Nano modeling workshop

Workshop:	Nano modeling workshop
DATE / PLACE:	Tuesday, October 25, 2016 - Schloss Beuggen 11, Rheinfelden, Germany
TIME:	13:00-17:00

TITLE:	Extracting knowledge from data using the JaqPot Modelling Tool

SPEAKER:	Philip Doganis, Georgios Drakakis
AUTHORS:	Philip Doganis, Georgios Drakakis *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 metal oxide Nanoparticles, based on experimental data, publically available in the publication of Gajewicz et. al.[3].

Participants will go through the workflow of constructing a model from a dataset drawn from the eNanoMapper database into the Jaqpot platform, getting predictions based on the model and evaluating its efficiency through model validation. Finally, users will work on the creation of predictive models using statistical and machine learning algorithms.

Figure 1 Workshop Outline

1. Login & Authorization to the Jaqpot and eNanoMapper services

Users access the Jaqpot homepage by NTUA http://jaqpot.org (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.

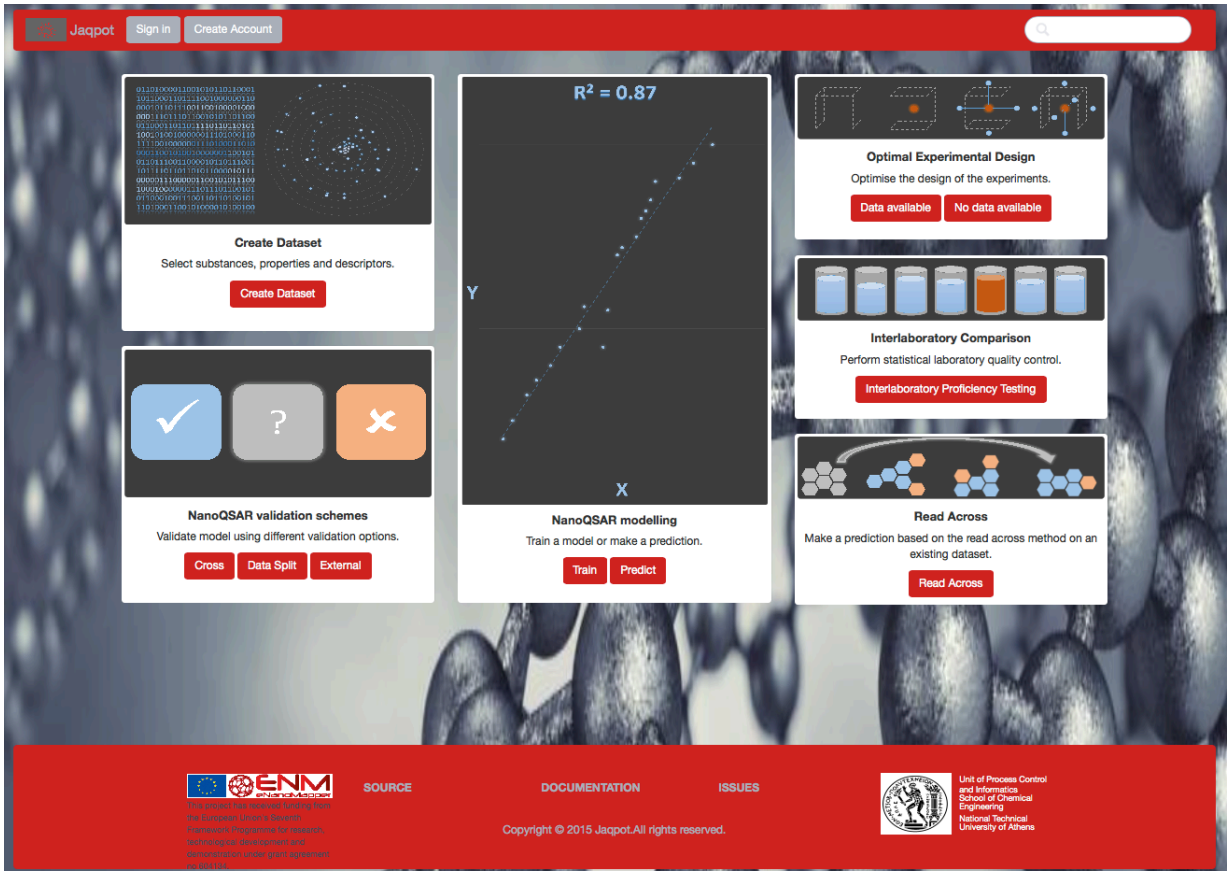

Figure 2 Homepage of eNanoMapper NTUA Modelling interface

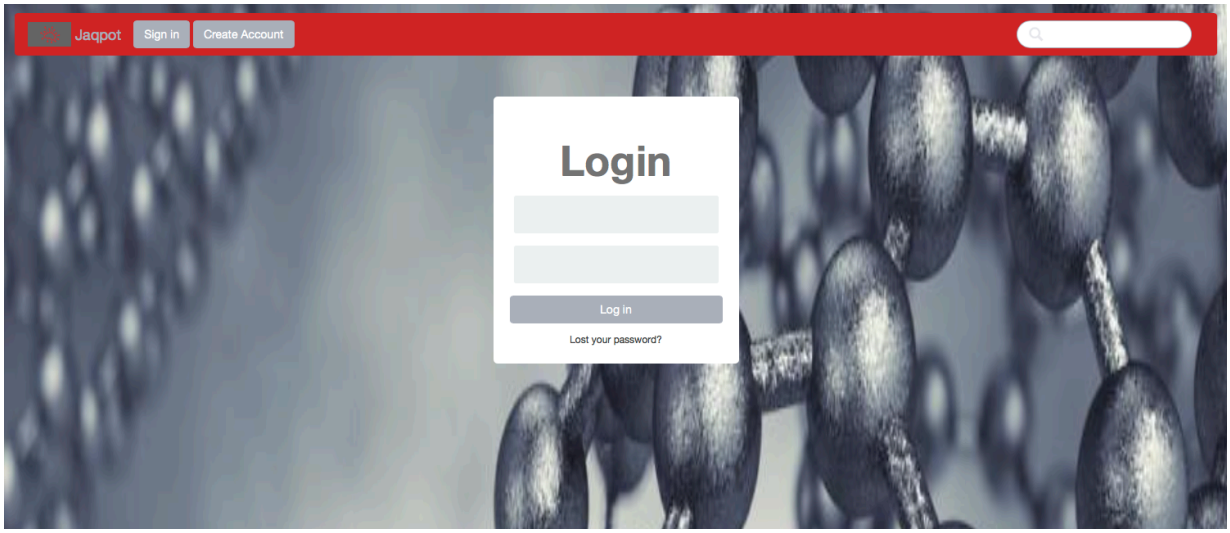

Figure 3 Login page

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://old.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).

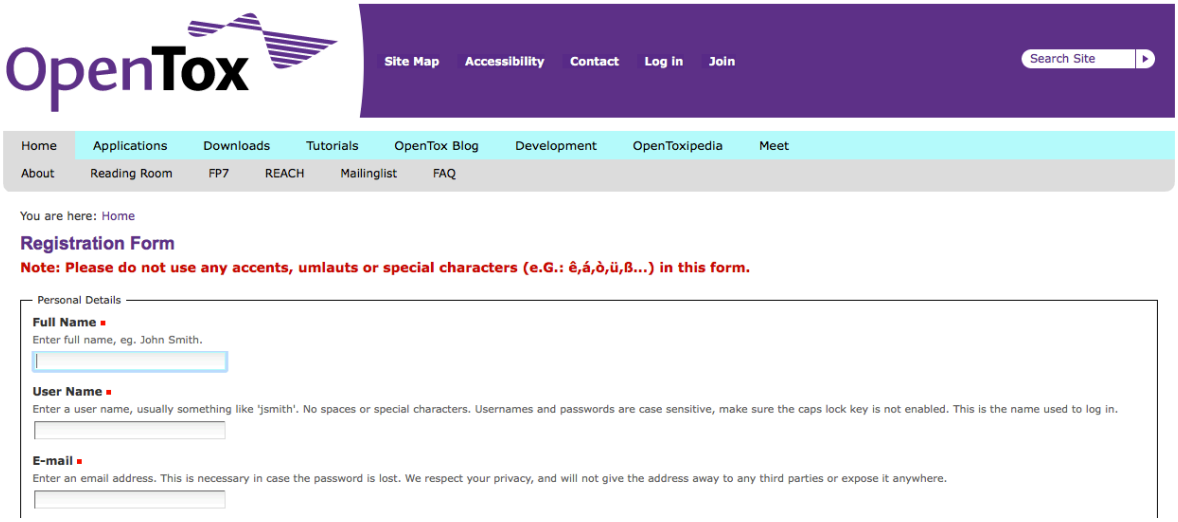

Figure 4 OpenTox Registration page (http://www.opentox.org/join_form)

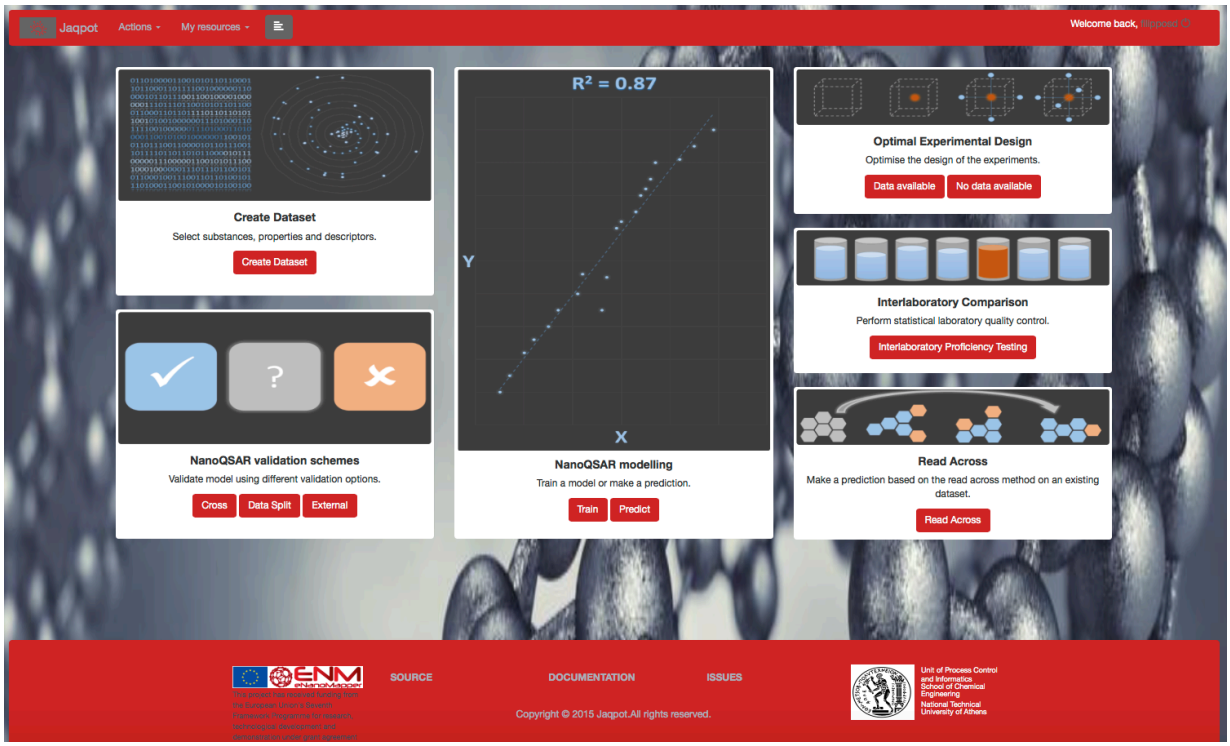

Figure 5 Starting page for registered users

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://jaqpot.org:8080/jaqpot/swagger/ to view the documentation of the API and experiment with its possibilities using the Swagger interface hosted there. As the eNanoMapper Jaqpot modelling interface leaves the testing phase and reaches finalization, a manual for its functionality will be released.

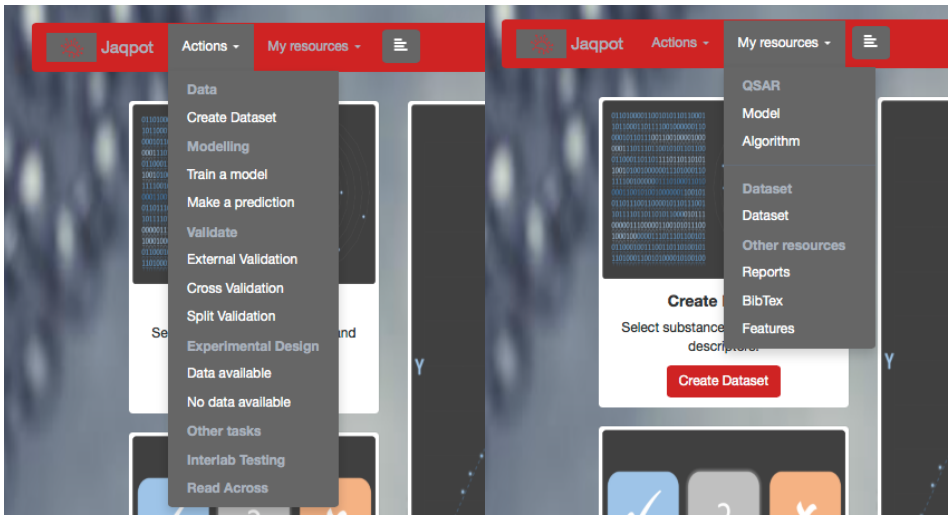

Figure 6 Jaqpot menu tree

The menu tree of the Jaqpot modelling interface is shown in Figure 6. 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 6):

- **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 Crystallographic data are included in the dataset, resulting to the creation of a local dataset.
- **Modelling** using a dataset in the data.enanomapper.net database or the local dataset and applying

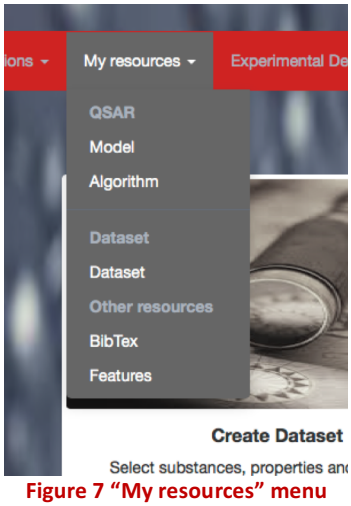

Figure 7 "My resources" menu

"My resources" menu

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 by Python
  - *python-lasso*: Lasso Regression by Python
  - *python-lm*: Linear regression by Python
- Classification
  - *weka-pls*: Partial Least Squares Classification by Weka
  - *python-id3-mci*: ID3 classification algorithm
  - *Bernoulli Naive Bayes*
  - *Generalised Naive Bayes*
  - *Multinomial Naive Bayes*
  - *CMI Decision Tree*
  - *Id3 - with MCI*

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.
- **Validate** model to review its efficiency by applying it to a dataset.
  Available choices are:
  - External validation: apply the model to another dataset
  - Cross validation: a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set, more info here: https://en.wikipedia.org/wiki/Cross-validation_(statistics)
  - Split validation: split the data into training and test, train a model and apply it to the test dataset.
- **Experimental design**
  - With data available
  - No data available
- **Other:** Interlaboratory testing and Read Across will soon exit the test phase and be released.

The resources that are available to the user are:

- **QSAR**
  - **Model**: a repository of models created by the user and example models
  - **Algorithm**: a list of the available algorithms offered by Jaqpot
- **Dataset**
  - A repository of datasets created by the user and example datasets
- **Reports, Bibtex, Features**: future capabilities

3. Model training and use for predictions

The workflow that will be followed in the three modelling cases is shown in Figure 8.

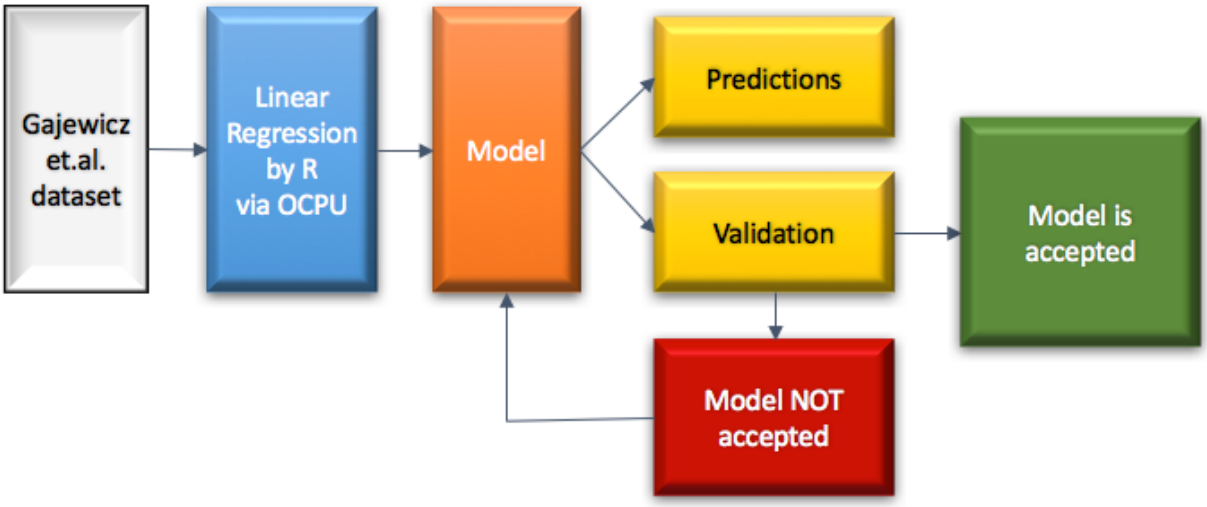

Figure 8 Workflow of the modelling cases

We will work on a dataset investigating the Cytotoxicity of Metal Oxides Nanoparticles by Gajewicz et.al.[3].
Please note that the dataset we will work on has undergone processing prior to modelling. The following steps guide the user to performing the modelling operations outlined in Figure 8.

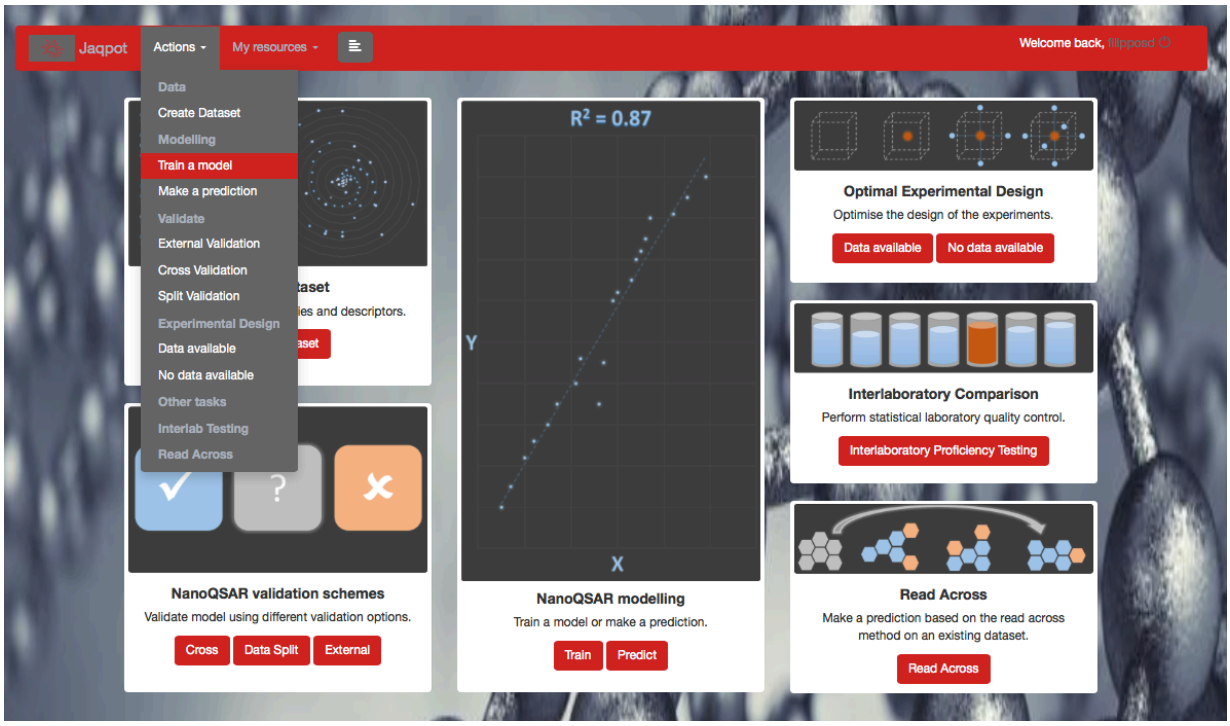

Figure 9 *Train a model* Action

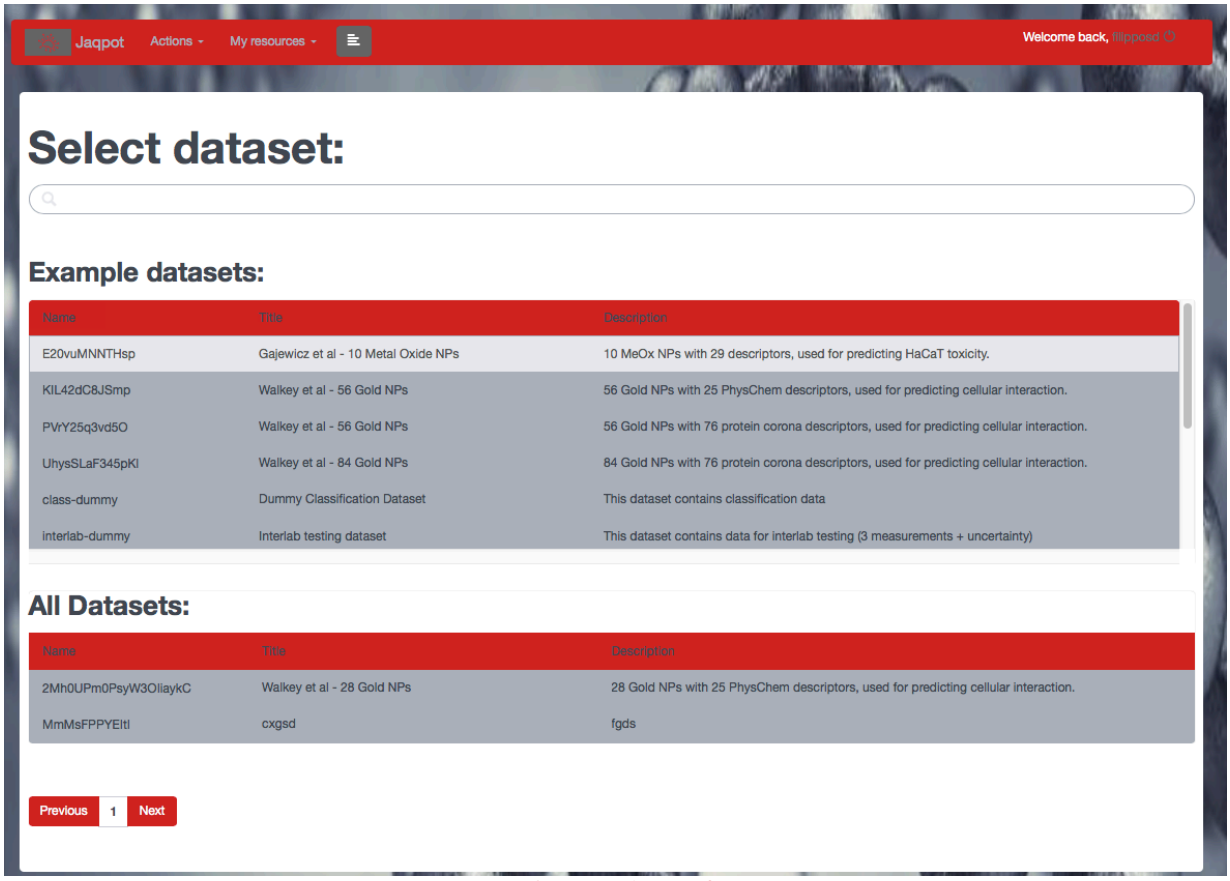

Figure 10 Selecting a dataset for training

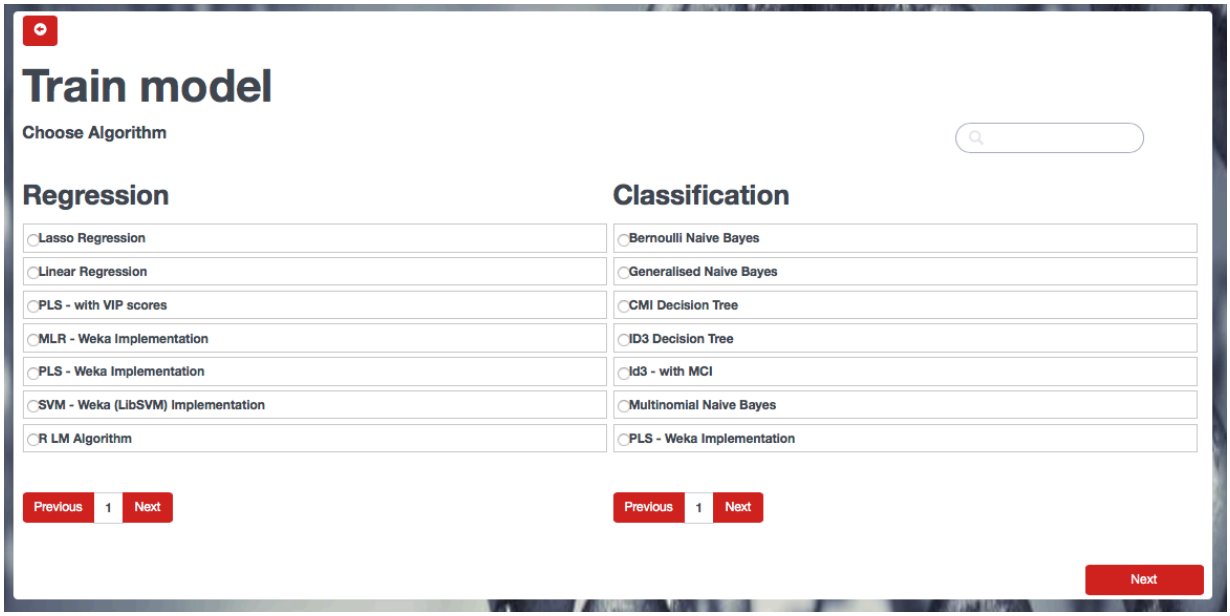

Figure 11 Train: Algorithm selection

Users should access the start page for registered users and select the "Actions→Train" option, as shown in Figure 9, where the **Gajewicz et al - 10 Metal Oxide NPs** dataset should be selected (Figure 10). The screen of Algorithm selection follows (Figure 11). Here **Lasso Regression** should be selected.

For details on the algorithms and their parameters please consult this document: http://www.enanomapper.net/deliverables/d4/150801_eNanoMapper-D4_3final.pdf .

In Figure 12 the appropriate parameters are given, as shown in the screen (all settings are kept to default). Here *Scaling* was not selected and no *Domain of Applicability* (DoA) calculations were requested.

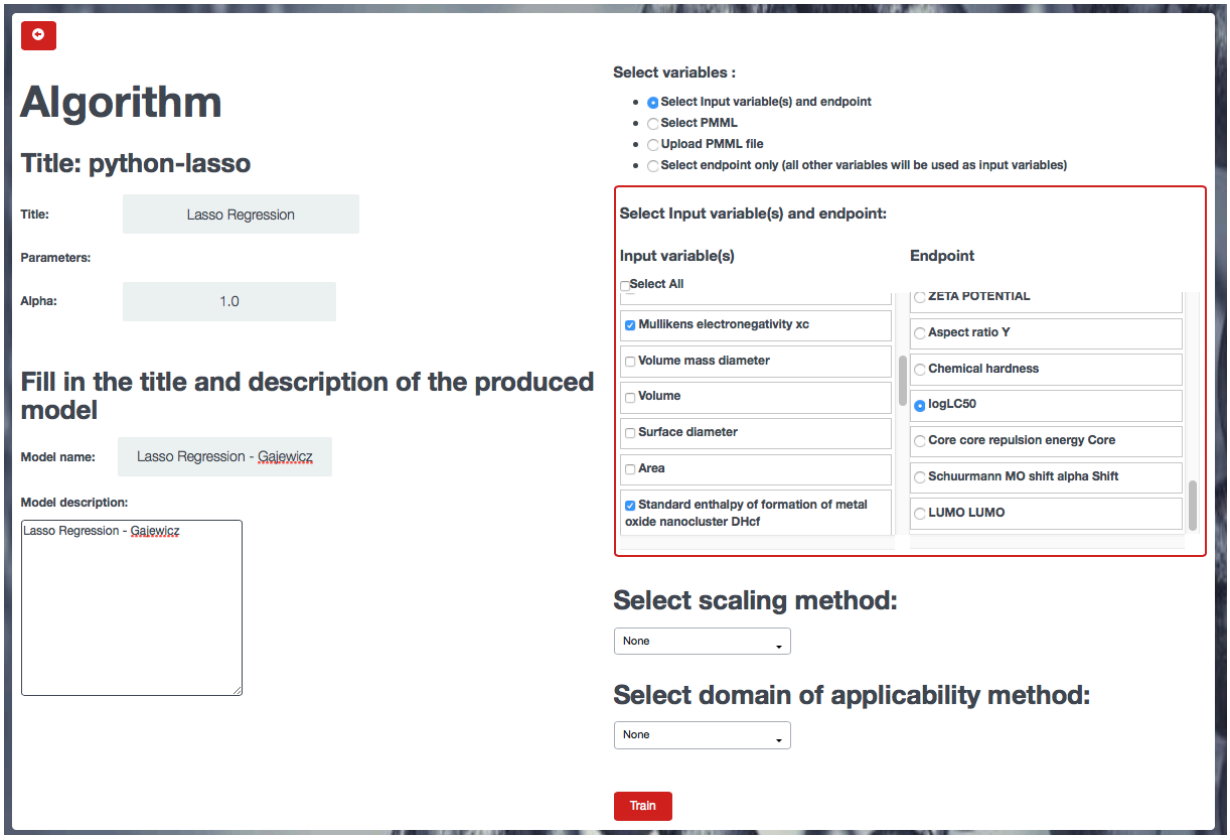

Figure 12 Train: Algorithm parameters – selection of input-output variables

After an intermediate Task page (Figure 13) the result of the Train action is given, which is the page of the model (Figure 14), which contains full information on the model.

Figure 13 Train: Training on algorithm task page

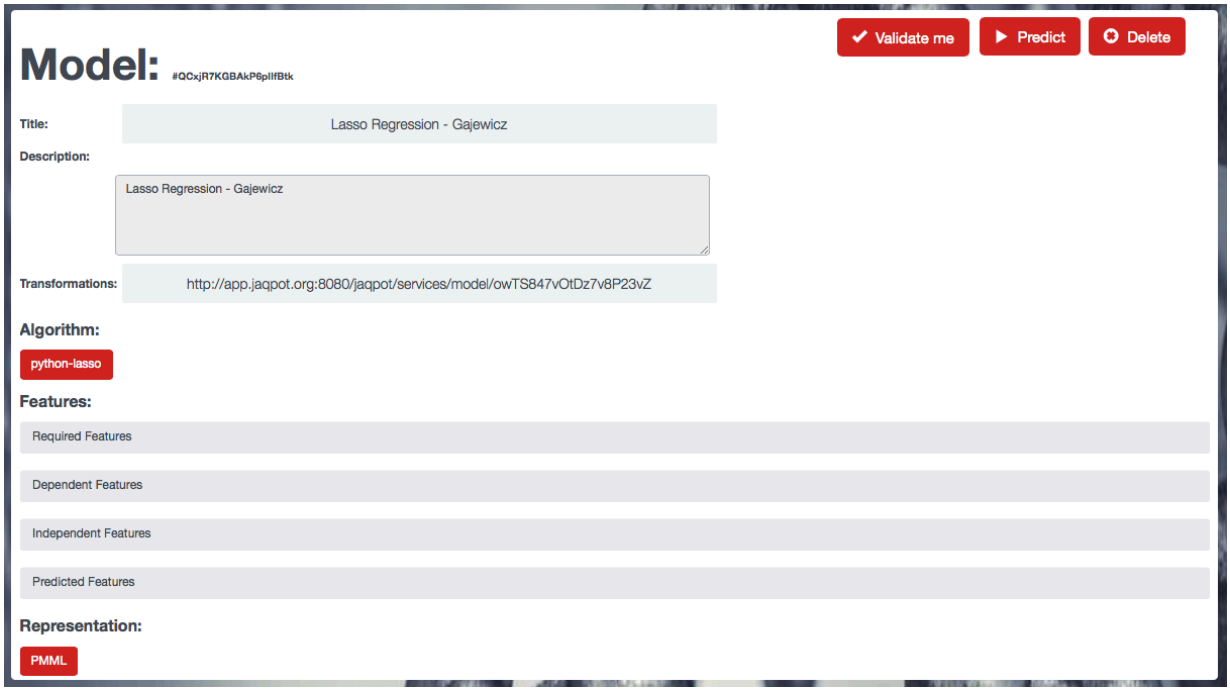

Figure 14 Train: Model page with details

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://jaqpot.org/m_detail?name=QCxjR7KGBAkP6pIIfBtk .

Predictions will be made by clicking the Predict button and selecting the **Gajewicz et al - 8 Metal Oxide NPs** Dataset (Figure 15). The dataset with predictions is given next (http://jaqpot.org/predicted_dataset?name=5RCO6mRmFU5hiz4zMIAv&model=QCxjR7KGBAkP6pIIfBtk, Figure 16).

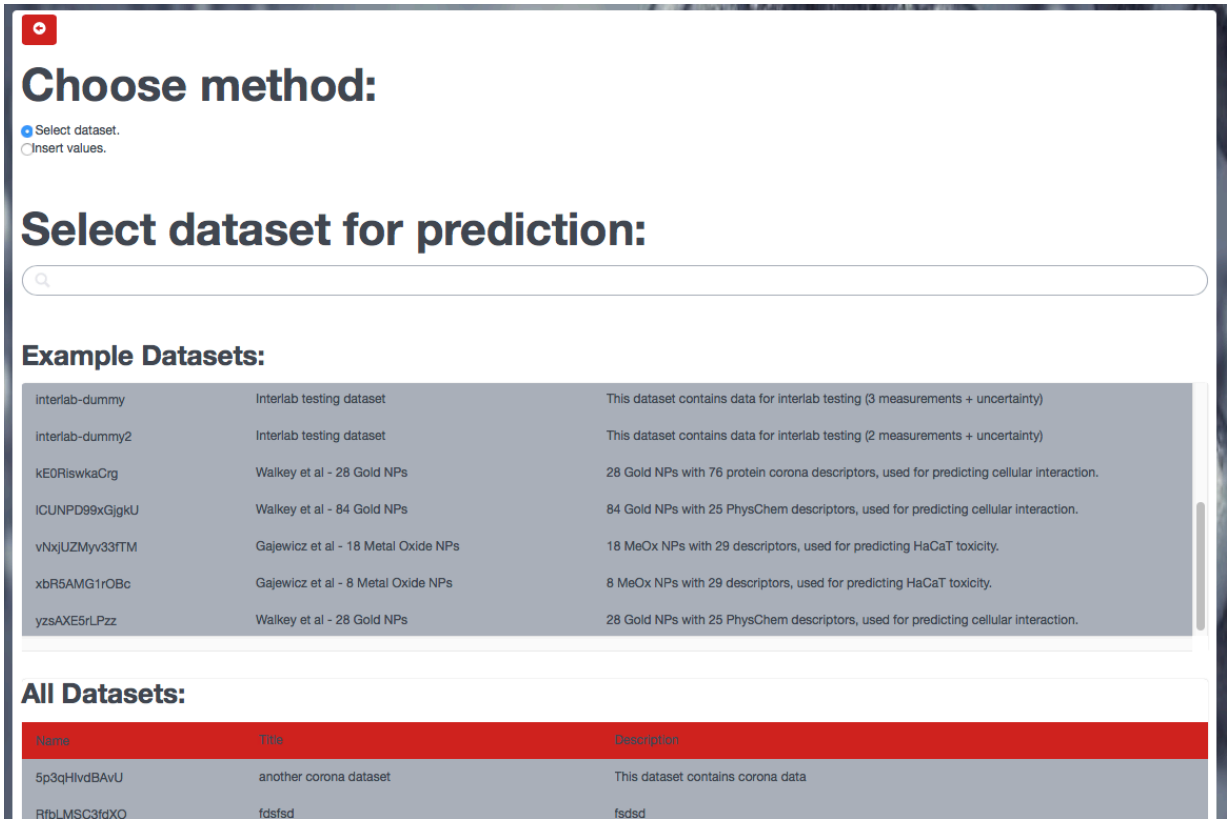

Figure 15 Predict: selection of a dataset

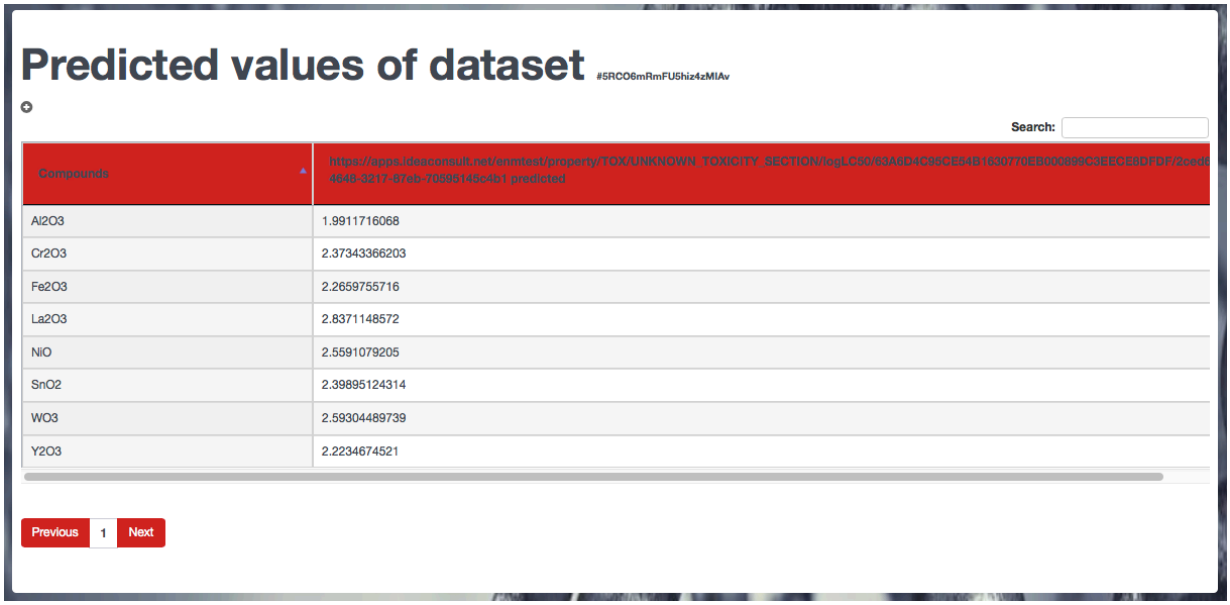

Figure 16 Predict: Predictions by model

4. Model validation

A very important step before accepting mathematical models for nanomaterials, and models in general, is model validation, defined as "*the process of deciding whether the numerical results quantifying hypothesized relationships between variables, obtained from regression analysis, are acceptable as descriptions of the data*" (https://en.wikipedia.org/wiki/Regression_validation).

The validation process can be performed on the Jaqpot platform by clicking the Validate button on the model's page (http://jaqpot.org/m_detail?name=QCxjR7KGBAkP6pIIfBtk, Figure 14), selecting an appropriate dataset, not the one used for training, but one that still contains the same properties and substances. Here the **Gajewicz et al - 8 Metal Oxide NPs** Dataset should be selected (Figure 15).

This leads to a report page http://jaqpot.org/report?name=6TgaNqGYrETZw9X (Figure 17).

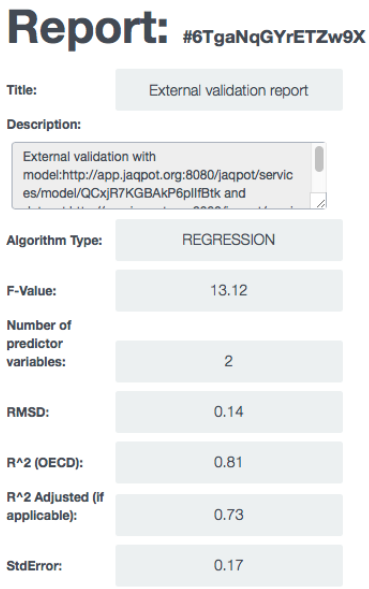

Figure 17 Model validation report

Please not the following on the definition of R^2 (https://en.wikipedia.org/wiki/Coefficient_of_determination):

In statistics, the **coefficient of determination**, denoted R^2 or r^2 and pronounced "R squared", is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable. If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

then the variability of the data set can be measured using three sums of squares formulas:

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

5. Exercise: Explore algorithms

It is now your exercise to:

- navigate the Jaqpot platform
- try out different models
- validate your models
- evaluate model performance

What have you found?

REFERENCES

1. NTUA eNanoMapper Web services, available at http://jaqpot.org, described at http://jaqpot.org:8080/jaqpot/swagger/
2. Hardy, B. et al, Collaborative Development of Predictive Toxicology Applications, Journal of Cheminformatics 2010, 2:7.
3. Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. Nanotoxicology 2014, 5390 (April 2016), 1–13 DOI: 10.3109/17435390.2014.930195.