Collector v1.5.0 setup and quickstart manual

Installation of collector

Collector is distributed as a tgz compressed file.

Uncompress it and follow the installation instructions in the README file.

Starting Collector web application server

Open a terminal window.

Move to the collector installation path. For example:

```
[collector@collector ~]$ cd /opt/collector
```

Start the web application by typing:

```
[collector@collector ~]$ startCollector
```

Now you will see on the terminal the Collector starting messages.

Stopping collector web application

Open a terminal window.

Move to the collector installation path. For example:

```
[collector@collector ~]$ cd /opt/collector
```

Stop the web application by typing:

```
[collector@collector ~]$ stopCollector
```

Troubleshooting starting web application

If you find difficulties starting the application you can try to stop previous collector instances that were already running:

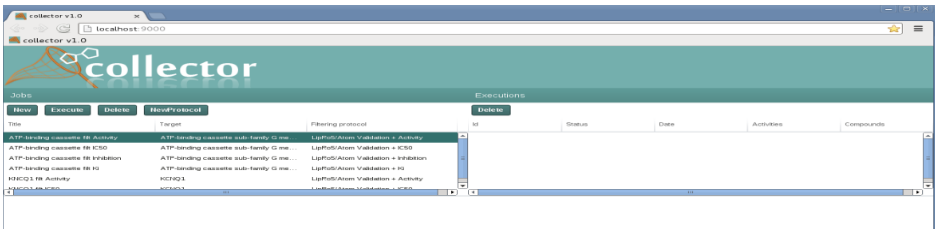
```
[collector@collector ~]$ stopCollector
[collector@collector ~]$ startCollector
```

Accessing Collector web application

Open a web browser and enter the web address:

http://localhost:9001

After loading the page you should see something like:



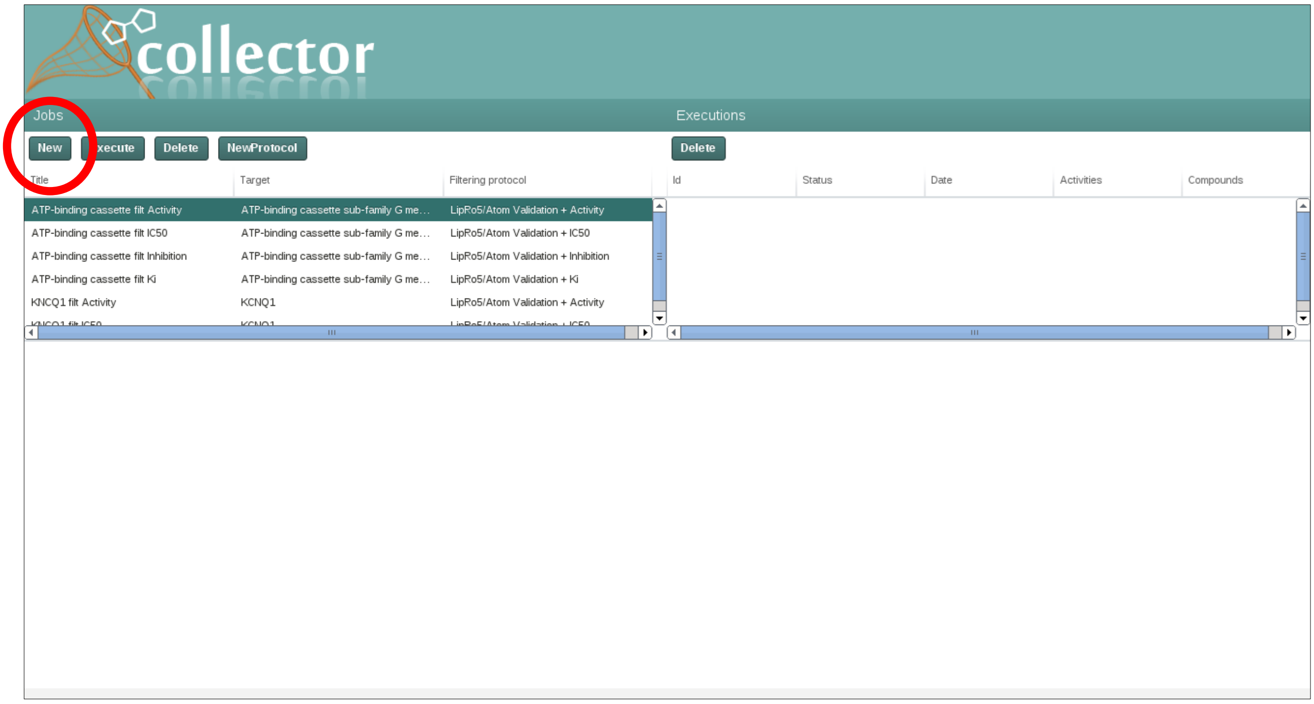
Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me...	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me...	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me...	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me...	LipRo5/Atom Validation + Ki
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity
KNCQ1 filt IC50	KCNQ1	LipRo5/Atom Validation + IC50

Id	Status	Date	Activities	Compounds

Collector web user functions

Defining a new job

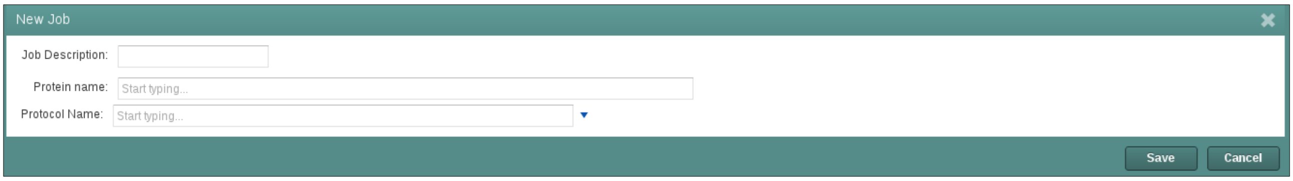
To define a new job push the button “New”:



Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Ki
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity
KNCQ1 filt IC50	KCNQ1	LipRo5/Atom Validation + IC50

Id	Status	Date	Activities	Compounds

The following form appears:



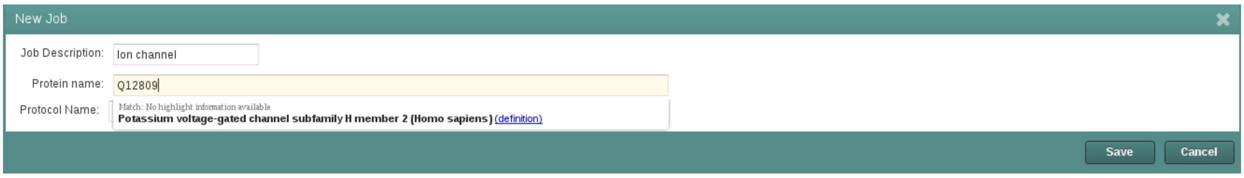
Job Description:

Protein name: Start typing...

Protocol Name: Start typing...

You have to complete:

- Job Description
- Protein Name: Enter the name of the target you we are interested in. You can use an UNIPROT accession number like Q12809 or a target name:
- The protocol to apply to the obtained data. You can select it from a list of predefined filtering protocols:



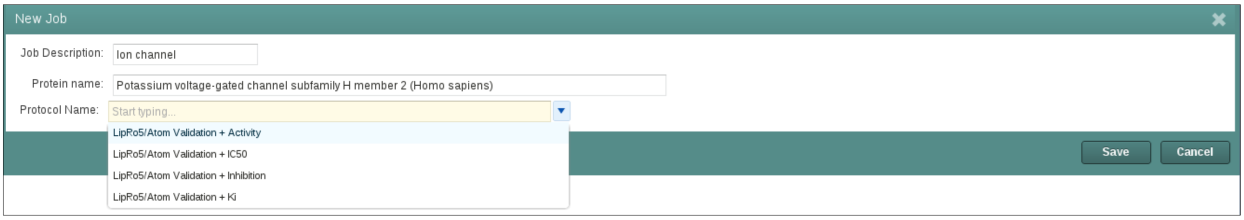
Job Description: Ion channel

Protein name: Q12809

Protocol Name:

Match: No highlight information available
Potassium voltage-gated channel subfamily H member 2 (Homo sapiens) (definition)

- LipRo5/Atom Validation + Activity



Job Description: Ion channel

Protein name: Potassium voltage-gated channel subfamily H member 2 (Homo sapiens)

Protocol Name: Start typing...

LipRo5/Atom Validation + Activity
LipRo5/Atom Validation + IC50
LipRo5/Atom Validation + Inhibition
LipRo5/Atom Validation + Ki

- LipRo5/Atom Validation + IC50
- LipRo5/Atom Validation + Inhibition
- LipRo5/Atom Validation + Ki

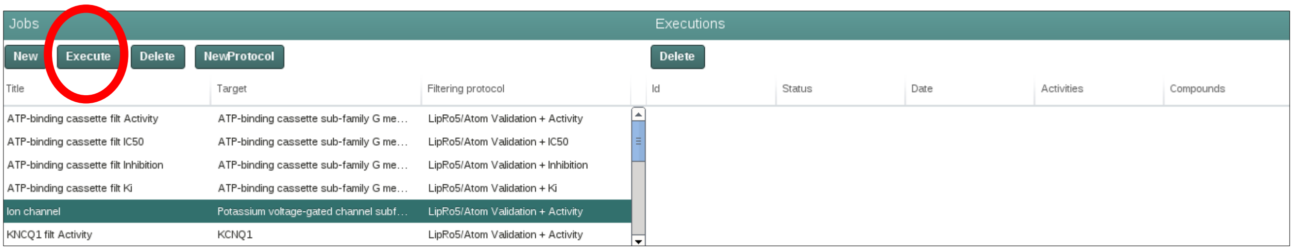
Each filtering protocol filters by a different activity type: Activity, IC50, Inhibition and Ki.

All these protocols apply two additional filters based on chemical properties:

- Lipinski Rule of 5: compounds that do not meet “Lipinski Rule of 5” are filtered out
- Atom validation: compound containing any atom different from H, C, N, O, S, P, Cl, I, Br, F are filtered out

Executing a job

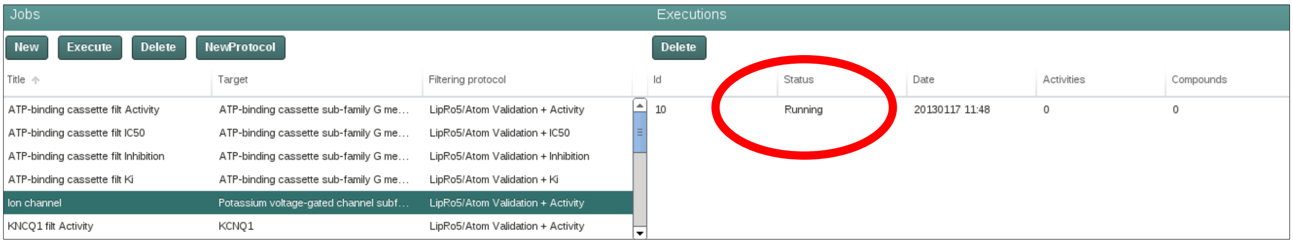
Once the job is defined you can execute it. You have to highlight the job in the “Jobs” panel and push the button “Execute”:



Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Ki
Ion channel	Potassium voltage-gated channel subf…	LipRo5/Atom Validation + Activity
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity

Id	Status	Date	Activities	Compounds

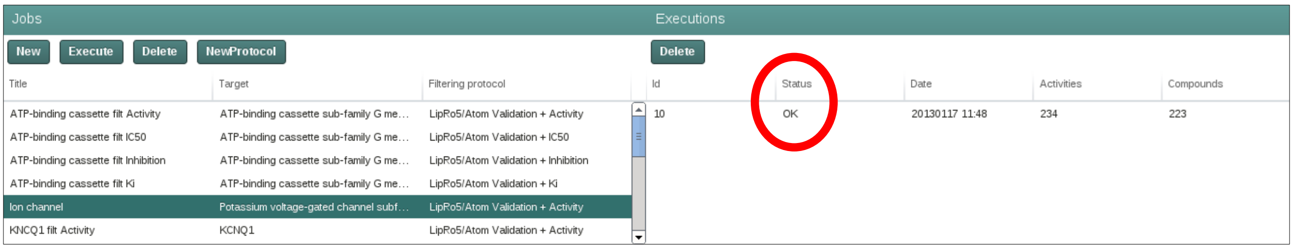
The job execution starts:



Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Ki
Ion channel	Potassium voltage-gated channel subf…	LipRo5/Atom Validation + Activity
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity

Id	Status	Date	Activities	Compounds
10	Running	20130117 11:48	0	0

After some time the execution finishes showing “Status” OK:



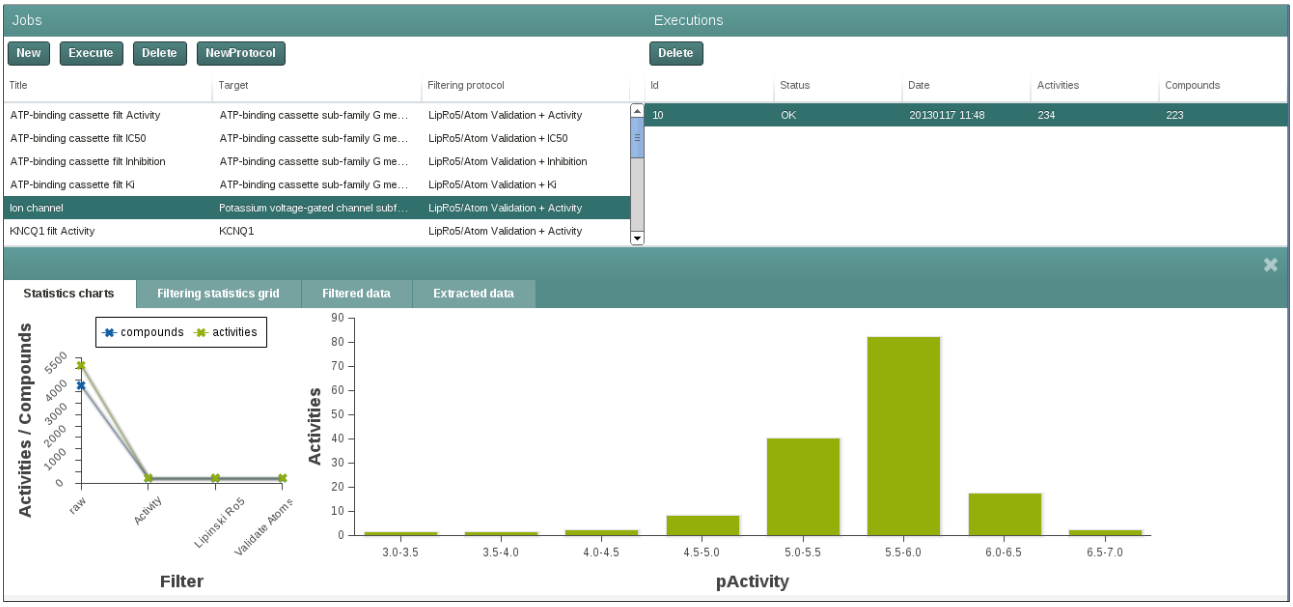
Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Ki
Ion channel	Potassium voltage-gated channel subf…	LipRo5/Atom Validation + Activity
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity

Id	Status	Date	Activities	Compounds
10	OK	20130117 11:48	234	223

Browse and export job execution data

Now you can browse the extracted data clicking on the panel “Executions”:

In the first tab you can see, on the left hand side, a line chart showing the number of compounds that passed the different filters of the protocol. On the right hand side you can see a histogram of the distribution of the parameter of activity obtained.

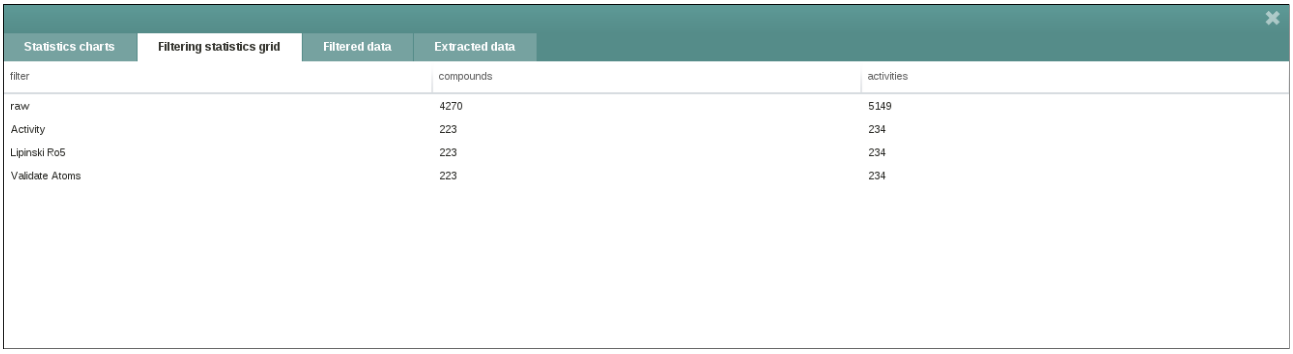


Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Ki
Ion channel	Potassium voltage-gated channel subf…	LipRo5/Atom Validation + Activity
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity

Id	Status	Date	Activities	Compounds
10	OK	20130117 11:48	234	223

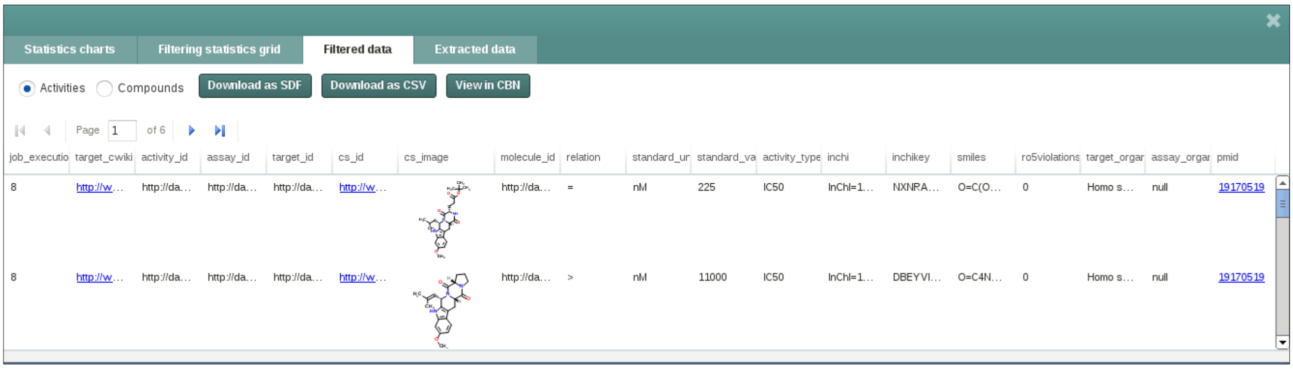
The second tab shows a table representing the same data of the line chart in tabular form:

Please note the different “compounds” and “activities” counts; compound indicates the number of different compounds obtained and activities the number of activity annotations extracted. Very often both figures are different, since the most databases contain several activity annotations for the same compounds.

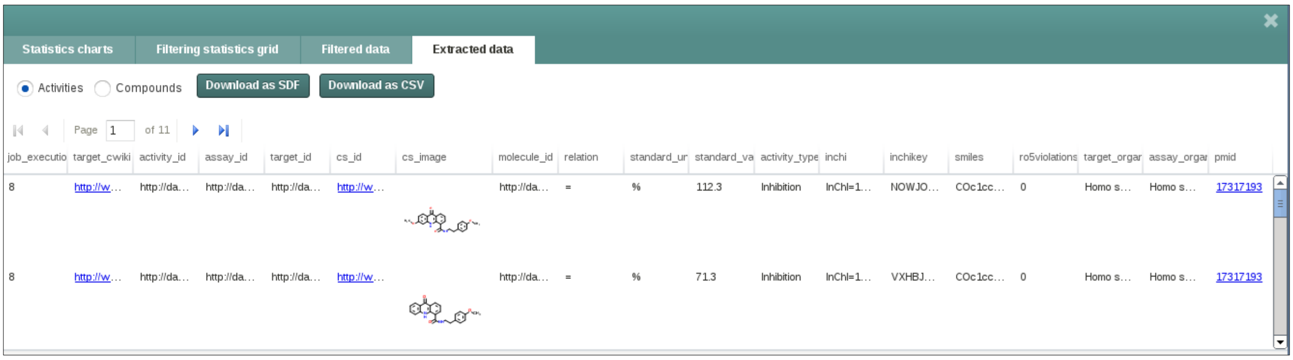


filter	compounds	activities
raw	4270	5149
Activity	223	234
Lipinski Ro5	223	234
Validate Atoms	223	234

In the third and fourth tab we can browse the raw data extracted or the filtered data:



job_executio	target_cwiki	activity_id	assay_id	target_id	cs_id	cs_image	molecule_id	relation	standard_ur	standard_va	activity_type	inchi	inchikey	smiles	ro5violations	target_orgar	assay_orgar	pmid
8	http://w…	http://da…	http://da…	http://da…	http://w…		http://da…	=	nM	225	IC50	InChI=1…	NXNRA…	O=C(O…	0	Homo s…	null	19170519
8	http://w…	http://da…	http://da…	http://da…	http://w…		http://da…	>	nM	11000	IC50	InChI=1…	DBEYVI…	O=C4N…	0	Homo s…	null	19170519



job_executio	target_cwiki	activity_id	assay_id	target_id	cs_id	cs_image	molecule_id	relation	standard_ur	standard_va	activity_type	inchi	inchikey	smiles	ro5violations	target_orgar	assay_orgar	pmid
8	http://w…	http://da…	http://da…	http://da…	http://w…		http://da…	=	%	112.3	Inhibition	InChI=1…	NOWJO…	COc1cc…	0	Homo s…	Homo s…	17317193
8	http://w…	http://da…	http://da…	http://da…	http://w…		http://da…	=	%	71.3	Inhibition	InChI=1…	VXHBJ…	COc1cc…	0	Homo s…	Homo s…	17317193

You can decide what to export by pressing Activities/Compounds radio button:

In the case of compounds, when we have different measurements for the same compound, Collector computes the median of the different values

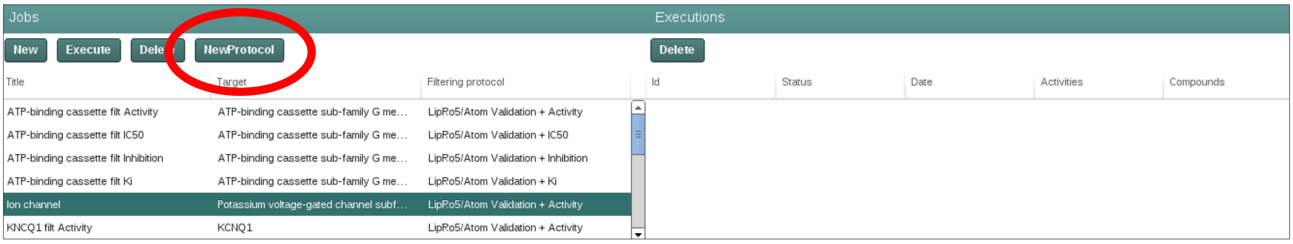
For exporting the results you can press the buttons:

- Download as CSV: if you want the data in plain text tab-separated format
- Download as SDF: in 2D SDF format

Define new Protocols

You can define a new protocol by pushing the NewProtocol button:

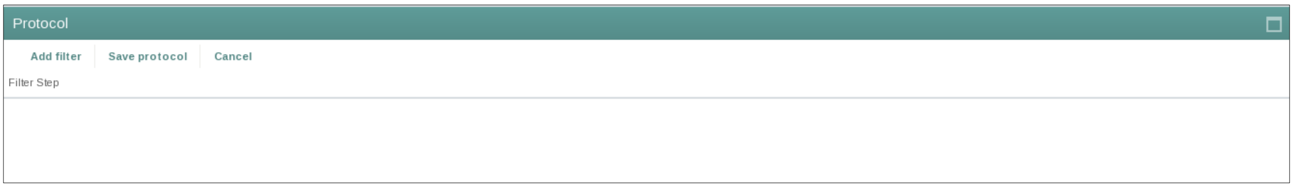
You can see a new form:



Title	Target	Filtering protocol
ATP-binding cassette filt Activity	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Activity
ATP-binding cassette filt IC50	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + IC50
ATP-binding cassette filt Inhibition	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Inhibition
ATP-binding cassette filt Ki	ATP-binding cassette sub-family G me…	LipRo5/Atom Validation + Ki
Ion channel	Potassium voltage-gated channel subf…	LipRo5/Atom Validation + Activity
KNCQ1 filt Activity	KCNQ1	LipRo5/Atom Validation + Activity

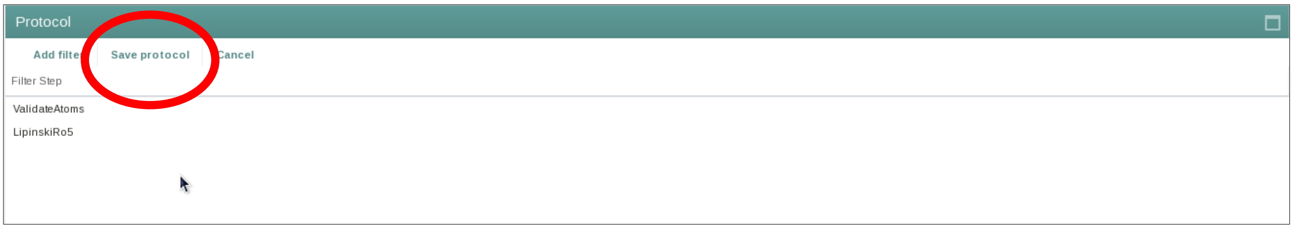
Id	Status	Date	Activities	Compounds

This form allows creating a new protocol by adding filters to the protocol. For example we can add a new filter:



Filter Step

When the user finishes adding the filters to the protocol, you can save it by pushing the “Save protocol” button:

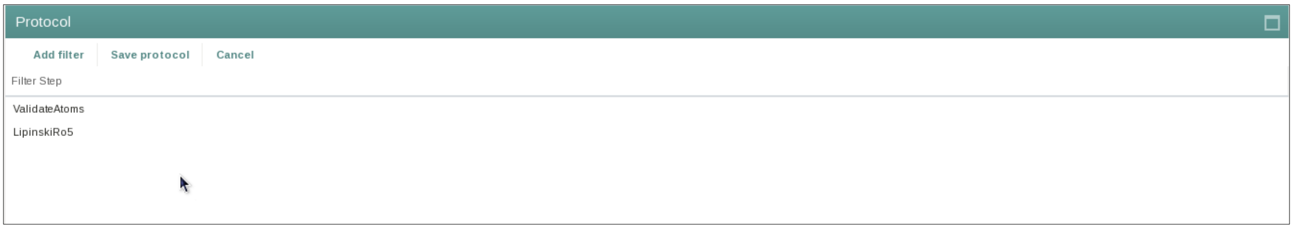


Filter Step

ValidateAtoms

LipinskiRo5

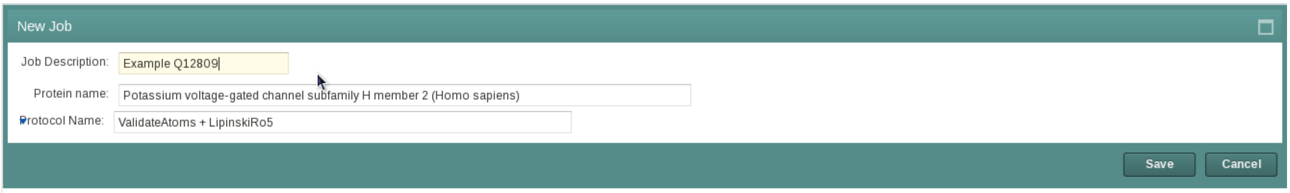
Once the protocol is saved you can define a new job that uses the newly defined protocol:



Filter Step

ValidateAtoms

LipinskiRo5



Job Description: Example Q12809

Protein name: Potassium voltage-gated channel subfamily H member 2 (Homo sapiens)

Protocol Name: ValidateAtoms + LipinskiRo5

Collector Command line

Collector has a full command line interface.

To access it you have to open a terminal. Type “collector” to get a brief description of the commands available:

```
[collector@collector ~]$ collector
```

Collector command line commands:

```
collector listprotocols
```

- Lists the protocols available in the system. In the current implementation:

job_filtering_id	job_filtering_description	filter_description	curation_order
1	No filtering	NoFiltering	1
2	LipRo5/Atom Validation + Activity	Activity	1
2	LipRo5/Atom Validation + Activity	LipinskiRo5	2
2	LipRo5/Atom Validation + Activity	ValidateAtoms	3
3	LipRo5/Atom Validation + IC50	IC50	1
3	LipRo5/Atom Validation + IC50	LipinskiRo5	2
3	LipRo5/Atom Validation + IC50	ValidateAtoms	3
4	LipRo5/Atom Validation + Inhibition	Inhibition	1
4	LipRo5/Atom Validation + Inhibition	LipinskiRo5	2
4	LipRo5/Atom Validation + Inhibition	ValidateAtoms	3
5	LipRo5/Atom Validation + Ki	Ki	1
5	LipRo5/Atom Validation + Ki	LipinskiRo5	2
5	LipRo5/Atom Validation + Ki	ValidateAtoms	3

Every filtering protocol is identified by job_filtering_id. The curation_order defines the order by which filters are applied. For example, the protocol 2 applies sequentially the filters

- Activity
- LipinskiRo5
- ValidateAtoms

The “No filtering” is protocol that applies no filtering; it only extracts the raw data from the source.

```
collector newjobuniprotid --protocolid <protocol_id> --uniprotid
<uniprotaccession> --jobdescription <job_description>
```

Defines a new job based on:

- `<protocol_id>:` the protocol to apply (obtained in listprotocols call)
- the `<uniprotaccession>` of the target
- `<job_description>: a descriptive text of the job`

```
collector executejob –jobdescription <job_description>
```

Executes the job with `<job_description>`

```
collector listjobexecutions --jobid <job_id>
```

Lists all the executions of the given job_id

```
collector export --raw --jobexecutionid <job_execution_id> --datatoexport
activities|compounds –exportformat sdf|csv –filename <filename>
```

Exports the data obtained in the job_execution_id depending on:

- activities: extracts the detailed activity data
- compounds: extracts the detailed activity data aggregated at compound level (median of all the activities reported for the same compound)
- sdf or csv: SDF 2D format or plain text tab-separated plain text file.
- <filename>: Filename to export.