Hands-on Workshop on Computational Nanotoxicology using eNanoMapper web services by NTUA

Workshop:	NMP project eNanoMapper
DATE / PLACE**:**	March 17, 2016 / National Technical University of Athens, Zografou Campus:
TIME:	13.30

TITLE:	Developing Nano-QSAR predictive toxicology models

SPEAKER:	Philip Doganis and Haralambos Sarimveis
AUTHORS:	Philip Doganis, Haralambos Sarimveis *School of Chemical Engineering, National Technical University of Athens*

ABSTRACT

The workshop will offer hands-on work on the development of nanoQSAR models based on data available from the data.enanomapper.net server, making use of the eNanoMapper computational infrastructure from NTUA[1] that extends the OpenTox API[2]. The focus is on the use case of predicting cell association of Gold Nanoparticles, based on experimental data, publically available in the publication of Walkey et. al.[3].

Participants will go through the workflow of constructing a dataset from a selected one from the eNanoMapper database, performing preparatory operations on it, such as selection of a subset of properties or calculated descriptors and examine using diverse sets of nanoparticle descriptors, including descriptors derived from protein-corona information. Three model building cases will be explored towards the creation of predictive models using statistical and machine learning algorithms, firstly on physicochemical data, secondly on proteomics data and finally a case where the use of PMML (Predictive Model Markup Language) in XML-based files for definitions of descriptor transformations and model output will be presented based on physicochemical data.

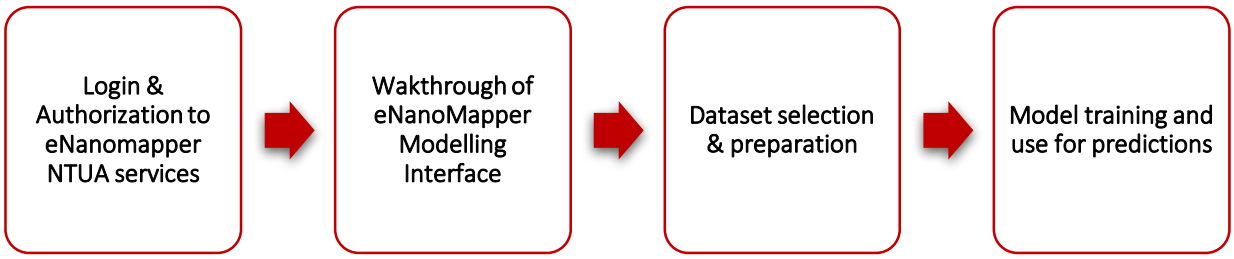


Figure 1 Workshop Outline

1. Login & Authorization to the Jaqpot and eNanoMapper services

Users access the Jaqpot Quattro homepage by NTUA http://app.jaqpot.org:8000/ (Figure 2), currently as a test instance as is under development. To gain access to the services, users should click the **Sign in** button, to be transferred to the screen in Figure 3.

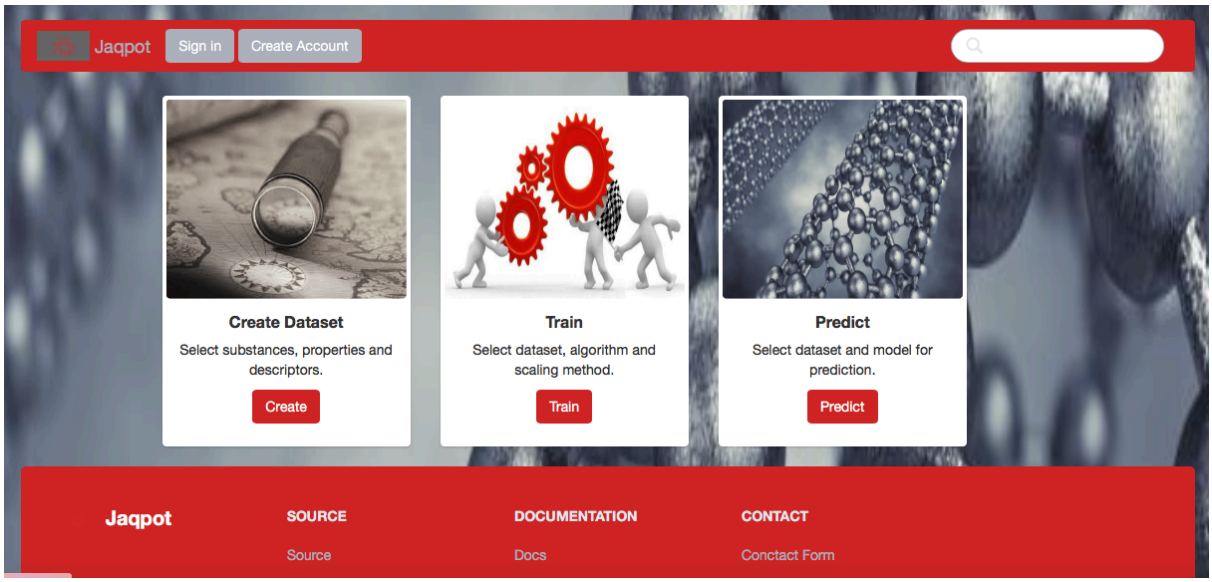


Figure 2 Homepage of eNanoMapper NTUA Modelling interface

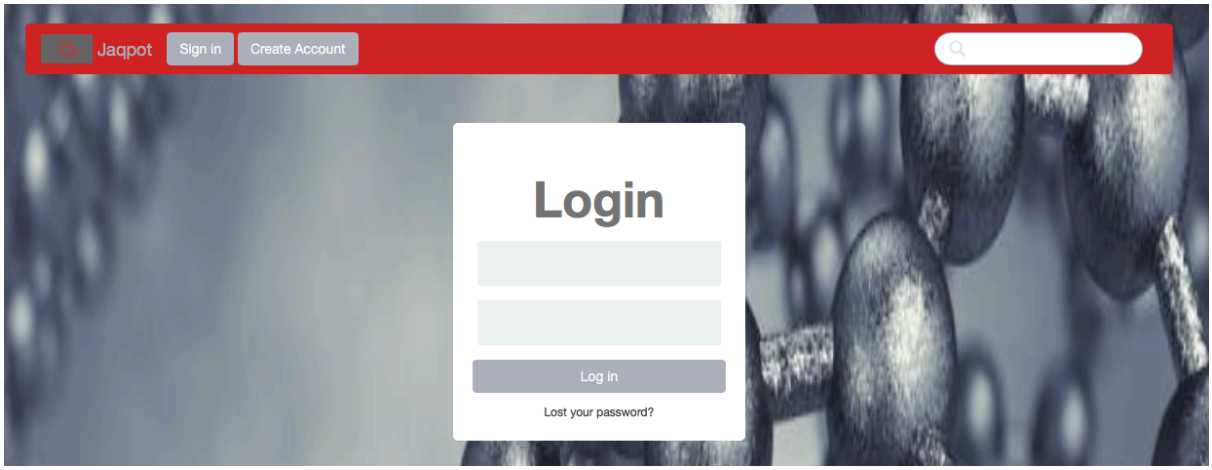


Figure 3 Login page

Log in to Jaqpot Quattro using OpenTox credentials. You can get an OpenTox user account by clicking on "Create Account" on the start page (Figure 2) at: http://opentox.org/join_form (Figure 4), login with your own credentials or just type "**guest**" in both fields. Successful entry of credentials leads to the homepage for registered users (Figure 5).

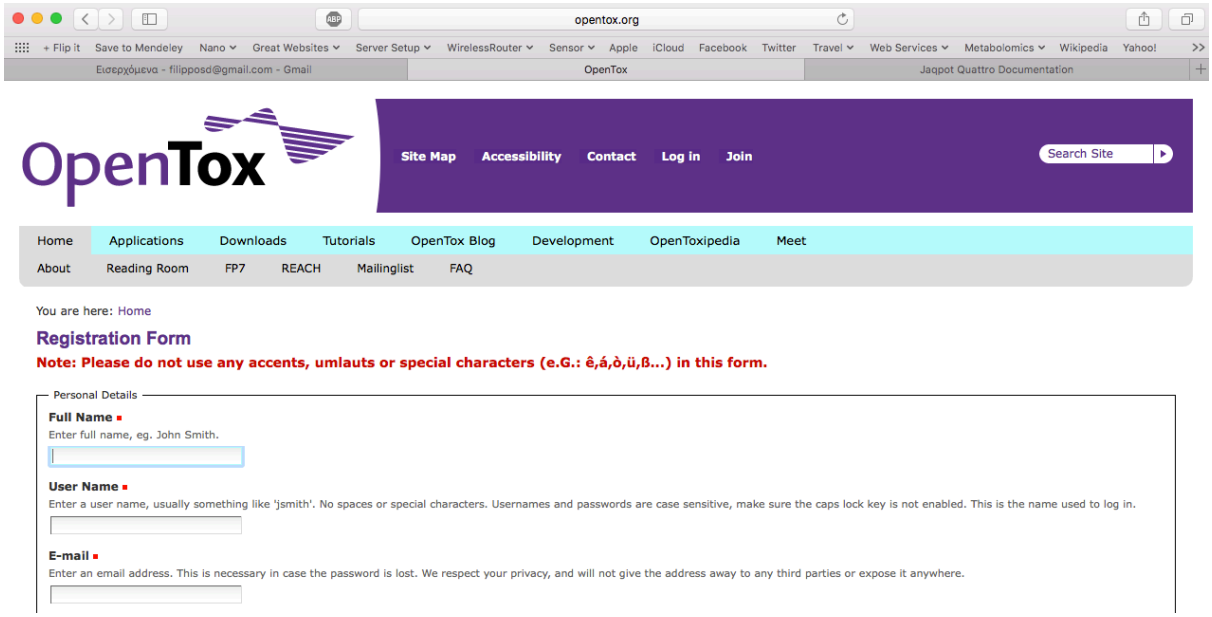


Figure 4 OpenTox Registration page (http://www.opentox.org/join_form)

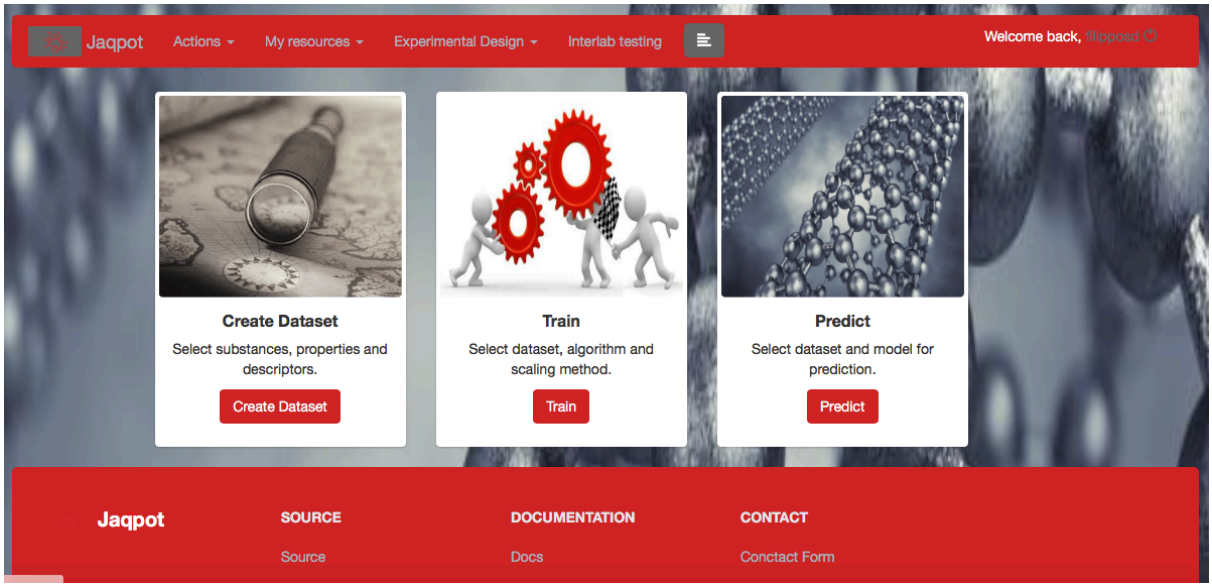


Figure 5 Starting page for registered users

2. eNanoMapper Jaqpot modelling services – A walkthrough

In this section we will have a brief walkthrough of the eNanoMapper Jaqpot modelling interface. Please note that this a test User interface for the Modelling API (Application programming interface), which can be examined by more proficient users who would like to use the Web Services exposed through it in order to achieve more automated modelling workflow. Users can visit http://app.jaqpot.org:8080/jaqpot/swagger/ to view the documentation of the API and experiment with its possibilities using the Swagger interface hosted there (Figure 6). As the eNanoMapper Jaqpot

modelling interface leaves the testing phase and reaches finalization, a manual for its functionality will be released.

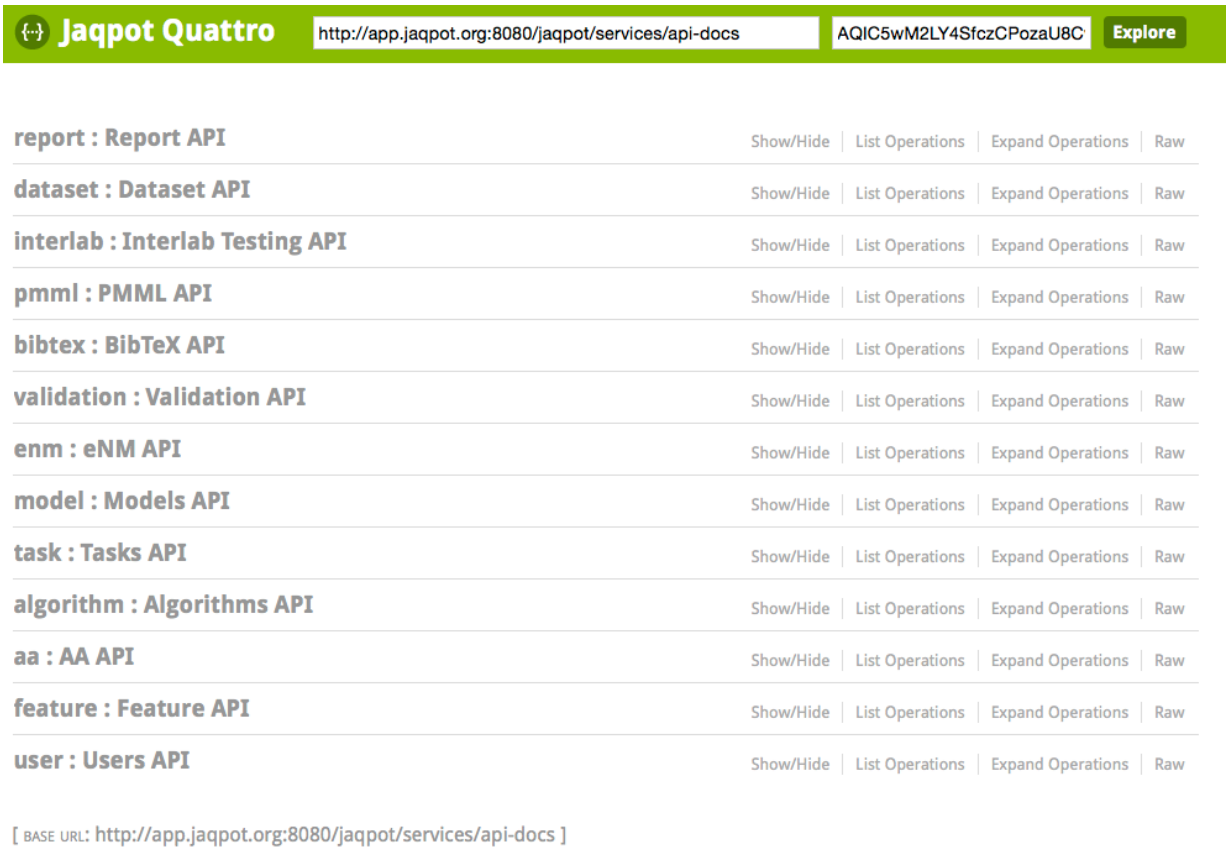


Figure 6 *eNanoMapper Modelling API:* Screenshot of Swagger interface for API documentation

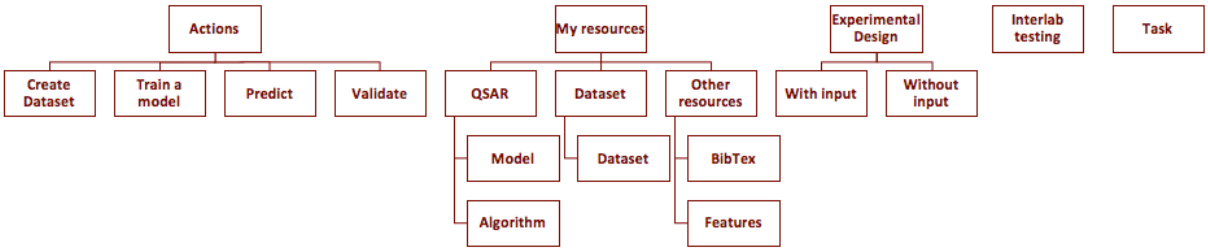


Figure 8 Jaqpot menu tree

The menu tree of the Jaqpot modelling interface is shown in Figure 87. Only the items of interest to this workshop will be described; please also note that this is test release and tools are under development.

The modelling Actions are (see Figure 78):

- **Dataset Creation** from an existing dataset in the data.enanomapper.net database, allowing selection a subset of substances or properties and execution of descriptor calculations when image files or

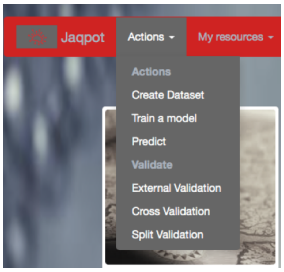


Figure 7 Available modelling actions

Crystallographic data are included in the dataset, resulting to the creation of a local dataset.

- **Model training** using a dataset in the data.enanomapper.net database or the local dataset and applying one of the available algorithms, currently:
  - Regression
    - *ocpu-lm*: Linear Regression by R offered through OpenCPU
    - *weka-mlr*: Linear Regression by Weka
    - *weka-pls*: Partial Least Squares by Weka
    - *weka-svm*: Support Vector Machines by Weka
    - *python-pls-vip*: Partial Least Squares with "Variable Importance in Projection" scores
  - Classification
    - *weka-pls*: Partial Least Squares Classification by Weka
    - *python-id3-mci*: ID3 classification algorithm

  Users can export their model in PMML format (explained in following chapter) and distribute it online using the unique URI assigned to it for easy deployment to other systems/web services. More detailed descriptions on the algorithms, their implementation and their parameters can be found in the openly available Deliverable 4.3 of eNanoMapper.
- **Predict** by applying a model to a dataset with matching properties and receive a dataset with predictions and Domain of Applicability calculations, if that choice has been made.

The resources that are available to the user are:

- **QSAR**
  - **Model**: a repository of models created by the user and public models
  - **Algorithm**: a list of the available algorithms offered by Jaqpot
- **Dataset**
  - A repository of datasets created by the user and public models

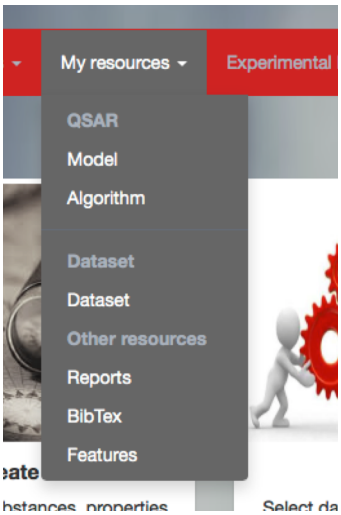


Figure 9 "My resources" menu

3. Dataset selection & preparation

Three cases will be examined, all involving the data from Walkey et.al.[3]. The **first** case involves a dataset with only Physicochemical data and two toxicity related endpoints, Net Cell Association and its log2 value (only one will be selected for use in the next). The **second** case involves a dataset with only Proteomics data and a toxicity related endpoint. The **third** case involves using PMML input files for transformations and determination of input variables.

Case 1

In the first case, a dataset with 84 Gold nanoparticles will be used after having been split into two sets, a Training and a Prediction set with 56 and 28 nanoparticles respectively. Among the available properties, only the Physicochemical data will be selected, thus leaving out the Proteomics experimental data, as shown in Figure 10.

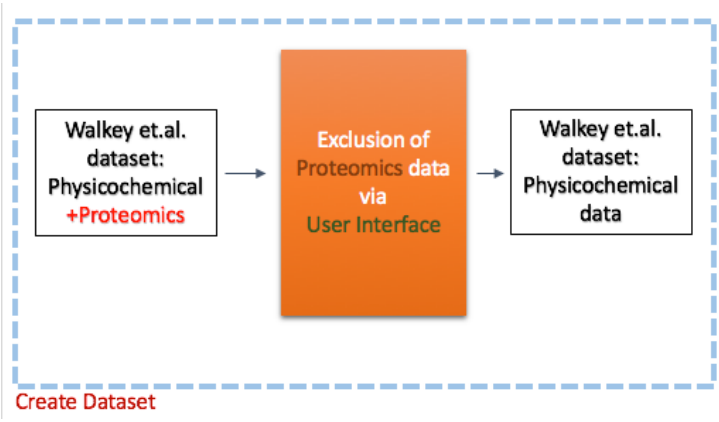


Figure 10 Create Dataset actions for Case 1

After clicking the "Create Dataset" option in the start screen (Figure 5) a view of datasets available on the https://apps.ideaconsult.net/enmtest/ui database is shown. Please select the ***WalkeyGoldTrain.csv*** file and hit Next, as in Figure 11.

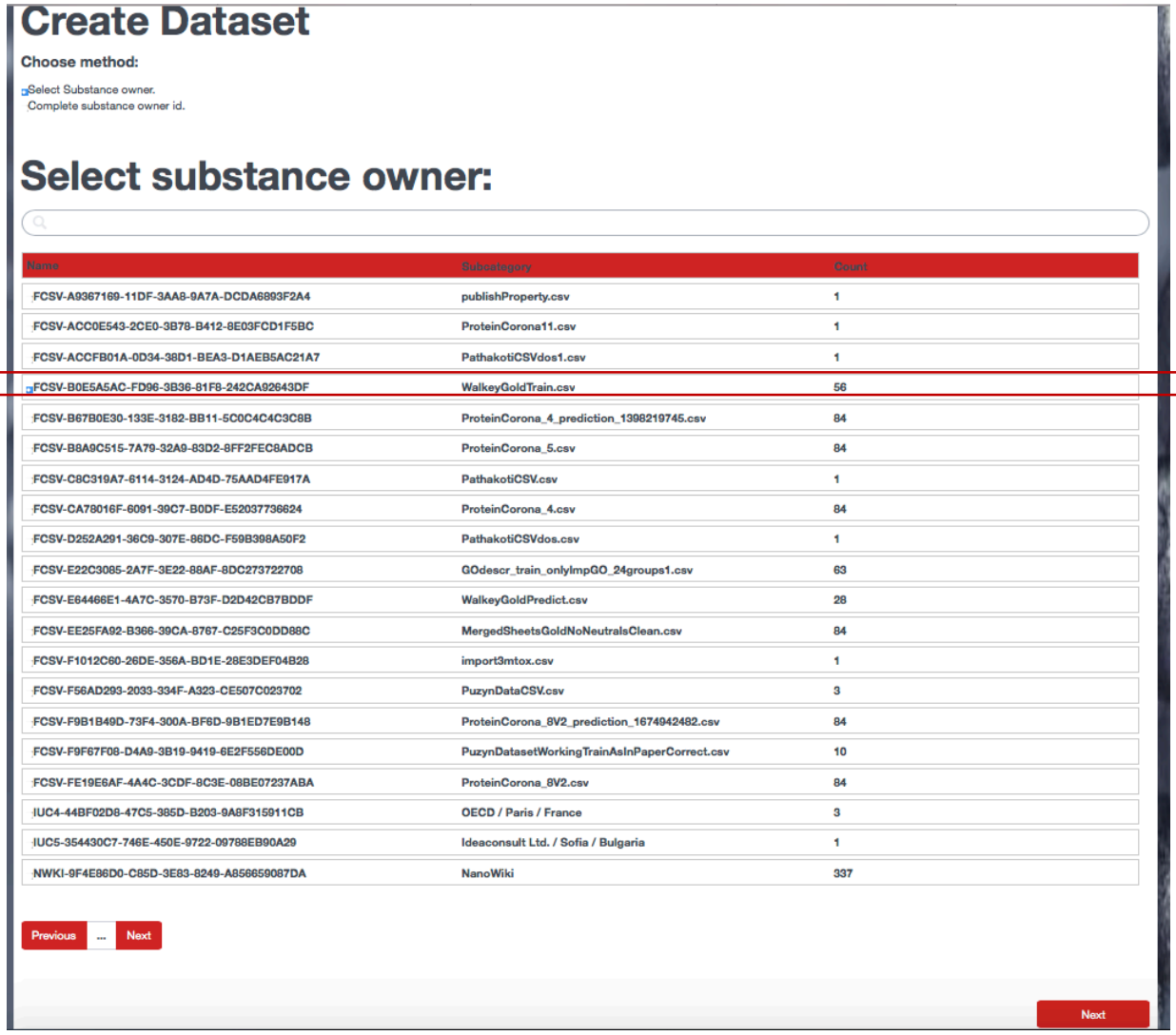


Figure 11 Create dataset: Selection of substance owner

In the **Substances** screen (Figure 12) users can select a subset of the available substances to be included in the dataset. Here all the substances should be selected.

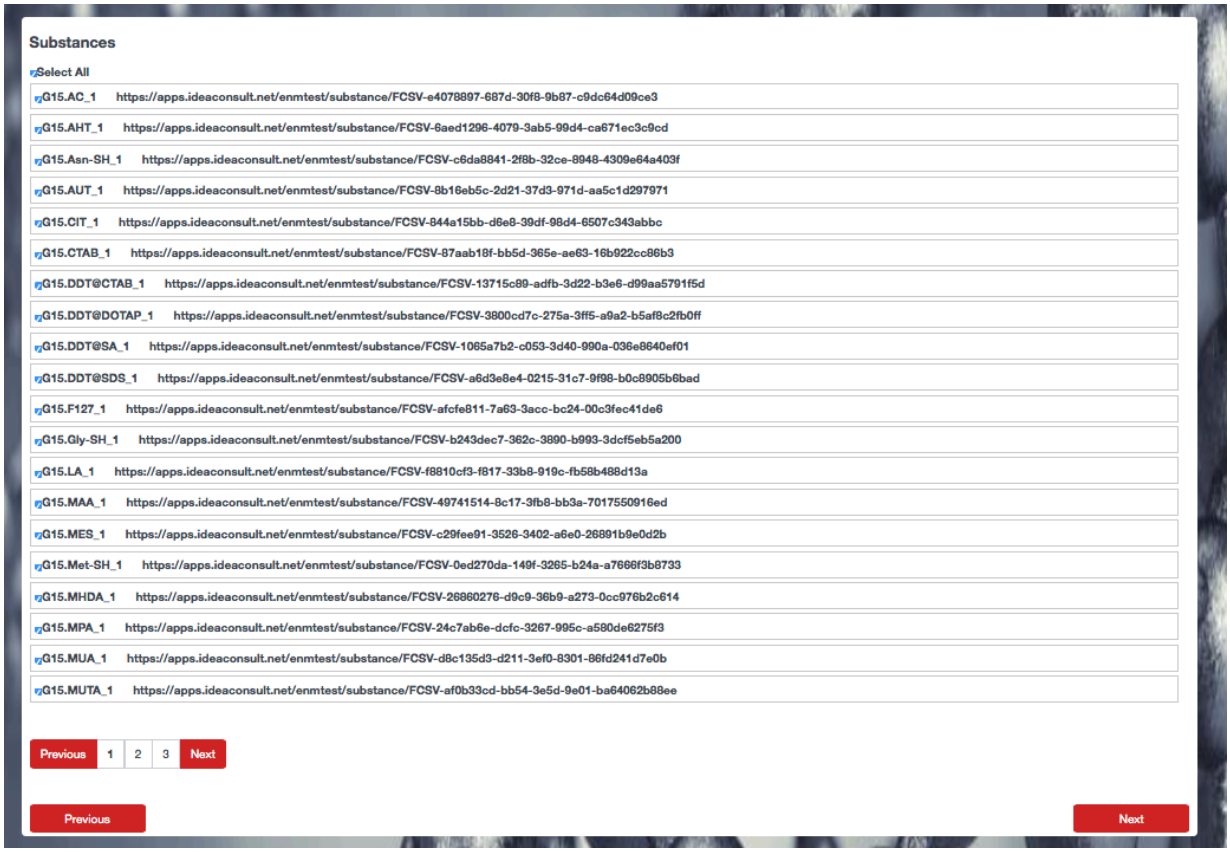


Figure 12 Create dataset: Substance select page

In the next screen (Figure 13), users may select the experimental properties to be included in the dataset, which have been grouped into 4 categories, Physicochemical, Environmental Fate, Toxicity and Ecotoxicity.

Figure 13 Create dataset: Selection of experimental properties to be included in the dataset

In the case studied, the Proteomics section in the Toxicity Category should be unchecked as in Figure 14, as the scope of the case is to study the Physicochemical section.

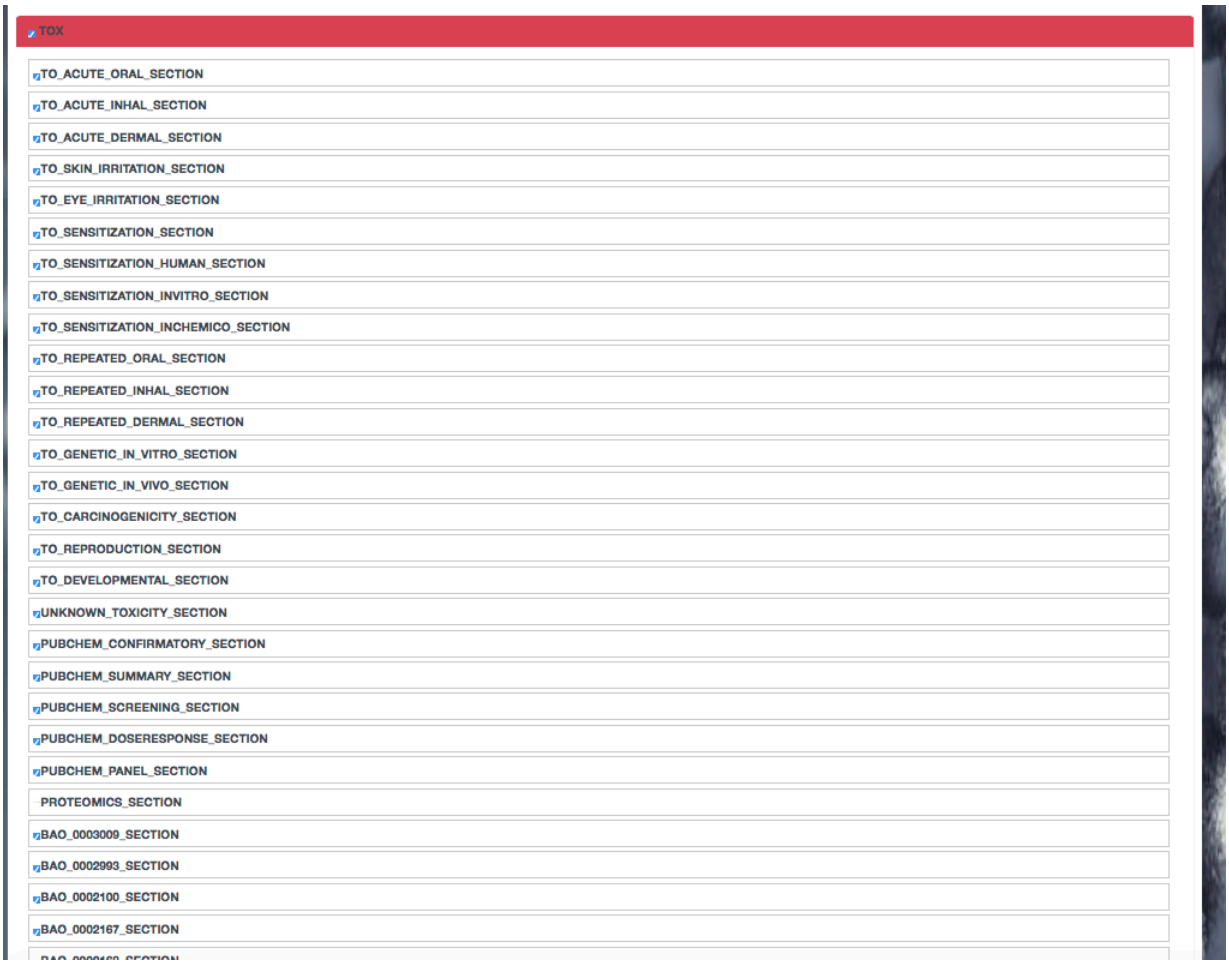


Figure 14 Create dataset: Some of the Toxicity properties available for selection. Please notice that the Proteomics Section is unchecked.

In the next stage (Figure 15), users can select the descriptors to be transferred, in the case of (pre-existing) Experimental values or to be calculated, in case the dataset contains nanoparticle images, Crystallographic data, Proteomics data and descriptors based on CDK or algorithms can be calculated. Additionally, the user must select a representative title and a description for the dataset.

Clicking the Next button at this stage leads to a notification page that the task has been initiated and an intermediate page showing the status of the task until completion (both pages shown in Figure 16).

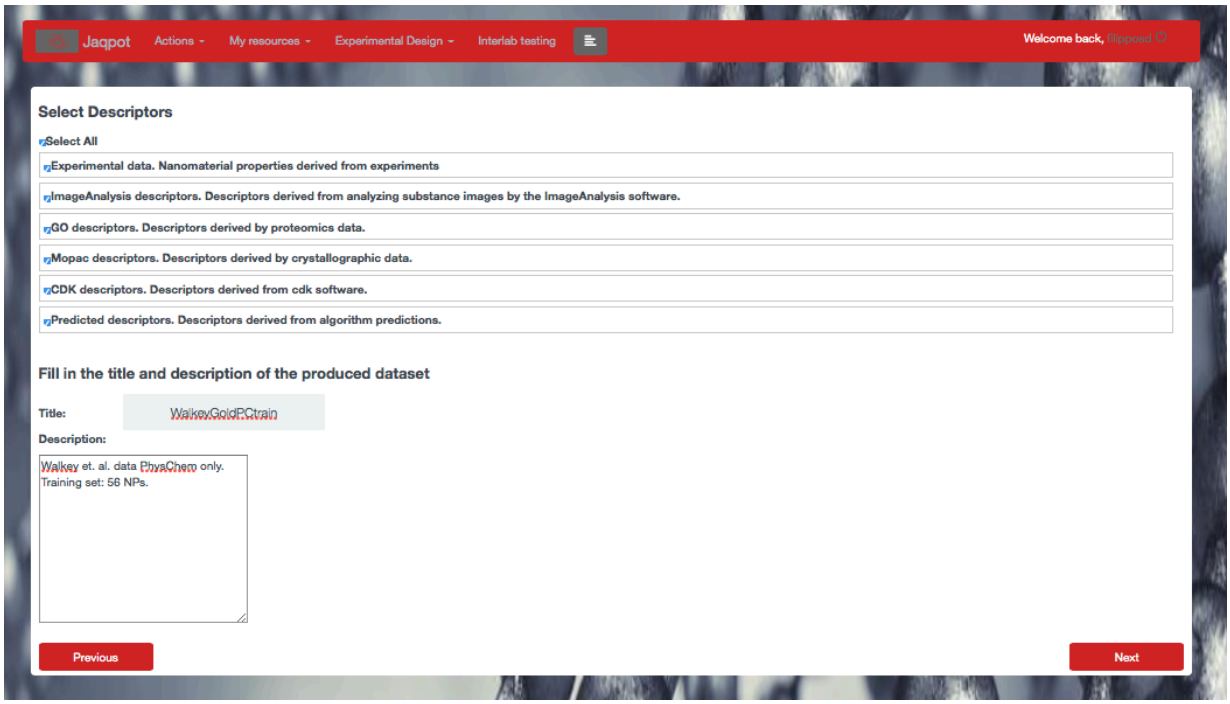


Figure 15 Create dataset: Selection of descriptors to be transferred or calculated, assignment of title and description

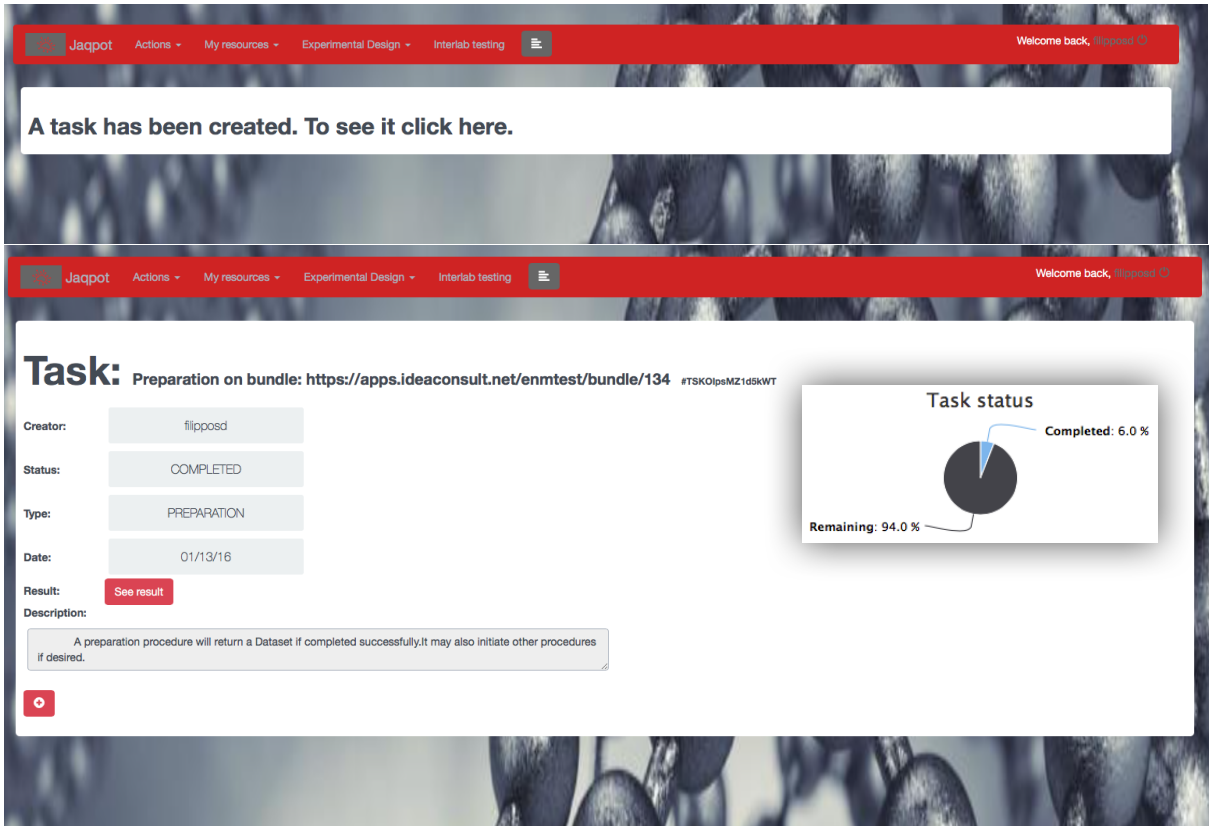
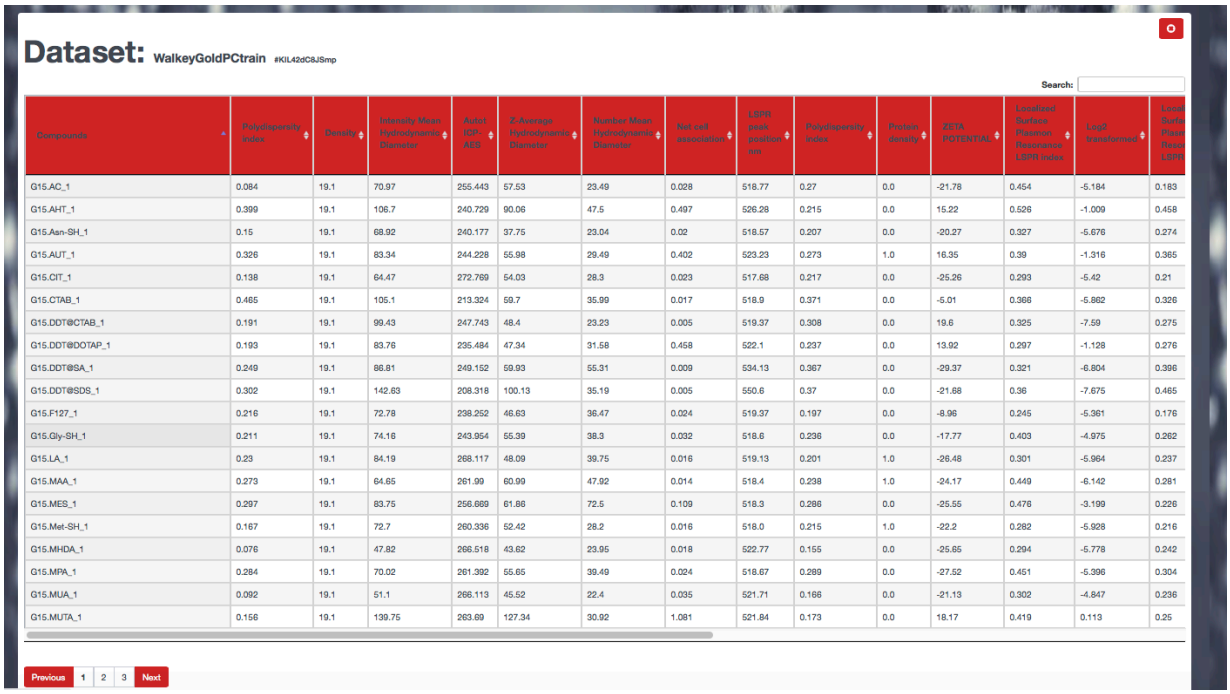


Figure 16 Intermediate task screens: *Above*: notification of task initiation. *Below*: Task progress page. While task is running, the status pie chart updates without refresh.

The next page shows the view of the dataset in the Jaqpot modelling interface (Figure 17).



Dataset: WalkeyGoldPCtrain #KIL42dC8JSmp

Search:

Compounds	Polydispersity Index	Density	Intensity Mean Hydrodynamic Diameter	Auto DP AES	Z-Average Hydrodynamic Diameter	Number Mean Hydrodynamic Diameter	Net cell association	LSPR peak position nm	Polydispersity Index	Protein density	ZETA POTENTIAL	Localized Surface Plasmon Resonance LSPR index	Log2 transformed	Localized Surface Plasmon Resonance LSPR
G15.AC_1	0.084	19.1	70.97	255.443	57.53	23.49	0.028	518.77	0.27	0.0	-21.78	0.454	-5.184	0.183
G15.AHT_1	0.399	19.1	106.7	240.729	90.06	47.5	0.497	526.28	0.215	0.0	15.22	0.526	-1.009	0.458
G15.Asn-SH_1	0.15	19.1	68.92	240.177	37.75	23.04	0.02	518.57	0.207	0.0	-20.27	0.327	-5.676	0.274
G15.AUT_1	0.326	19.1	83.34	244.228	55.98	29.49	0.402	523.23	0.273	1.0	16.35	0.39	-1.316	0.365
G15.CIT_1	0.138	19.1	64.47	272.769	54.03	28.3	0.023	517.68	0.217	0.0	-25.26	0.293	-5.42	0.21
G15.CTAB_1	0.465	19.1	105.1	213.324	59.7	35.99	0.017	518.9	0.371	0.0	-5.01	0.366	-5.862	0.326
G15.DDT@CTAB_1	0.191	19.1	99.43	247.743	48.4	23.23	0.005	519.37	0.308	0.0	19.6	0.325	-7.59	0.275
G15.DDT@DOTAP_1	0.193	19.1	83.76	235.484	47.34	31.58	0.458	522.1	0.237	0.0	13.92	0.297	-1.128	0.276
G15.DDT@SA_1	0.249	19.1	86.81	249.152	59.93	55.31	0.009	534.13	0.367	0.0	-29.37	0.321	-6.804	0.396
G15.DDT@SDS_1	0.302	19.1	142.83	208.318	100.13	35.19	0.005	550.6	0.37	0.0	-21.68	0.36	-7.675	0.465
G15.F127_1	0.216	19.1	72.78	238.252	46.63	36.47	0.024	519.37	0.197	0.0	-8.96	0.245	-5.361	0.176
G15.Gly-SH_1	0.211	19.1	74.16	243.954	55.39	38.3	0.032	518.6	0.236	0.0	-17.77	0.403	-4.975	0.262
G15.LA_1	0.23	19.1	84.19	268.117	48.09	39.75	0.016	519.13	0.201	1.0	-26.48	0.301	-5.964	0.237
G15.MAA_1	0.273	19.1	64.65	261.99	60.99	47.92	0.014	518.4	0.238	1.0	-24.17	0.449	-6.142	0.281
G15.MES_1	0.297	19.1	83.75	256.869	61.86	72.5	0.109	518.3	0.286	0.0	-25.55	0.476	-3.199	0.226
G15.Met-SH_1	0.167	19.1	72.7	260.336	52.42	28.2	0.016	518.0	0.215	1.0	-22.2	0.282	-5.928	0.216
G15.MHDA_1	0.076	19.1	47.82	266.518	43.62	23.95	0.018	522.77	0.155	0.0	-25.65	0.294	-5.778	0.242
G15.MPA_1	0.284	19.1	70.02	261.392	55.65	39.49	0.024	518.67	0.289	0.0	-27.52	0.451	-5.396	0.304
G15.MUA_1	0.092	19.1	51.1	266.113	45.52	22.4	0.035	521.71	0.166	0.0	-21.13	0.302	-4.847	0.236
G15.MUTA_1	0.156	19.1	139.75	263.69	127.34	30.92	1.081	521.84	0.173	0.0	18.17	0.419	0.113	0.25

Previous 1 2 3 Next

Figure 17 View of the dataset produced

Both the Train and the Predict Dataset have been uploaded already and are available in the Jaqpot modelling interface as Example datasets under the names:

Walkey et al - 56 Gold NPs with description: "56 Gold NPs with 25 PhysChem descriptors"

and

Walkey et al - 28 Gold NPs with description: "28 Gold NPs with 25 PhysChem descriptors".

They can be accessed at http://app.jaqpot.org:8000/data.

Case 2

In the second case, a dataset originating from the same publication, having the same 84 Gold nanoparticles as in Case 1, but this time with relative abundance data for 76 proteins will be used, after having been split into two sets, a Training and a Prediction set with 56 and 28 nanoparticles respectively.

Similarly to the previous case, users can follow the Create Dataset route to input the data into the Jaqpot Quattro Modelling Interface. For the sake of brevity, both the Train and the Predict Dataset have been uploaded already and are available in the Jaqpot modelling interface as Example datasets under the names:

Walkey et al - 56 Gold NPs with description: "56 Gold NPs with 76 protein corona descriptors"

and

Walkey et al - 28 Gold NPs with description: "28 Gold NPs with 76 protein corona descriptors".

They can be accessed at http://app.jaqpot.org:8000/data.

4. Model training and use for predictions

The workflow that will be followed in the three modelling cases is shown in Figure 18.

Figure 18 Workflow of the modelling cases

Case 1

Please note that the dataset of Case 1 has undergone processing prior to modelling (dotted "Create dataset" area in figure). The following steps guide the user to performing the modelling operations outlined in Figure 18.

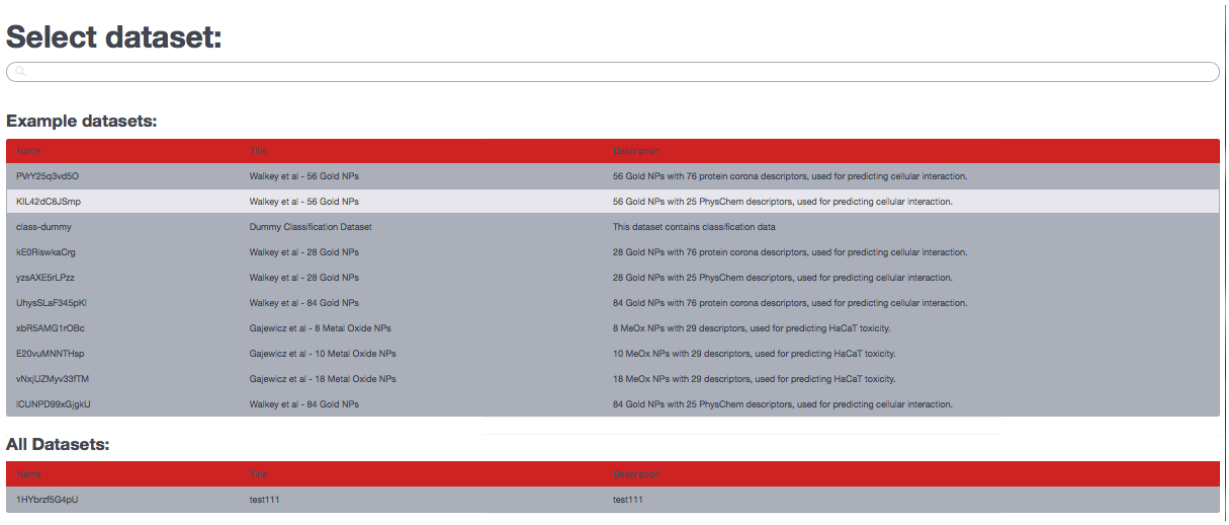


Figure 19 Train: Dataset selection

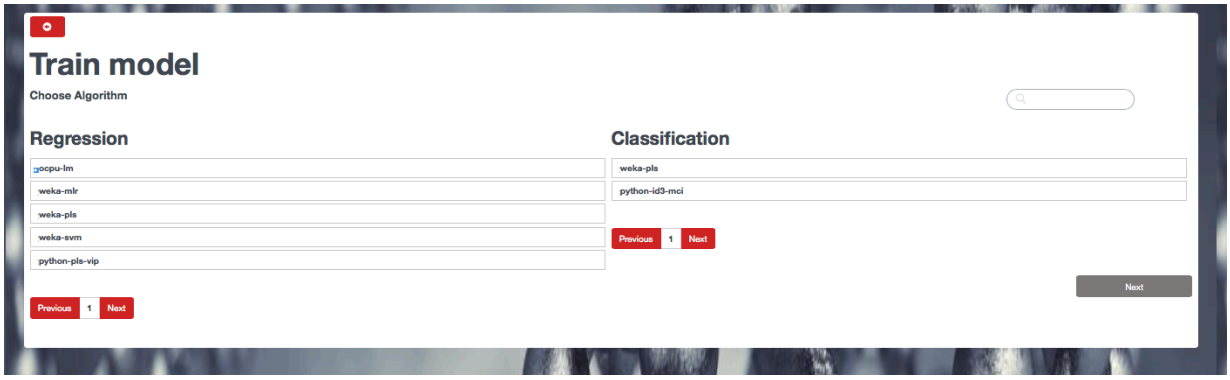


Figure 20 Train: Algorithm selection

Users should access the start page for registered users and select the "**Train a model**" option in **Actions**. This leads to the screen of Figure 19, where the ***Walkey et al - 56 Gold NPs*** dataset with description: "56 Gold NPs with 25 PhysChem descriptors" should be selected. The screen of Algorithm selection follows (Figure 20). Here **OCPU-lm** should be selected. In Figure 21 the appropriate parameters are given, as shown in the screen (all settings are kept to default). Please note that "Log2 transformed" property has been selected as the end-point, so it need to be excluded from the input variables, as well as the "Net cell association" property, as shown in the figure. Here *Scaling* was not selected and *Domain of Applicability* (DoA) calculations using the *Leverage method* were requested (DoA values closer to 1 show good applicability of model).

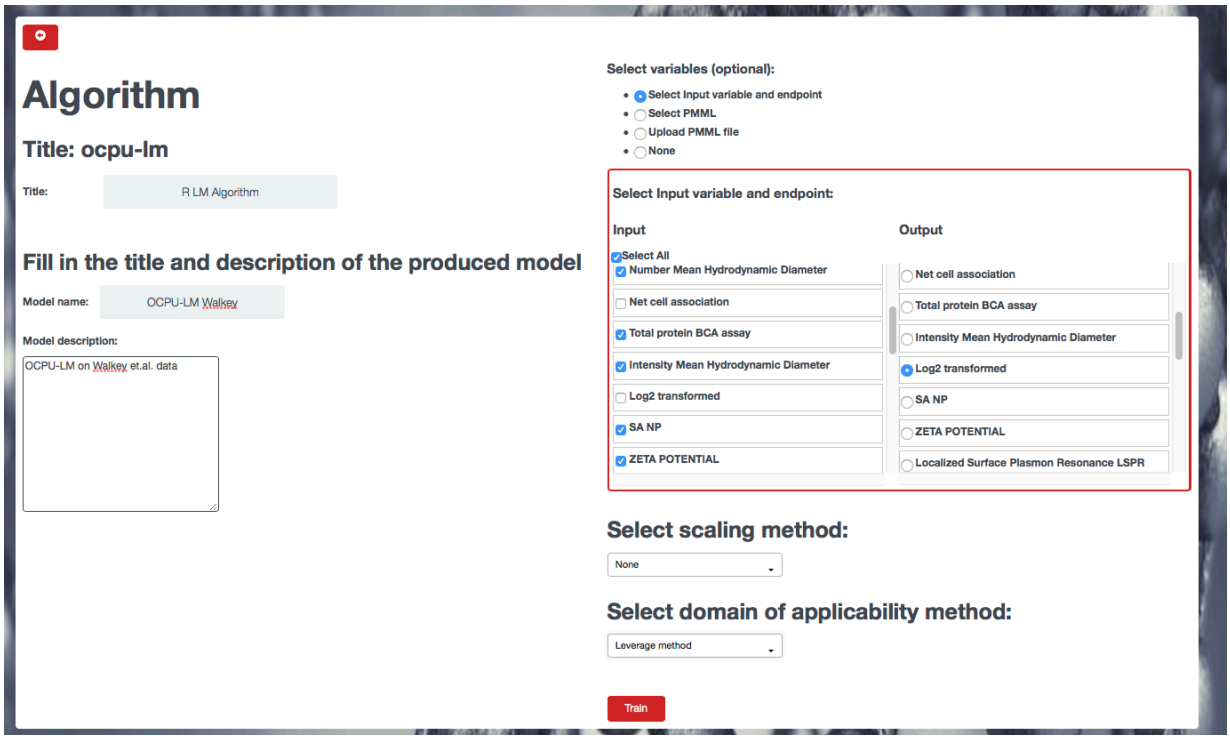


Figure 21 Train: Algorithm parameters – selection of input-output variables

After an intermediate Task page (Figure 22) the result of the Train action is given, which is the page of the model (Figure 23), which contains full information on the model.

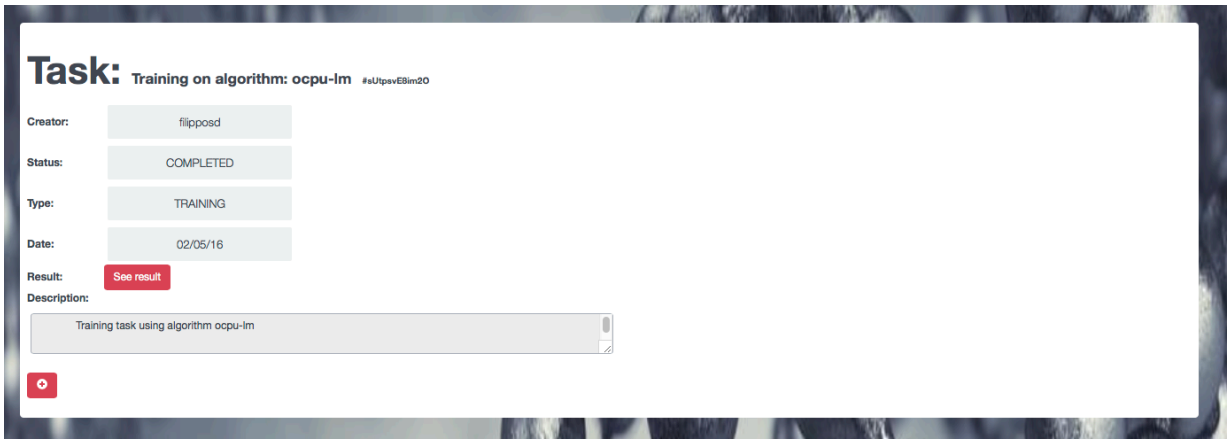


Figure 22 Train: Training on algorithm task page

Figure 23 Train: Model page with details

A very important aspect here is the generation of the PMML representation of the model, which allows the seamless integration of models into various platforms. This model's PMML file is accessible by users who have logged in (at least as guests) at http://app.jaqpot.org:8000/m_detail?name=LXjNxxY3I3J1.

Predictions will be made by clicking the Predict button and selecting the ***Walkey et al - 28 Gold NPs*** dataset with description: "28 Gold NPs with 25 PhysChem descriptors".
The dataset with predictions is given next (Figure 25).

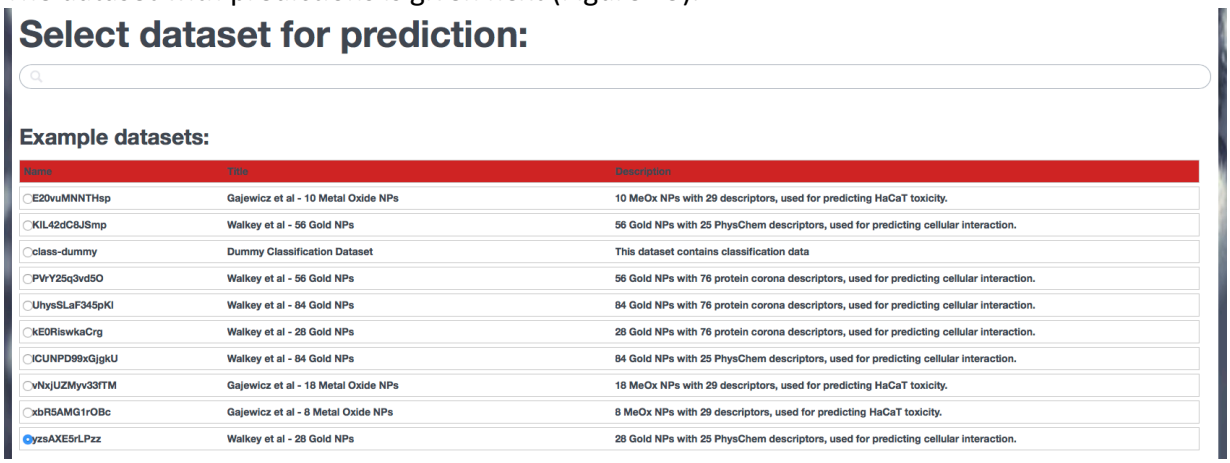


Figure 24 Predict: selection of a dataset

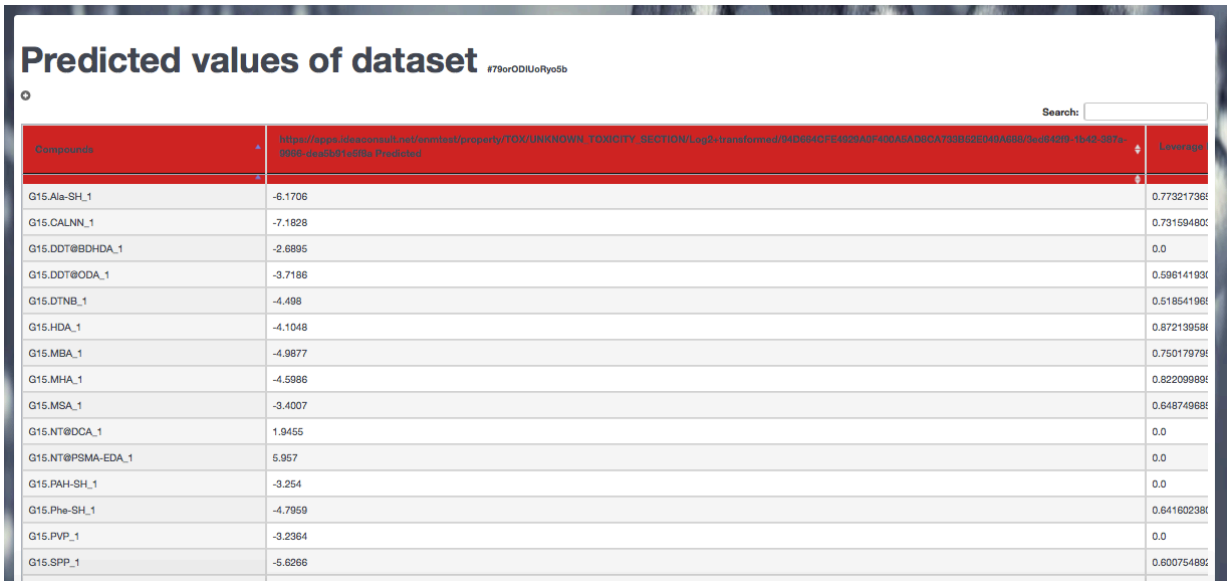


Figure 25 Predict: Predictions by model

Case 2

Similarly to Case 1, the ***Walkey et al - 56 Gold NPs*** dataset with description: "56 Gold NPs with 76 protein corona descriptors" should be selected, using the **python-pls-vip** algorithm (Figure 26) with the parameters shown in Figure 27, which leads to the model shown in http://app.jaqpot.org:8000/m_detail?name=jJPN5UrWvTu6.

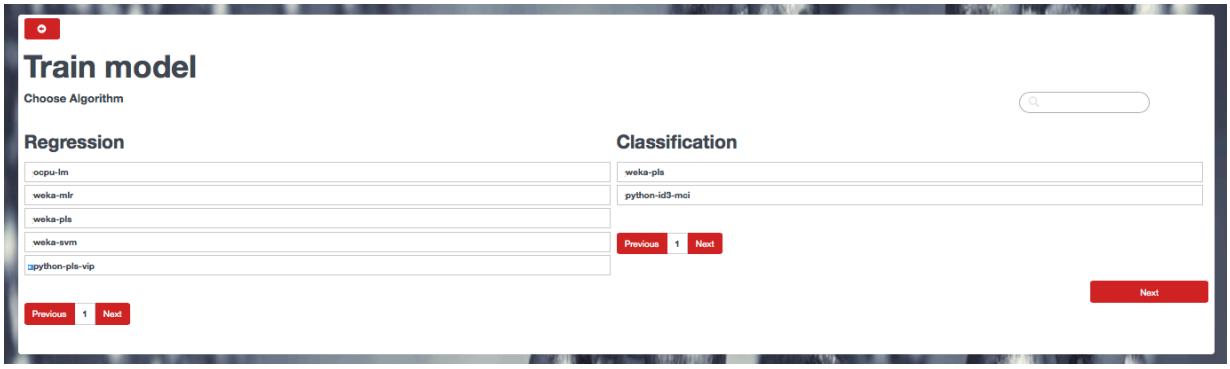


Figure 26 Train: Algorithm selection

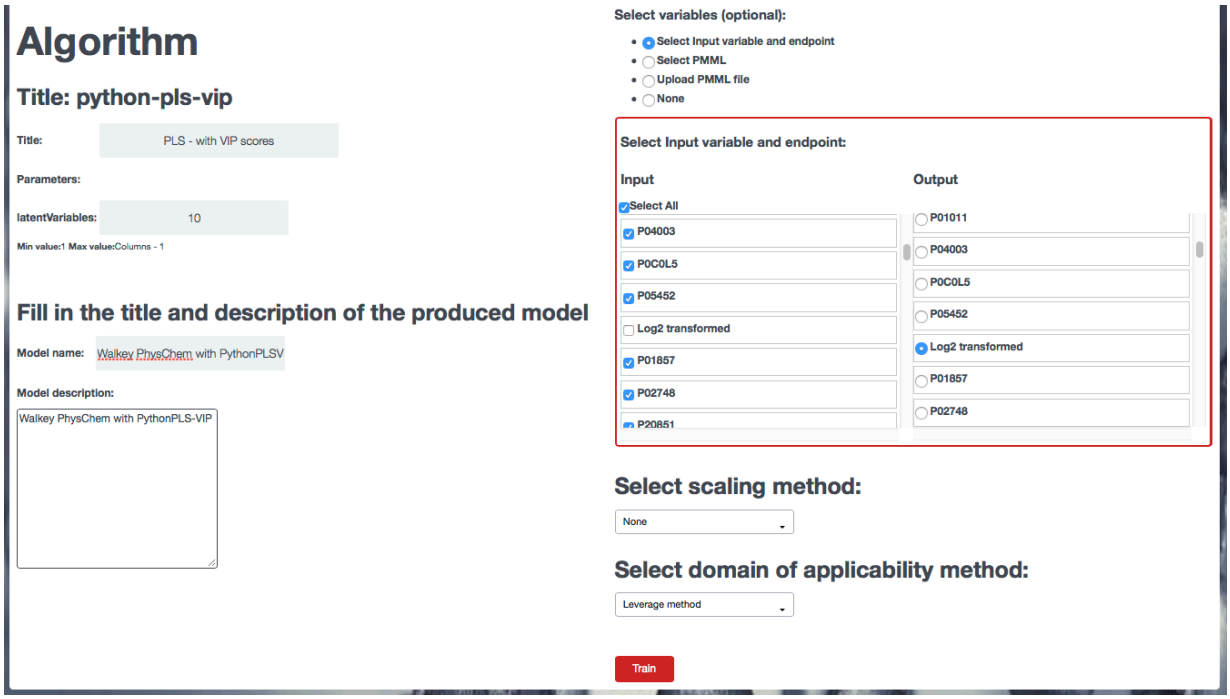
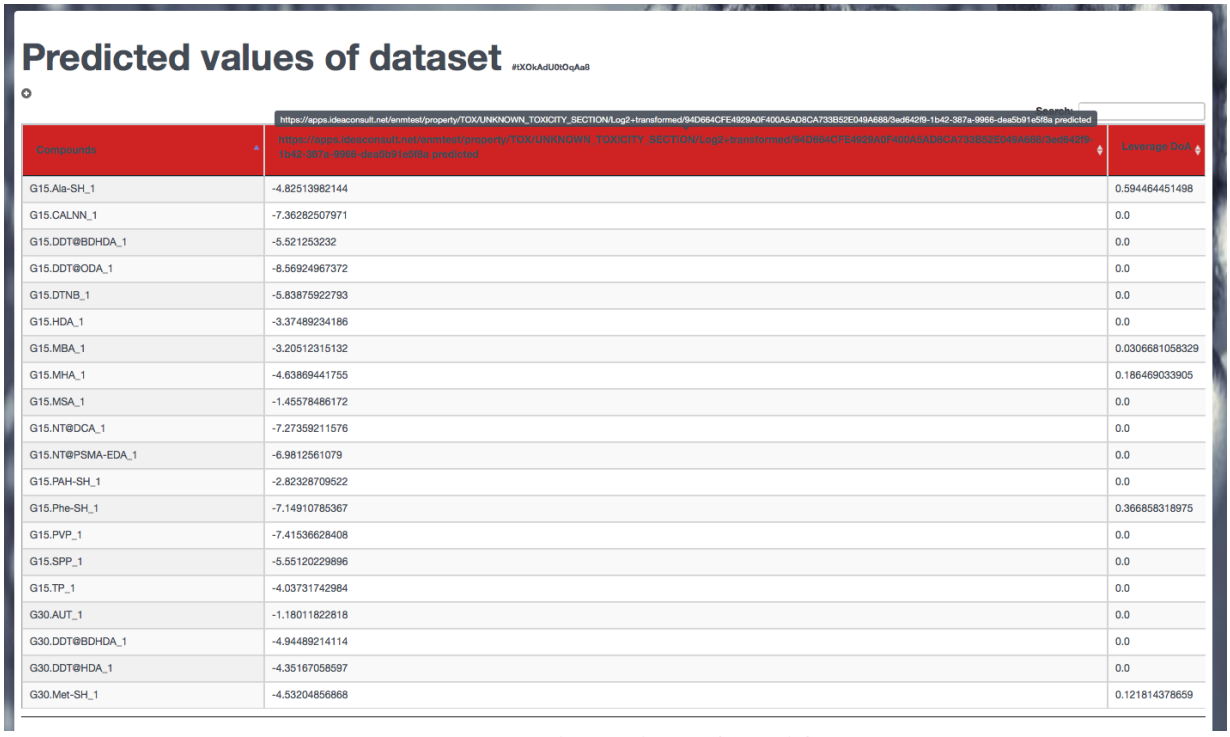


Figure 27 Train: Algorithm parameters – selection of input-output variables



Predicted values of dataset

Compounds	https://apps.ideaconsult.net/enmtest/property/TOX/UNKNOWN_TOXICITY_SECTION/Log2+transformed/94D664CFE4929A0F400A5AD8CA733B52E049A688/3ed642f9-1b42-387a-9966-dea5b91e5f6a predicted	Leverage DoA
G15.Ala-SH_1	-4.82513982144	0.594464451498
G15.CALNN_1	-7.36282507971	0.0
G15.DDT@BDHDA_1	-5.521253232	0.0
G15.DDT@ODA_1	-8.56924967372	0.0
G15.DTNB_1	-5.83875922793	0.0
G15.HDA_1	-3.37489234186	0.0
G15.MBA_1	-3.20512315132	0.0306681058329
G15.MHA_1	-4.63869441755	0.186469033905
G15.MSA_1	-1.45578486172	0.0
G15.NT@DCA_1	-7.27359211576	0.0
G15.NT@PSMA-EDA_1	-6.9812561079	0.0
G15.PAH-SH_1	-2.82328709522	0.0
G15.Phe-SH_1	-7.14910785367	0.366858318975
G15.PVP_1	-7.41536628408	0.0
G15.SPP_1	-5.55120229896	0.0
G15.TP_1	-4.03731742984	0.0
G30.AUT_1	-1.18011822818	0.0
G30.DDT@BDHDA_1	-4.94489214114	0.0
G30.DDT@HDA_1	-4.35167058597	0.0
G30.Met-SH_1	-4.53204856868	0.121814378659

Figure 28 Predict: Predictions by model

Case 3

Use of PMML for determination of input variables and transformations

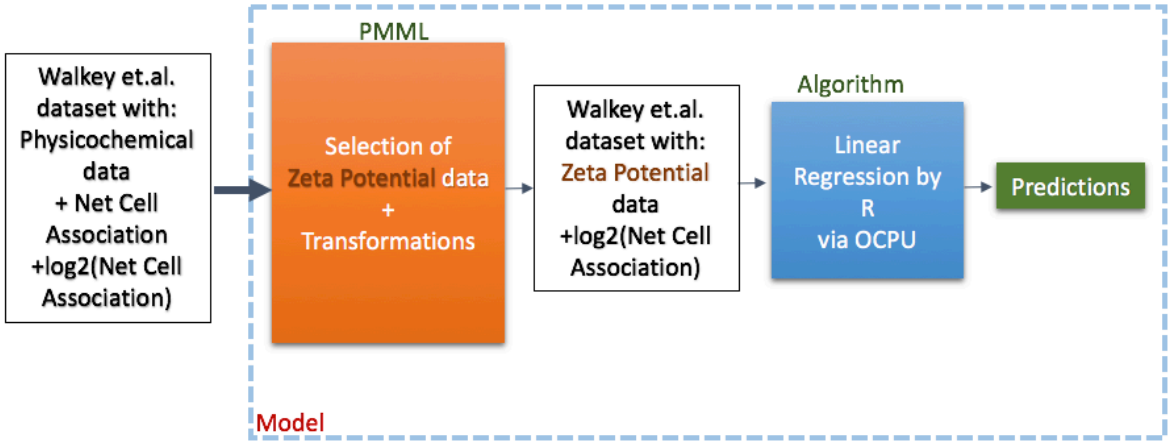


Figure 29 Input variables determination via UI and PMML transformations on the Case 1 dataset

In this case 3, users should use the PMML file in the Workshop webpage in order to select a subset of properties, as shown in the above figure. The input dataset is ***Walkey et al - 56 Gold NPs*** with description: "56 Gold NPs with 25 PhysChem descriptors" and ***Walkey et al - 28 Gold NPs*** with description: "28 Gold NPs with 25 PhysChem descriptors" should be used for predictions.

Additionally, another way to get predictions without previously uploading a dataset is by using the "Upload dataset" option shown in Figure 30.

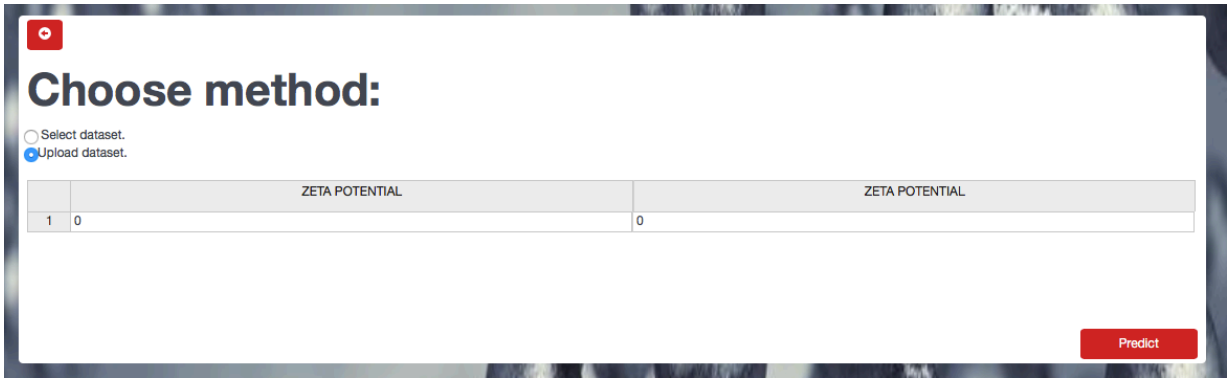


Figure 30 Field to upload a new dataset fore predictions by pasting it using the interface

1. NTUA eNanoMapper Web services, available at http://app.jaqpot.org:8000, described at http://app.jaqpot.org:8080/jaqpot/swagger/
2. Hardy, B. et al, Collaborative Development of Predictive Toxicology Applications, Journal of Cheminformatics 2010, 2:7.
3. Walkey, C. D., Olsen, J. B., Song, F., Liu, R., Guo, H., Olsen, D. W. H., Chan, W. C. W. (2014). Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. ACS Nano, 8:3.