by the software. This can be done with the *Data Input* module.



To start the pre-processing step select an appropriate data file either from already processed data (data exported by PIPINO) or from raw data (unprocessed data from your experiment) and specify the needed values for the protein of interest. Afterwards, you can push the *Import* button to structure your document with the import dialog.

3.1 Pre-processing

First you need to specify the document structure and field delimiters. You can skip leading lines at the start of your file (e.g. comments or version numbers) and after the header (e.g. comments or for changing previewed data rows). The header size should normally be a single line, but if you have multiline headers you can increase this number. The preview size changes the amount of lines parsed from your document into the preview section. A relatively small



number can be used to reduce the processing time for document previewing. Furthermore you should specify the field delimiter in your document. Simple delimiters (e.g. tab, comma, ...) can be selected via the appropriate checkboxes, while a custom delimiter (possibly a regular expression[1]) can be specified in the corresponding textfield. You can evaluate your parameters by pressing the *Update document preview* button that activates the document preview. Adjust the document structure until it represents your data best.

In the next step the actual transformation (respectively conversion) is configured. You need to adjust the amount of different experiments and samples per experiment first. By pushing the *Update data model* button you can apply these changes. Subsequently you need to map your experiment data to the internal data model. This is done by selecting a source column in the table by clicking on the cell and choosing a column from the preview. Afterwards you

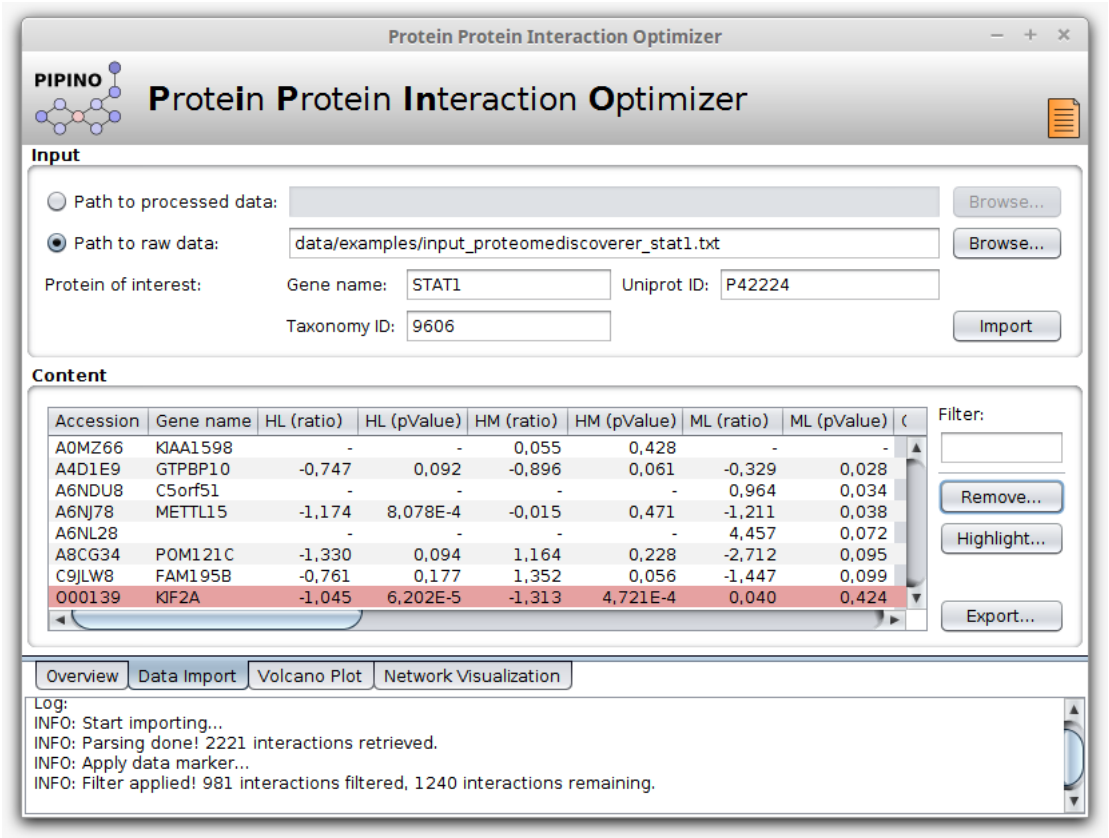
[1]If you are uncommon with regular expressions please refer to an online tutorial like `http://www.regular-expressions.info` or a live tool like `http://regex101.com`

should adjust the extraction pattern by clicking into the cell. This pattern is described by a regular expression (also called regex). To specify a default value you can simply enter the value (mostly useful for experiment names). To use a value directly from your data you can use this expression: „(.*)“, which means: take everything. To extract only a specific part of your columns you need to specify an appropriate regular expression with a single capturing group (e.g. GN=(.*) P). You can normalize your data by checking the box below the table.

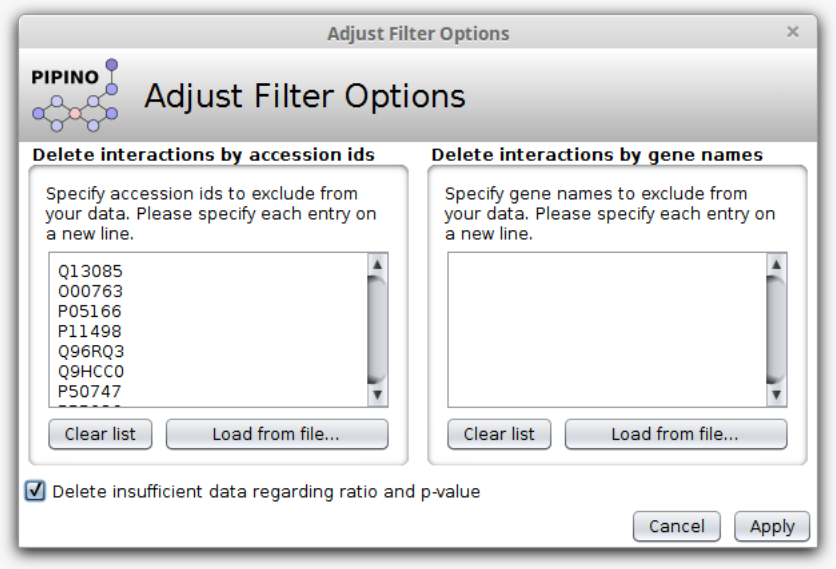
As soon as all required fields are extracted, the converted data model preview table should display calculated values depending on your preview data. Please be aware that these values may be incorrect due to the small preview size. You can now apply the data conversion to your whole data by pushing the *Apply* button or additionally save all parameters in a template file with the *Save template & Apply* button. Consecutive data imports from the same source file format can be accelerated by loading a previously saved template with the *Load template* button.

ATTENTION: If you change the document structure or update the data model, all your previous changes will be discarded.

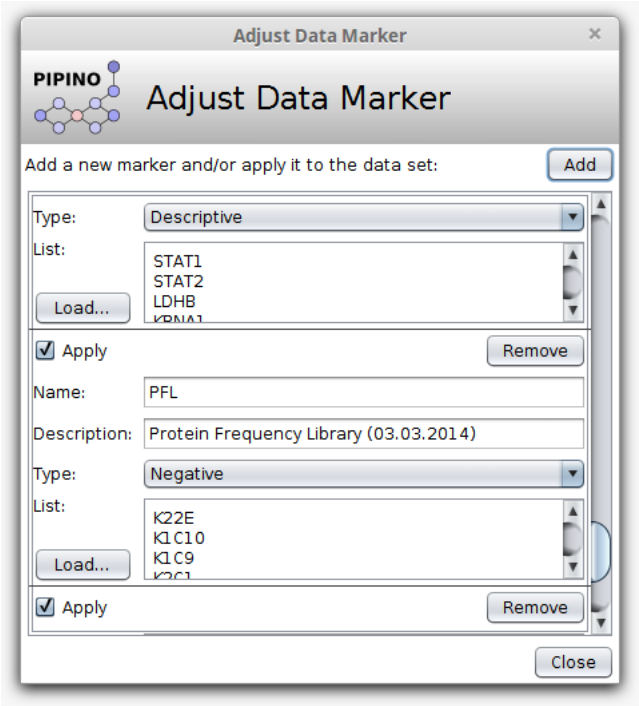
3.2 Post-processing



After a successful data conversion, the content table should be populated with the intermediate data. You can now inspect all parsed interactions or display only specific entries by applying a filter (this is a regex filter) with the appropriate textfield. Additional, data enhancements can be performed via the *Remove...* and *Highlight...* buttons.



You can remove unwanted entries (e.g. chaperone, proteolysis, biotinylation, ...) by specifying accession ids or gene names. Insufficient data entries regarding missing enrichment ratios or probability values can be deleted by selecting the appropriate checkbox.



Highlighting data entries and marking them with visual feedback can be a useful annotation. The data marker dialog provides the capabilities to add several markers. You can specify a name for identification purposes as well as an optional description to remember the purpose of this marker. There are three selectable types:

Descriptive	Neither positive nor negative, simply a marker
Negative	These entries may be considered as negative
Positive	These entries seem positive

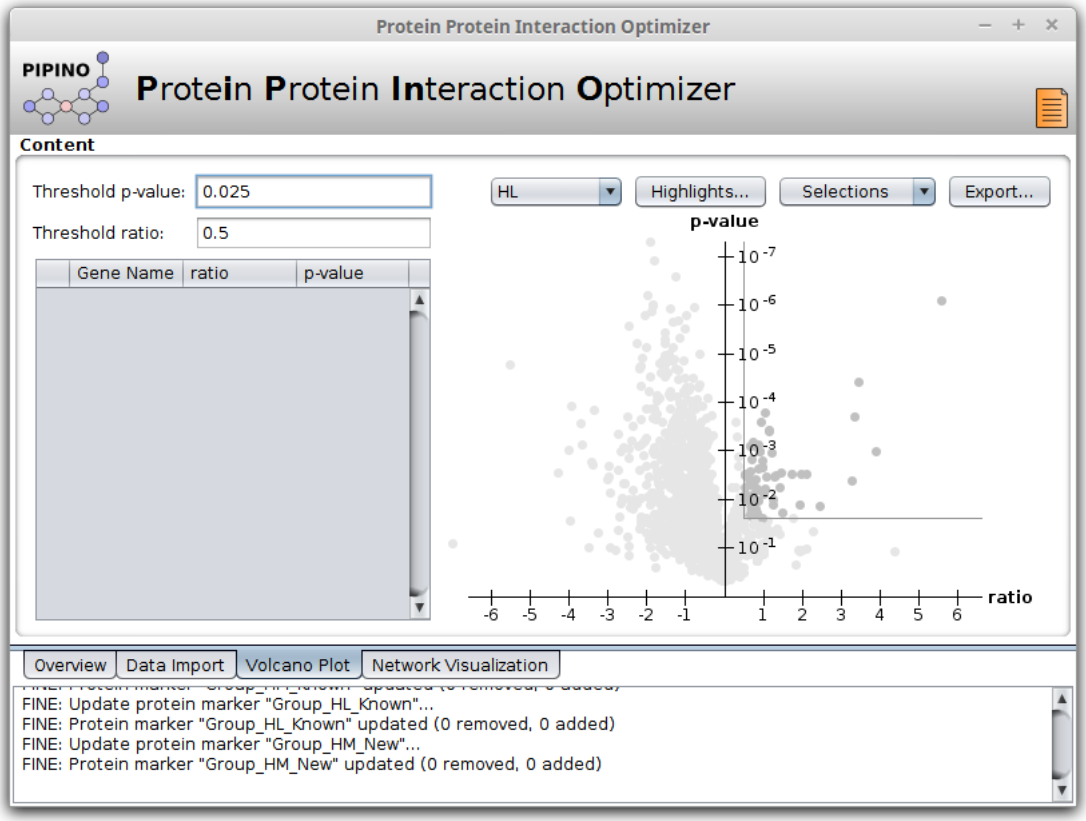
To add entries to a marker, insert either the gene name or the accession id into the text field. Multiple entries can be separated by comma or newline. To load a predefined list the appropriate *Load...* button can be pushed and any text file can be inserted.

Generated markers will be persisted within PIPINO and automatically applied to new imported data sets. If you want to apply only certain markers to your data, you can uncheck the *Apply* box for every marker or permanently remove them by pushing the *Remove* button.

Finally, you can export the processed data either as a tabstop separated file (*.tsv) for further external processing or as a serialized file (*.msd) which can be used as input for subsequent sessions (processed data).

4 The Volcano Plot

The *Volcano Plot* module displays the intermediate data in a scaled scatter plot. It presents the probability values in reversed order in combination with the logarithmic enrichment ratios. Therefore enriched data will tend to the right of the plot and low p-values will tend to the top of the plot. It is possible to display different experiments by selecting them from the dropdown menu above the plot.



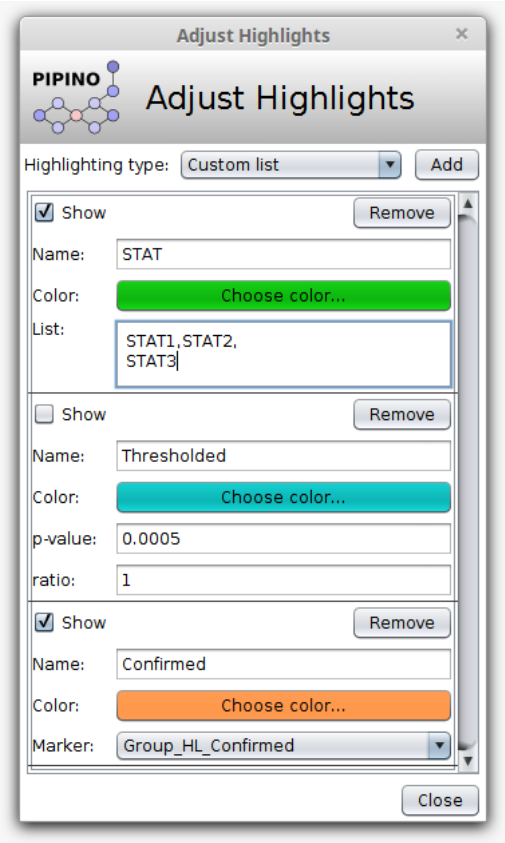
There is a threshold sector overlayed to inspect high interesting data entries in the top right corner of the plot. This sector can be adjusted by dragging the boundary lines or by manually inserting the desired values into the two threshold textfields in the top left corner of the module.

There are two methods to emphasize specific parts of the data. First, you can specify one or more highlight sets. These sets will be labeled and colored within the plot to visually separate the data. Currently there are 3 types of highlights possible:

Custom list	A custom gene name list separated by comma or newline
Thresholds	A area of interest bounded by threshold values
Marker	A marker group defined in the *Data Import* module

The highlight dialog can be triggered with the appropriate button above the plot. To hide a

highlight set simply uncheck the *Show* checkbox or to permanently remove it, push the *Remove* button.

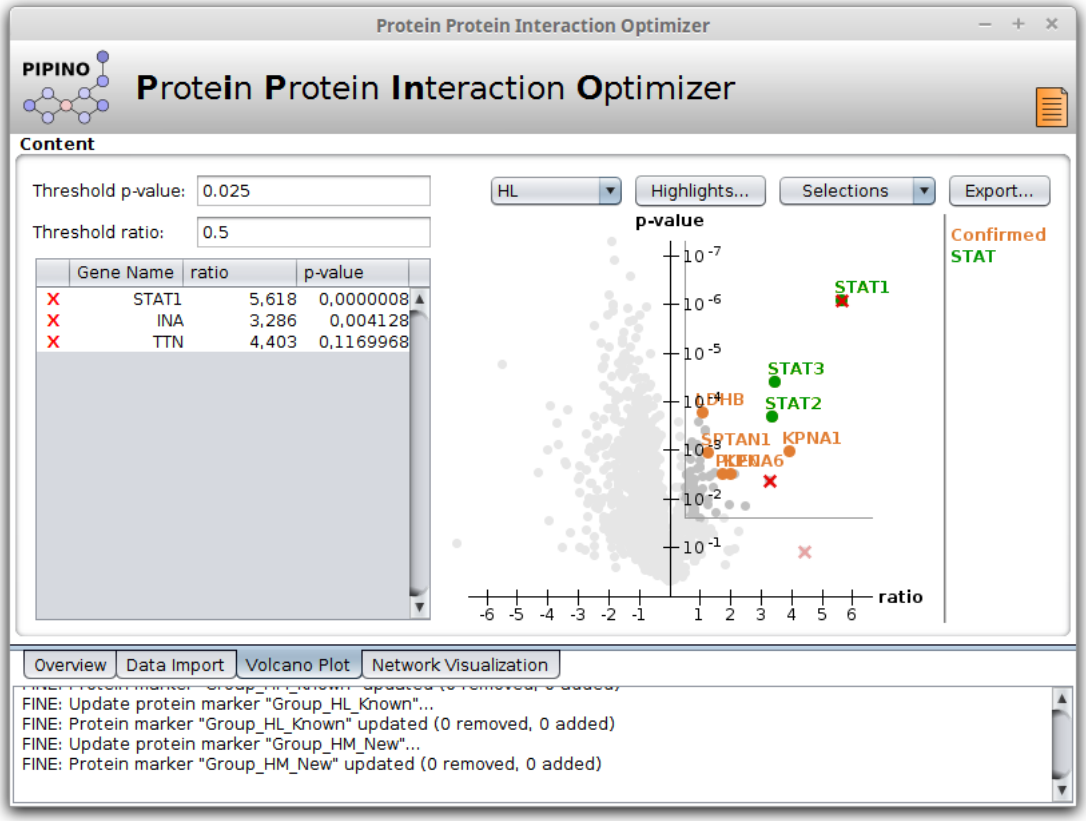


The second method to emphasize specific parts of the data is the data selection. Data entries can be selected by using the left mouse button or by choosing one of the option from the selection dropdown menu:

Clear selection	Resets all selected entries
Above thresholds	Selects all entries above the current thresholds
Highlighted	Selects all data entries currently highlighted
Highlighted above thresholds	Selects all highlighted data entries above the threshold

Selected data is overlayed with a red cross and the detail table on the left is used to display these entries with their values. You can deselect data by clicking on the red cross either in the table or in the plot.

The export option of the *Volcano Plot* module can be used to save the visualization as an image or to export selected data entries as text. The image export can be done in different formats (png, svg, tiff) and user defined dimensions. For large image dimensions you should consider adjusting the font size with the appropriate controls. The textual export can be formatted as plain gene name list or as complete export with all fields. The format of the complete export follows these guidelines to simplify parsing with other tools:



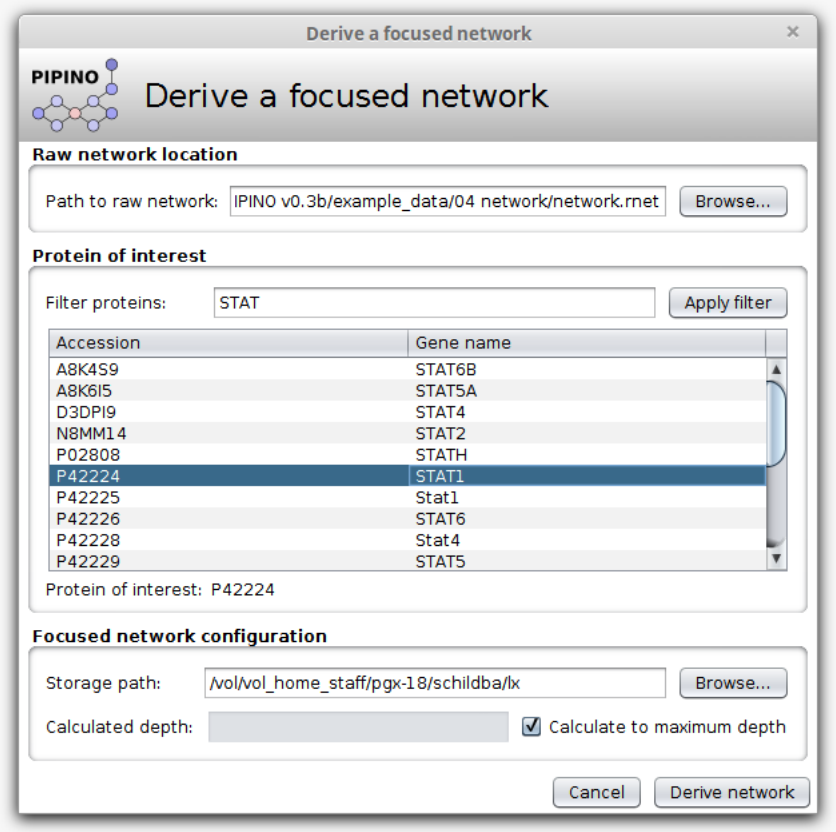

```
# Version
# Protein of interest (POI)
# Taxonomy of POI
# Amount of experiments
# Amount of markers
Marker descriptions with name and type (tabstop separated)
Table header (tabstop separated)
Data (tabstop separated)
```

5 Network Analysis

ATTENTION: The network analysis and visualization needs a lot of resources and may be slow depending on your hardware settings and provided data.

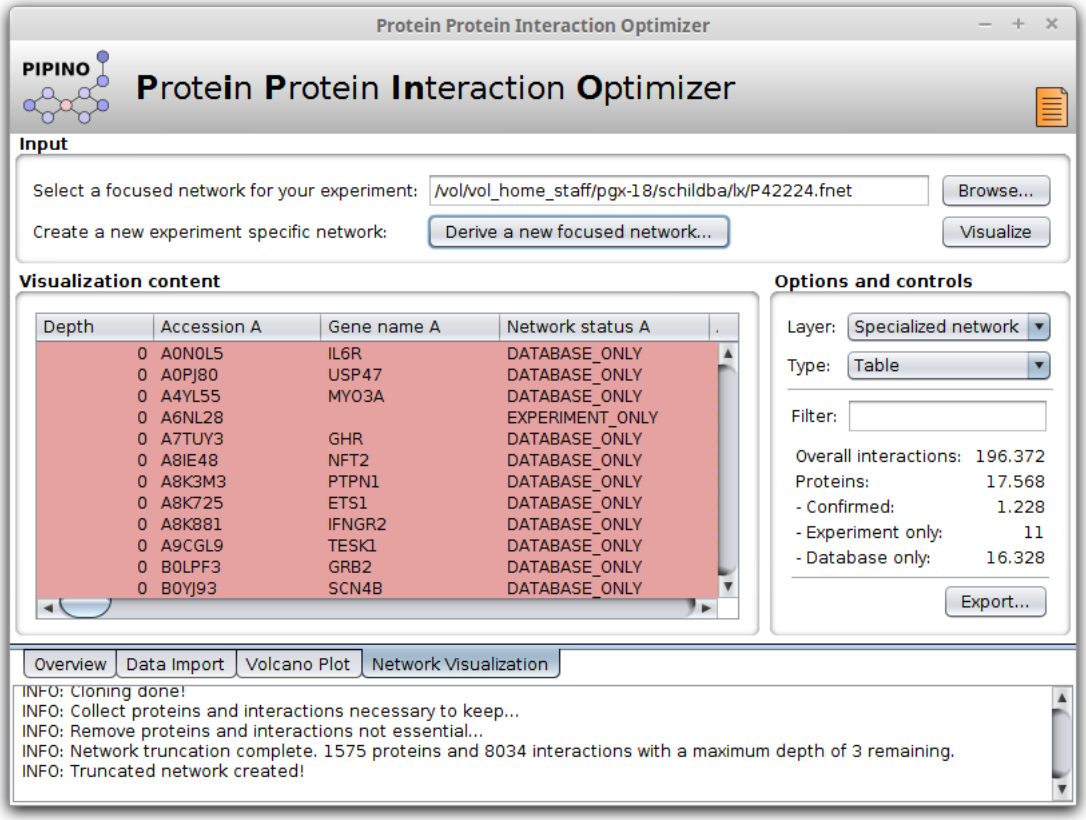
To start a network analysis on your data you can either use a previously focused network file by specifying an appropriate file and pushing the *Visualize* button or you can create a new focused network with the *Derive a new focused network...* button.

A focused network is a protein protein interaction network centered on a specific protein of interest and enriched by your experiment data. Therefore you need a complete network to operate on, which is provided as a prepared raw network file. This file is extracted from interaction databases and needs some computational efforts to be created – it is therefore not yet computable with the standalone version of PIPINO. Please be aware, that if you change, adjust or edit your experiment data, you need to create a new focused network.



After specifying the path to the raw network you need to select the protein of interest from the network. Due to different styles of writing a gene name, you need to select the right one for your experiment. You can apply a simple filter to narrow the search space. Select a protein of interest from the list and check if it is correctly assigned by verifying the label below the list. The last step for network creation is to specify a storage path and the calculation depth. It is

recommended to calculate to the maximum depth unless you only need a specific depth. The network creation can finally be started by using the *Derive network* button. This may take a while, depending on your hardware and provided data.



As soon as a focused network is loaded, the visualization content is updated and the table view is displayed by default. The table view displays all network interactions where the two interaction partners are labeled with *A* and *B*. The first column displays the network depth, where 0 means direct connection to the protein of interest and a larger number represents the amount of indirect interactions (e.g. depth of 2 means there are two proteins between the POI and the current protein). The fourth and seventh column represent the network status, which can be one of the following:

POI	The entry is the protein of interest
CONFIRMED	The entry is present in both database and experiment
EXPERIMENT_ONLY	The entry is only present in the experiment
DATABASE_ONLY	The entry is only present in the database

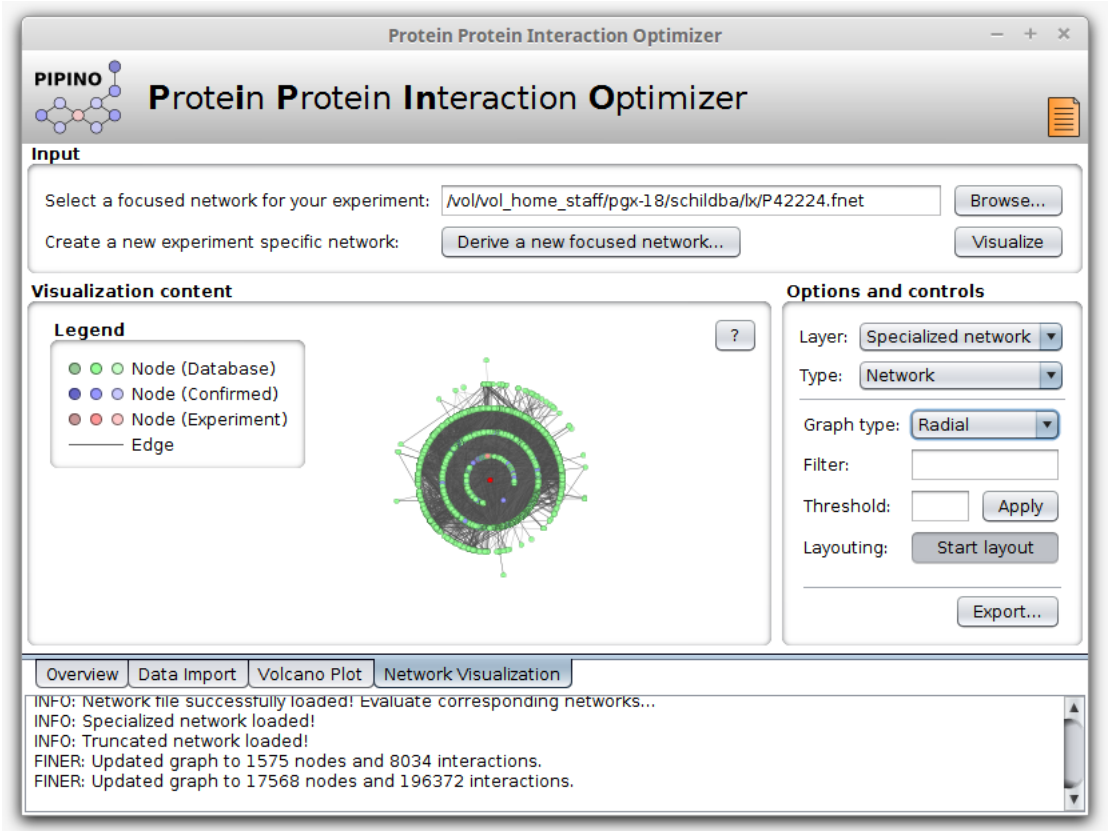
The following columns display the scoring values from the different interaction databases. While they are not directly comparable, they are all scaled between 0 (no interaction at all) and 1 (confident interaction). A missing value represents an unrated interaction. The column *Experiment* score* contains values for interactions only present in the experiment. The scores are followed by the experiment values and the data markers.

On the right hand side of the module there is the *Options and controls* area. To change the detail level of network observation you can change the network layer:

Focused network	The complete network focused around the protein of interest
Specialized network	The focused network enriched with experiment data
Truncated network	The specialized network with irrelevant leaf nodes removed

To switch the visualization content between table and network, you can select the appropriate entry from the type dropdown menu. The options below these two controls will change depending on the visualization content type.

For the table view, there is a case-sensitive filter to narrow the displayed table entries and some statistics of the displayed data. The export options offer a complete tabular data export of all columns and a data hub export. The latter export option sorts all proteins by the amount of interactions and returns a descending list containing the protein name, hub size and all interaction partners (tabstop separated).



The network view represents all data entries in an interactive network. This graph representation may cause heavy resource load for your system. The provided options can be used to change the graph type to one of the following:

Force field	A force field based network layout
Radial	A radial network layout
Hierarchical	A hierarchical network layout

Additionally the network can be highlighted with the case-sensitive filter, which will recolor corresponding nodes. To reduce the network size you can specify a score threshold. Only elements with a score (taking the highest available score) higher or equal to this threshold will be displayed. To stop the layouting process (only effecting the force field layout) you can trigger the *Stop layout* button. The displayed area can be exported as an image (png, svg, tiff) via the export option and scaled with the corresponding field.

The interactive network visualization content displays proteins as nodes and interactions as edges. The nodes are colored by their network status and the edges are colored by the highest available score (high (dark) to low (bright)). A quick guide how to interact with the network graph can be accessed by the question mark in the upper right corner:

Target	**User interaction**	**Graph reaction**
Background	Left click	Move the graph
Protein	Hover mouse	Highlight neighbors, Display tooltip, Display label
Protein	Left click	Select (hold)
Protein	Left click + Ctrl	Select (hold) multiple, Unselect selected
Protein	Left drag	Move node
Protein	Left drag + Ctrl	Move subtree
Everywhere	Right click	Zoom to fit
Everywhere	Mouse wheel	Zoom