

## Hands-on Workshop on eNanoMapper tools + services

<b>Workshop:</b>	Hands-on Workshop on eNanoMapper tools + services
<b>DATE / PLACE:</b>	29-30 September 2016 / National Technical University of Athens
<b>TIME:</b>	10:00-11:30

<b>TITLE:</b>	Extracting knowledge from data using the JaqPot Modelling Tool
---------------	--

<b>SPEAKER:</b>	Philip Doganis
<b>AUTHORS:</b>	Philip Doganis <i>School of Chemical Engineering, National Technical University of Athens</i>

## ABSTRACT

The workshop will offer hands-on work on the development of nanoQSAR models based on data available from the [data.enanomapper.net](http://data.enanomapper.net) server, making use of the eNanoMapper computational infrastructure from NTUA<sup>1</sup> that extends the OpenTox API<sup>2</sup>. 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.<sup>3</sup>.

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.

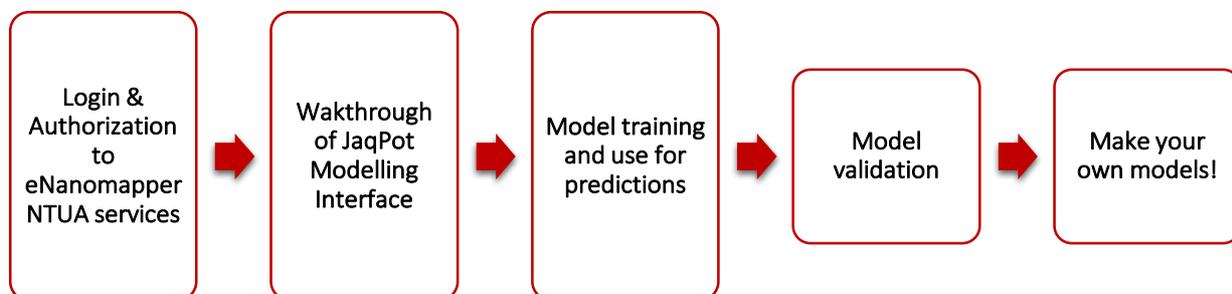
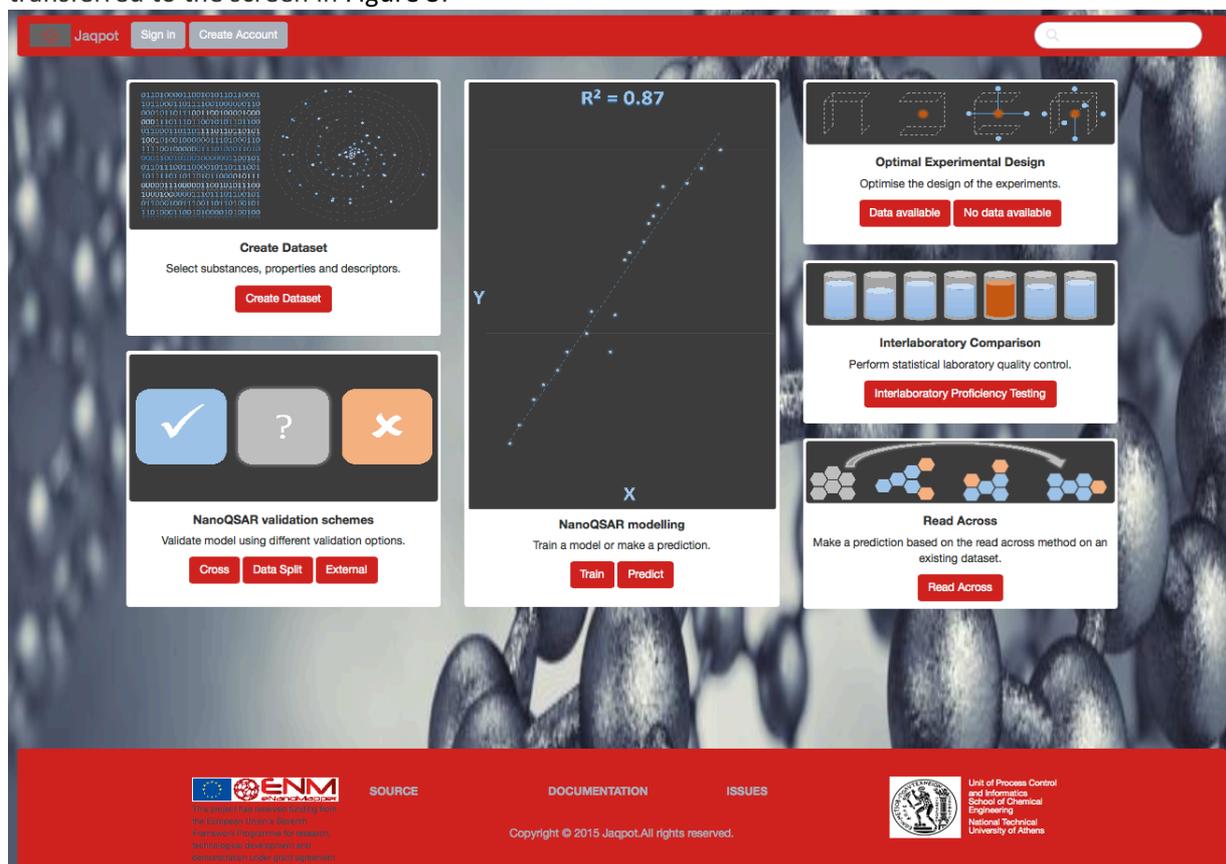


Figure 1 Workshop Outline

## 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.

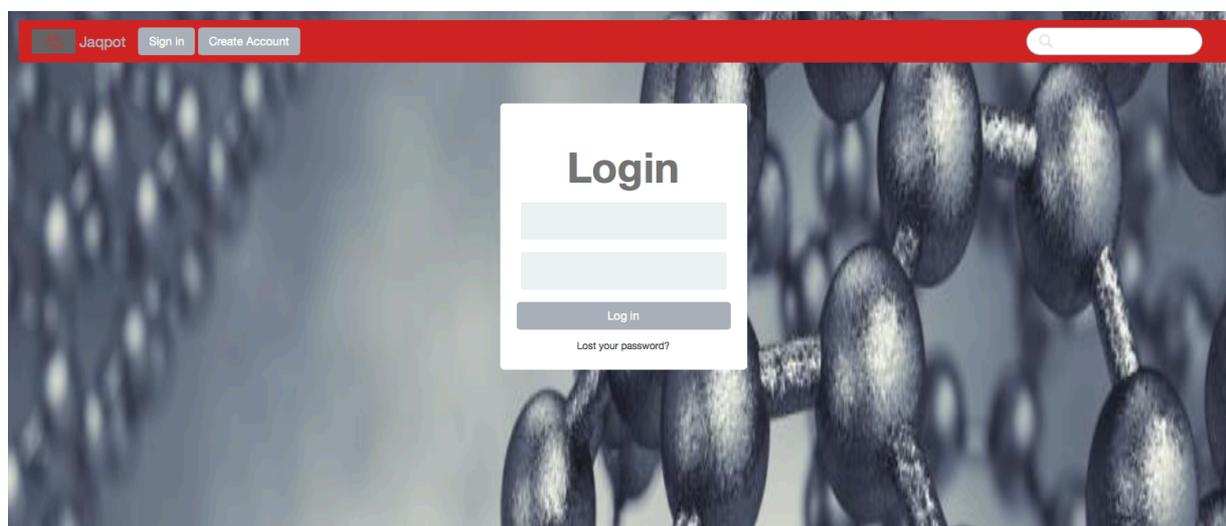


The screenshot shows the Jaqpot homepage with a navigation bar containing 'Jaqpot', 'Sign in', and 'Create Account' buttons. The main content area features several interactive panels:

- Create Dataset:** A panel with a grid of binary code and a scatter plot, with a 'Create Dataset' button.
- NanoQSAR validation schemes:** A panel with three buttons: a checkmark (Cross), a question mark (Data Split), and an 'X' (External).
- NanoQSAR modelling:** A central panel displaying a scatter plot with a regression line and the text  $R^2 = 0.87$ . It includes 'Train' and 'Predict' buttons.
- Optimal Experimental Design:** A panel with a network diagram and buttons for 'Data available' and 'No data available'.
- Interlaboratory Comparison:** A panel with a diagram of test tubes and a button for 'Interlaboratory Proficiency Testing'.
- Read Across:** A panel with a diagram of molecular structures and a 'Read Across' button.

The footer contains logos for ENM, SOURCE, DOCUMENTATION, ISSUES, and the National Technical University of Athens (NTUA).

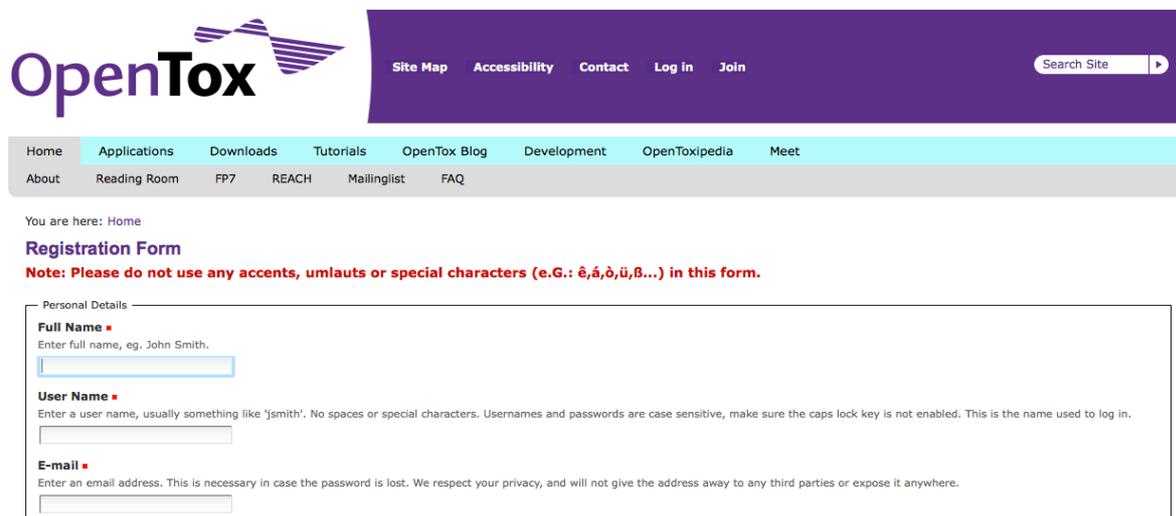
Figure 2 Homepage of eNanoMapper NTUA Modelling interface



The screenshot shows the Jaqpot login page. It features a central white 'Login' form with two input fields for username and password, and a 'Log in' button. Below the button is a link for 'Lost your password?'. The background is a dark image of a molecular structure.

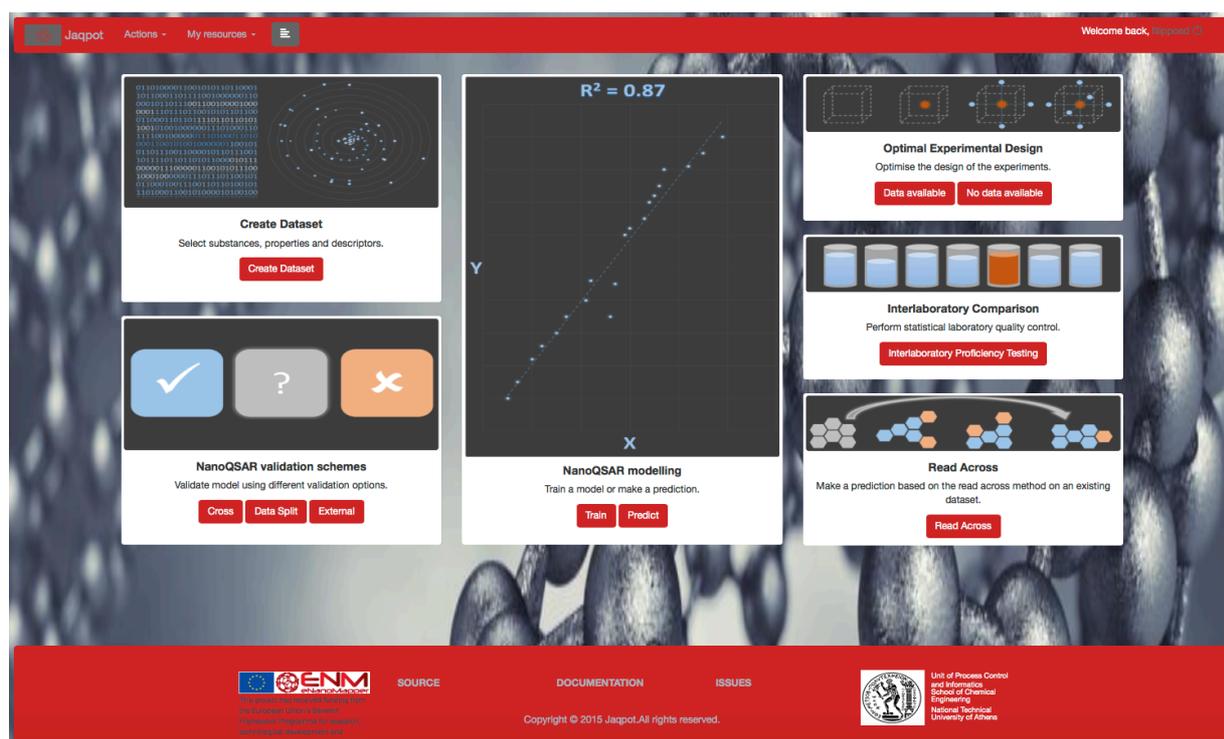
Figure 3 Login page

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](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).



The screenshot shows the OpenTox website's registration page. At the top, there is a navigation bar with links for Site Map, Accessibility, Contact, Log in, and Join, along with a search site input field. Below this is a secondary navigation bar with links for Home, Applications, Downloads, Tutorials, OpenTox Blog, Development, OpenToxipedia, and Meet. The main content area is titled "Registration Form" and includes a note: "Please do not use any accents, umlauts or special characters (e.g.: ê,á,ð,ü,ß...) in this form." The form is divided into "Personal Details" and contains three sections: "Full Name" with a text input field and a "Full Name" label; "User Name" with a text input field and a "User Name" label; and "E-mail" with a text input field and an "E-mail" label. Each section has a brief instruction on how to fill it out.

Figure 4 OpenTox Registration page ([http://www.opentox.org/join\\_form](http://www.opentox.org/join_form))



The screenshot shows the Jaqpot starting page for registered users. The page has a red header with the Jaqpot logo and navigation links. The main content area is divided into several sections: "Create Dataset" with a "Create Dataset" button; "NanoQSAR validation schemes" with buttons for "Cross", "Data Split", and "External"; "NanoQSAR modelling" with a scatter plot showing  $R^2 = 0.87$  and buttons for "Train" and "Predict"; "Optimal Experimental Design" with buttons for "Data available" and "No data available"; "Interlaboratory Comparison" with a button for "Interlaboratory Proficiency Testing"; and "Read Across" with a "Read Across" button. The footer contains logos for the European Union, ENM, SOURCE, DOCUMENTATION, ISSUES, and the National Technical University of Athens.

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 is 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.

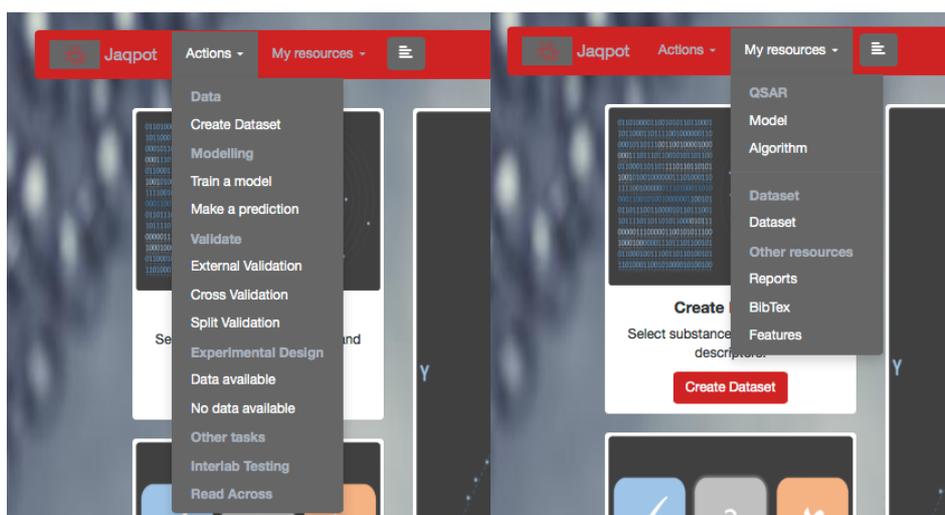


Figure 6 Jaqpot menu tree

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.enanmapper.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.enanmapper.net database or the local dataset and applying

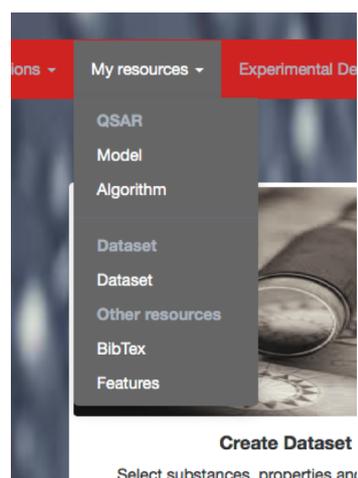


Figure 7 "My resources" menu

"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\\_\(statistics\)](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.

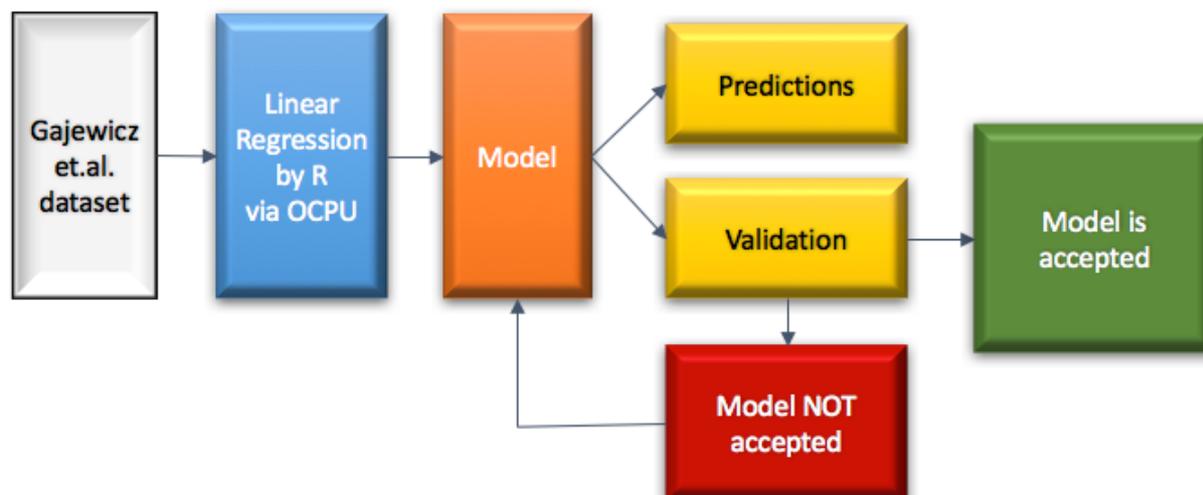


Figure 8 Workflow of the modelling cases

We will work on a dataset investigating the Cytotoxicity of Metal Oxides Nanoparticles by Gajewicz et.al.<sup>3</sup>

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.

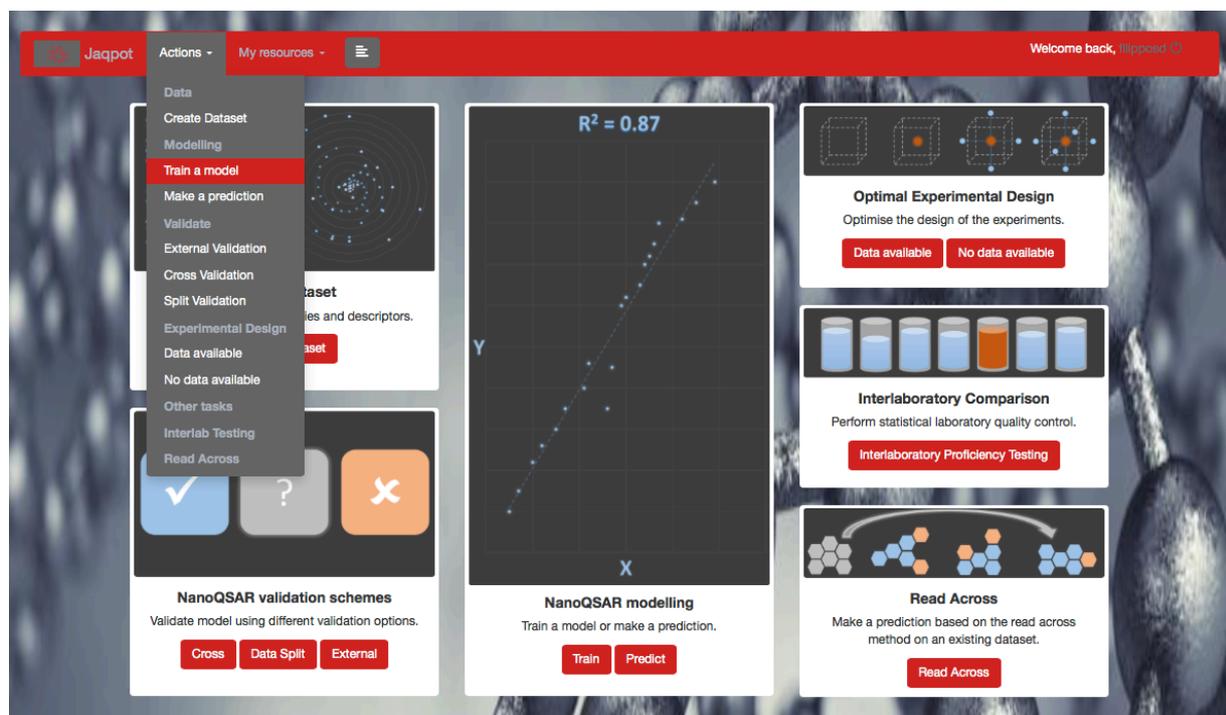


Figure 9 Train a model Action

Jaapot Actions - My resources - Welcome back, [User]

## Select dataset:

Example datasets:

Name	Title	Description
E20vuMNNThsp	Gajewicz et al - 10 Metal Oxide NPs	10 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.
KIL42dC8JSmp	Walkey et al - 56 Gold NPs	56 Gold NPs with 25 PhysChem descriptors, used for predicting cellular interaction.
PVvY25q3vd5O	Walkey et al - 56 Gold NPs	56 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.
UhysSLaF345pkI	Walkey et al - 84 Gold NPs	84 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.
class-dummy	Dummy Classification Dataset	This dataset contains classification data
interlab-dummy	Interlab testing dataset	This dataset contains data for interlab testing (3 measurements + uncertainty)

All Datasets:

Name	Title	Description
2Mh0UPm0PsyW3OllaykC	Walkey et al - 28 Gold NPs	28 Gold NPs with 25 PhysChem descriptors, used for predicting cellular interaction.
MmMsFPPYEItI	cxgsd	fgds

Previous 1 Next

Figure 10 Selecting a dataset for training

## Train model

Choose Algorithm

### Regression

- Lasso Regression
- Linear Regression
- PLS - with VIP scores
- MLR - Weka Implementation
- PLS - Weka Implementation
- SVM - Weka (LibSVM) Implementation
- R LM Algorithm

### Classification

- Bernoulli Naive Bayes
- Generalised Naive Bayes
- CMI Decision Tree
- ID3 Decision Tree
- Id3 - with MCI
- Multinomial Naive Bayes
- PLS - Weka Implementation

Previous 1 Next

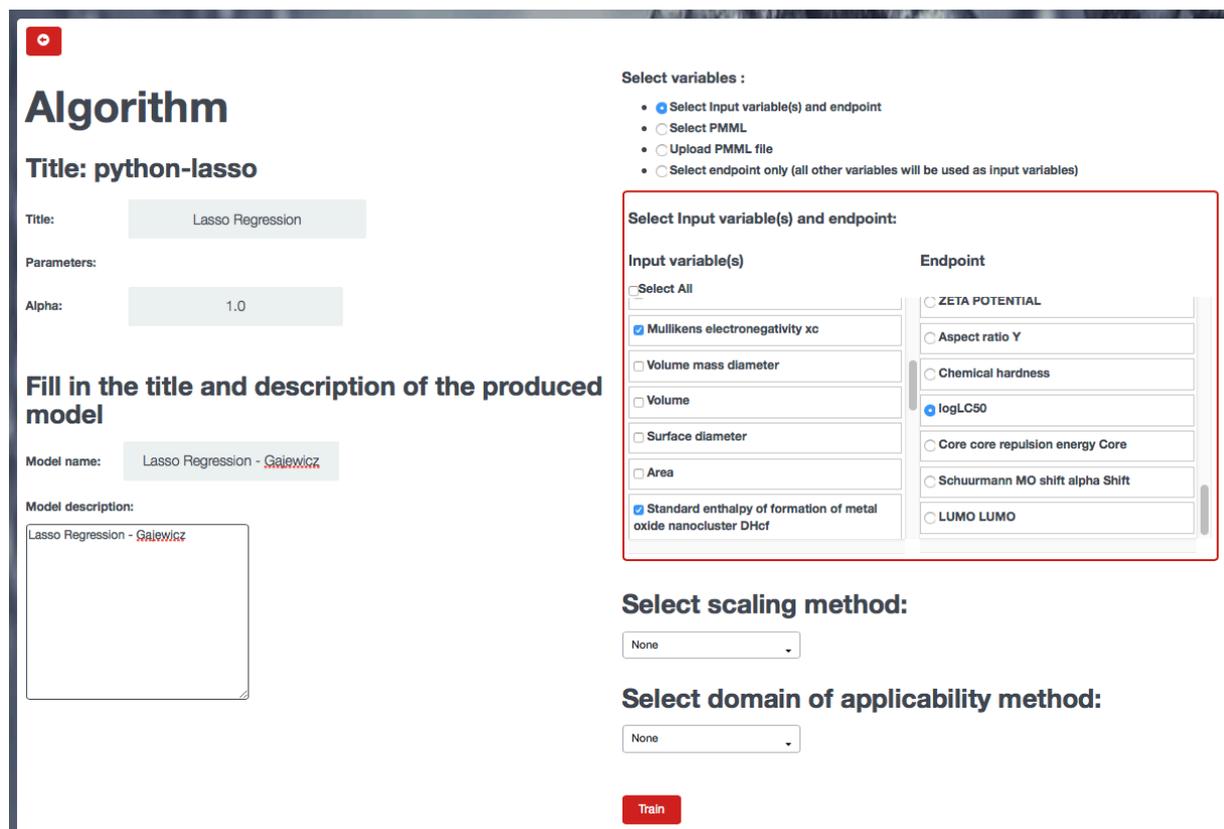
Next

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](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.



**Algorithm**

**Title:** python-lasso

**Title:** Lasso Regression

**Parameters:**

**Alpha:** 1.0

**Fill in the title and description of the produced model**

**Model name:** Lasso Regression - Galewicz

**Model description:**

Lasso Regression - Galewicz

**Select variables :**

- Select Input variable(s) and endpoint
- Select PMML
- Upload PMML file
- Select endpoint only (all other variables will be used as input variables)

**Select Input variable(s) and endpoint:**

Input variable(s)	Endpoint
<input type="checkbox"/> Select All	<input type="radio"/> ZETA POTENTIAL
<input checked="" type="checkbox"/> Mullikens electronegativity xc	<input type="radio"/> Aspect ratio Y
<input type="checkbox"/> Volume mass diameter	<input type="radio"/> Chemical hardness
<input type="checkbox"/> Volume	<input checked="" type="radio"/> logLC50
<input type="checkbox"/> Surface diameter	<input type="radio"/> Core core repulsion energy Core
<input type="checkbox"/> Area	<input type="radio"/> Schuurmann MO shift alpha Shift
<input checked="" type="checkbox"/> Standard enthalpy of formation of metal oxide nanocluster DHcf	<input type="radio"/> LUMO LUMO

**Select scaling method:**

None

**Select domain of applicability method:**

None

**Train**

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.



**Task:** Training on algorithm: python-lasso #T9zu8G8EgZnA

**Status:** COMPLETED

**Type:** TRAINING

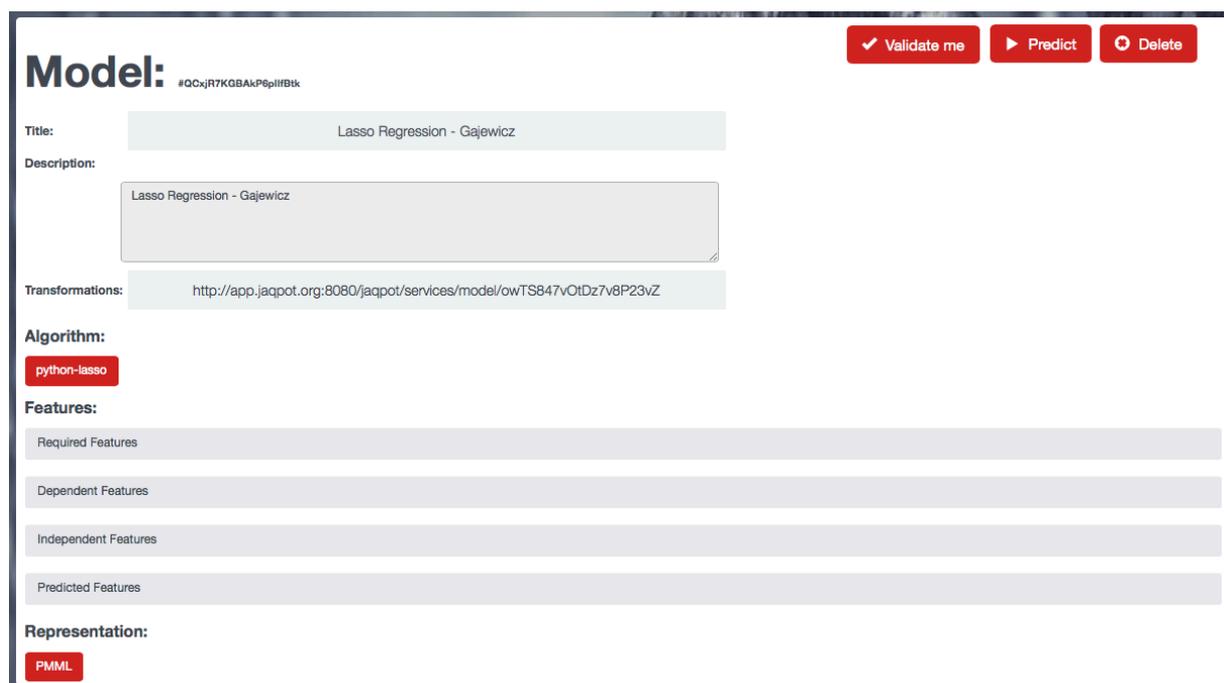
**Date:** 09/27/16

**Result:** [See result](#)

**Description:**

Training task using algorithm python-lasso

Figure 13 Train: Training on algorithm task page



**Model:** #QCxjR7KGBAkP6pIIfBtk

**Title:** Lasso Regression - Gajewicz

**Description:** Lasso Regression - Gajewicz

**Transformations:** <http://app.jaqpot.org:8080/jaqpot/services/model/owTS847vOIdz7v8P23vZ>

**Algorithm:** python-lasso

**Features:**

- Required Features
- Dependent Features
- Independent Features
- Predicted Features

**Representation:** PMML

Buttons: Validate me, Predict, Delete

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](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](http://jaqpot.org/predicted_dataset?name=5RCO6mRmFU5hiz4zMIAv&model=QCxjR7KGBAkP6pIIfBtk), Figure 16).

**Choose method:**

Select dataset.  
 Insert values.

**Select dataset for prediction:**

**Example Datasets:**

interlab-dummy	Interlab testing dataset	This dataset contains data for interlab testing (3 measurements + uncertainty)
interlab-dummy2	Interlab testing dataset	This dataset contains data for interlab testing (2 measurements + uncertainty)
kE0RiswkaCrg	Walkey et al - 28 Gold NPs	28 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.
ICUNPD99xGjgkU	Walkey et al - 84 Gold NPs	84 Gold NPs with 25 PhysChem descriptors, used for predicting cellular interaction.
vNxjUZMyv33ITM	Gajewicz et al - 18 Metal Oxide NPs	18 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.
xbR5AMG1rOBc	Gajewicz et al - 8 Metal Oxide NPs	8 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.
yzsAXE5rLPzz	Walkey et al - 28 Gold NPs	28 Gold NPs with 25 PhysChem descriptors, used for predicting cellular interaction.

**All Datasets:**

Name	Title	Description
5p3qHlvdBAvU	another corona dataset	This dataset contains corona data
RfbLMSC3tdXO	fdfsd	fdsd

Figure 15 Predict: selection of a dataset

**Predicted values of dataset** #5RC06mRmFUshiz4zMIAv

Search:

Compounds	
Al2O3	1.9911716068
Cr2O3	2.37343366203
Fe2O3	2.2659755716
La2O3	2.8371148572
NiO	2.5591079205
SnO2	2.39895124314
WO3	2.59304489739
Y2O3	2.2234674521

Previous 1 Next

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](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=QCxjR7KGBAkP6pIlfBtk](http://jaqpot.org/m_detail?name=QCxjR7KGBAkP6pIlfBtk), 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).

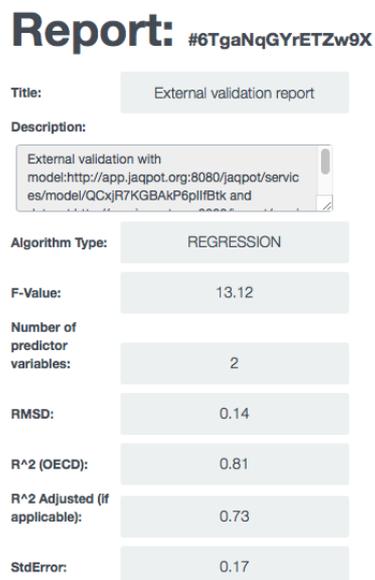


Figure 17 Model validation report

Please note the following on the definition of  $R^2$  ([https://en.wikipedia.org/wiki/Coefficient\\_of\\_determination](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.