



ENM TUTORIALS

JAQPOT QUATTRO
USER INTERFACE TUTORIAL
Nano-QSAR
Modelling infrastructure

RELEASE DATE:	01/06/16
USE:	How to use the JaqPot Quattro User Interface to create and validate NanoQSAR models
VERSION:	V.1.0
MAIN AUTHOR:	Georgios Drakakis
PARTNER:	NTUA
CONTACT DETAILS:	hsarimv@central.ntua.gr
AUTHORS:	G. Drakakis, C. Chomenidis, G. Tsiliki, P. Doganis, E. Anagnostopoulou, H. Sarimveis
LICENCE:	CC-BY 4.0



TABLE OF CONTENTS

- [1. INTRODUCTION](#)
- [2. GENERAL](#)
 - [2.1 LOG IN](#)
 - [2.2 CREATE ACCOUNT](#)
 - [2.3 MENU OPTIONS AVAILABLE](#)
- [3. CREATE DATASET](#)
- [4. TRAIN A MODEL](#)
- [5. MAKE A PREDICTION](#)
- [6. VALIDATION](#)
 - [6.1 EXTERNAL VALIDATION](#)
 - [6.2 CROSS VALIDATION](#)
 - [6.3 TRAINING SET SPLIT VALIDATION](#)
- [7. MY RESOURCES](#)
- [8. FUTURE FUNCTIONALITIES](#)
- [9. ACKNOWLEDGMENTS](#)
- [10. REFERENCES](#)
- [11. KEYWORDS](#)
 - [ENM TUTORIALS](#)

1. INTRODUCTION

This document provides a tutorial for the User Interface (UI) made available by the Jaqpot Quattro (JQ) modelling infrastructure. The resource has been made available at <http://www.jaqpot.org/>. At this location, users may create datasets containing nanoparticles and properties, apply PMML transformations, create and validate predictive NanoQSAR models and use the models for making predictions. Several other functionalities (optimal experimental design, interlaboratory testing, read-across methods) will be available in the next release of the UI.

2. GENERAL

The UI of JQ can be seen in Figure 1. It contains two options on the top left for the user, “Sing In” and “Create Account”. It also shows the six main menu options available, namely “Create Dataset”, “NanoQSAR modelling”, “NanoQSAR validation schemes”, Optimal Experimental Design, “Interlaboratory Comparison” and “Read Across”. For the current release, the first three options are made available, which comprise creating a dataset, training a model and making predictions, as well as cross-, split- and external validation. Should the user click on one of these options whilst not signed in, they will be prompted to log onto the system. On the bottom of the page, next to the eNanoMapper logo, the user can view the source (<https://github.com/KinkyDesign>) and documentation (<http://test.jaqpot.org:8000/documentation>), as well as report an issue on GitHub for the JQ UI (<https://github.com/KinkyDesign/JaqpotQuattroUI/issues>).

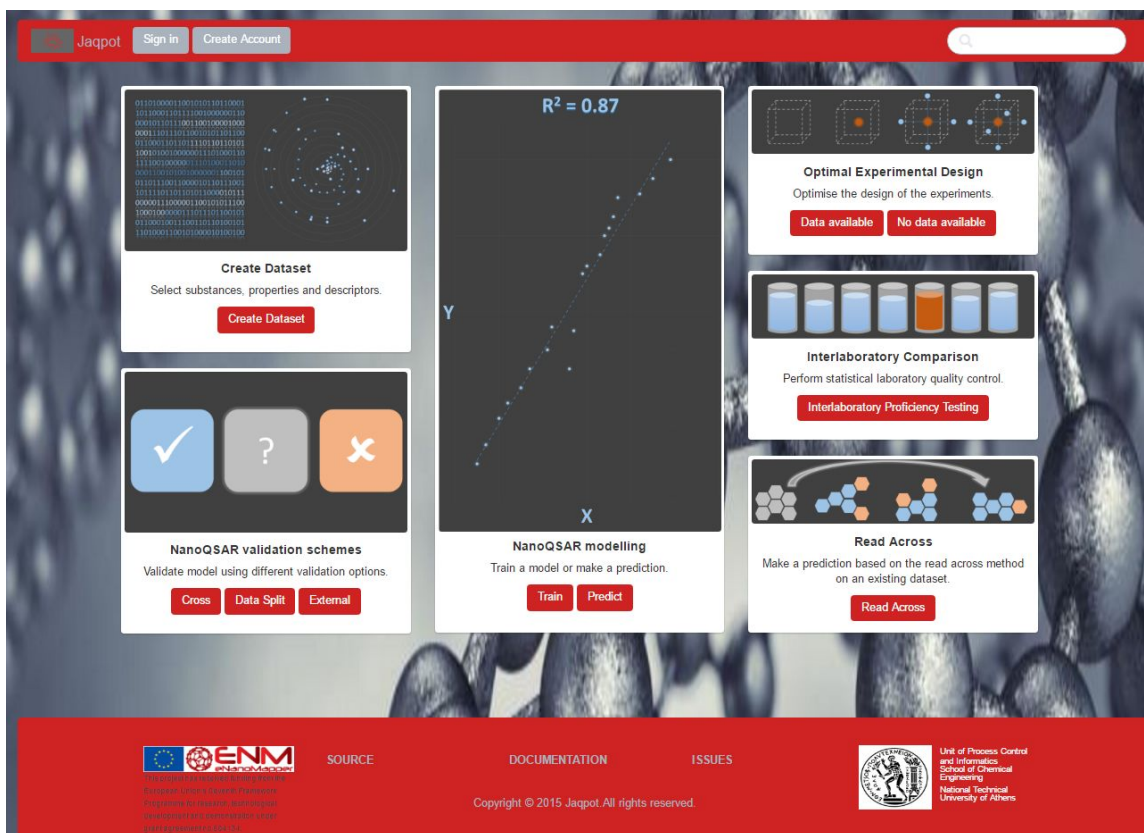


Figure 1: Main interface of Jaqpot UI available at <http://www.jaqpot.org/>.

2.1 LOG IN

The Login screen is shown in Figure 2. Here users must provide their credentials in order to be able to use the services. If users do not have an account, they must register using the “Create Account” button.

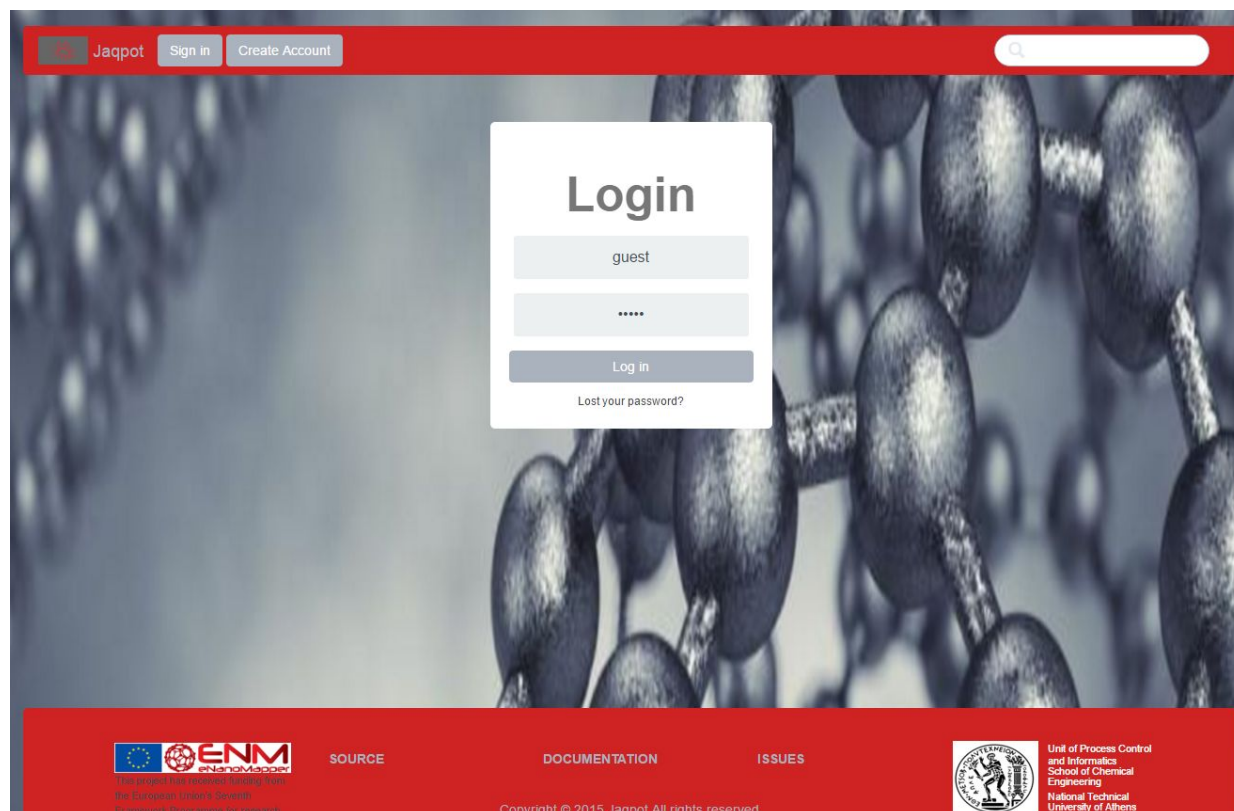
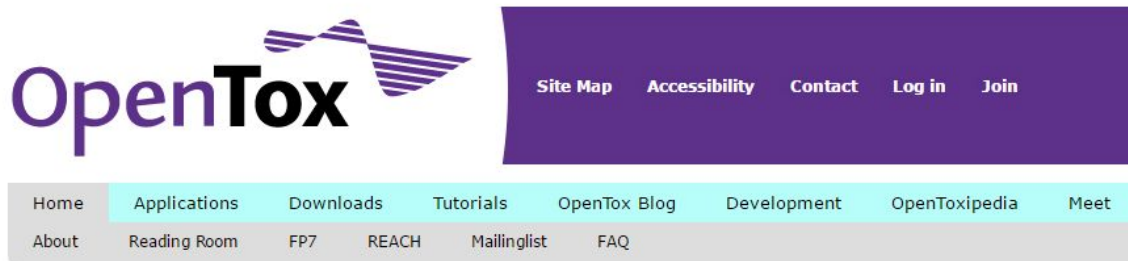


Figure 2: Login screen of Jaqpot UI

2.2 CREATE ACCOUNT

This option is made available for users not already registered. Should users click on this button, they will be redirected to the OpenTox website registration form, shown in Figure 3. Here, the user must provide his/her details, such as full name, email address and institution. Once the account is successfully created, users are notified by email at the address provided on the registration form.



You are here: Home

Registration Form

Note: Please do not use any accents, umlauts or special characters (e.G.: ê,á,ò,ü,ß...) in this form.

Personal Details

Full Name ■
Enter full name, eg. John Smith.

User Name ■
Enter a user name, usually something like 'jsmith'. No spaces or special characters. Usernames and passwords are case sensitive, make sure the caps

E-mail ■
Enter an email address. This is necessary in case the password is lost. We respect your privacy, and will not give the address away to any third parties

Figure 3: OpenTox registration form

2.3 MENU OPTIONS AVAILABLE

This section provides an overview of the options and methods available in JQ. Each service will be discussed in detail later in this document. Apart from the main options mentioned, two more categories appear next to the Jaqpot logo on the top menu bar, namely “Actions” and “My resources”. By clicking on “Actions” it can be seen that the main options have been made available in the form of categories and subcategories. For example, under “Validate” we can see the external, cross- and training set split validation options. This is demonstrated in Figure 4.

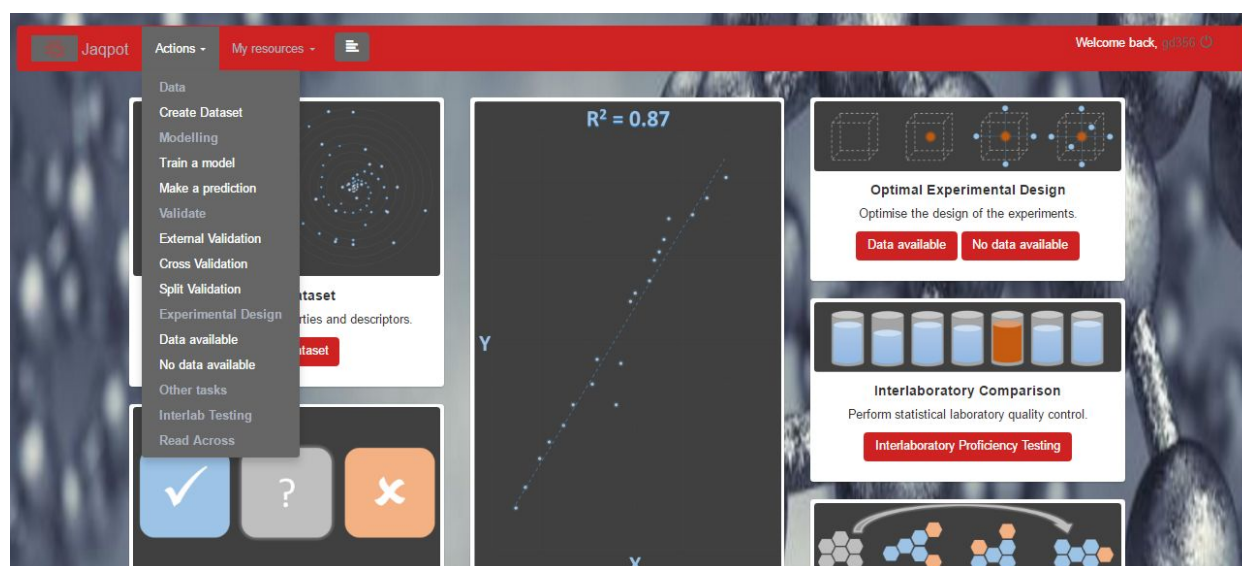


Figure 4: Options under Jaqpot “Actions” Menu

Similarly, additional options appear upon clicking on any of the remaining menu buttons. More specifically, under “My resources”, the user can view their own models, algorithms, datasets, reports, Bibtex entries and features. This is shown in Figure 5. In addition to the user’s resources, other models and datasets have been made available under the tags “Example Datasets”, “Example Models” etc. which will be shown later in this tutorial.

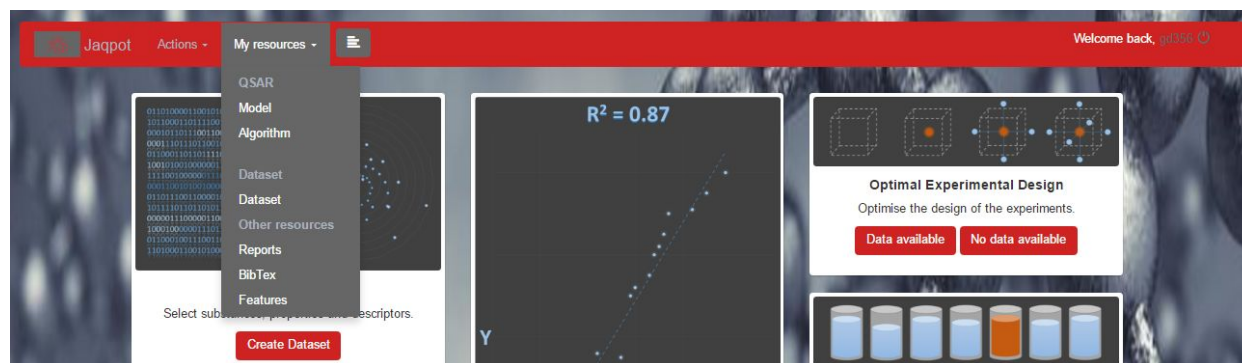


Figure 5: Options under Jaqpot “My Resources” Menu

3. CREATE DATASET

The first available menu option for the user is for creating a dataset on which to perform modelling or other additional tasks. This service allows the retrieval of data available on the eNanoMapper data repository (<https://apps.ideaconsult.net/enanomapper/>) by selecting a dataset creator (“Substance owner”) or manually inputting the substance owner URI given by AMBIT, shown respectively in Figures 6 and 7.

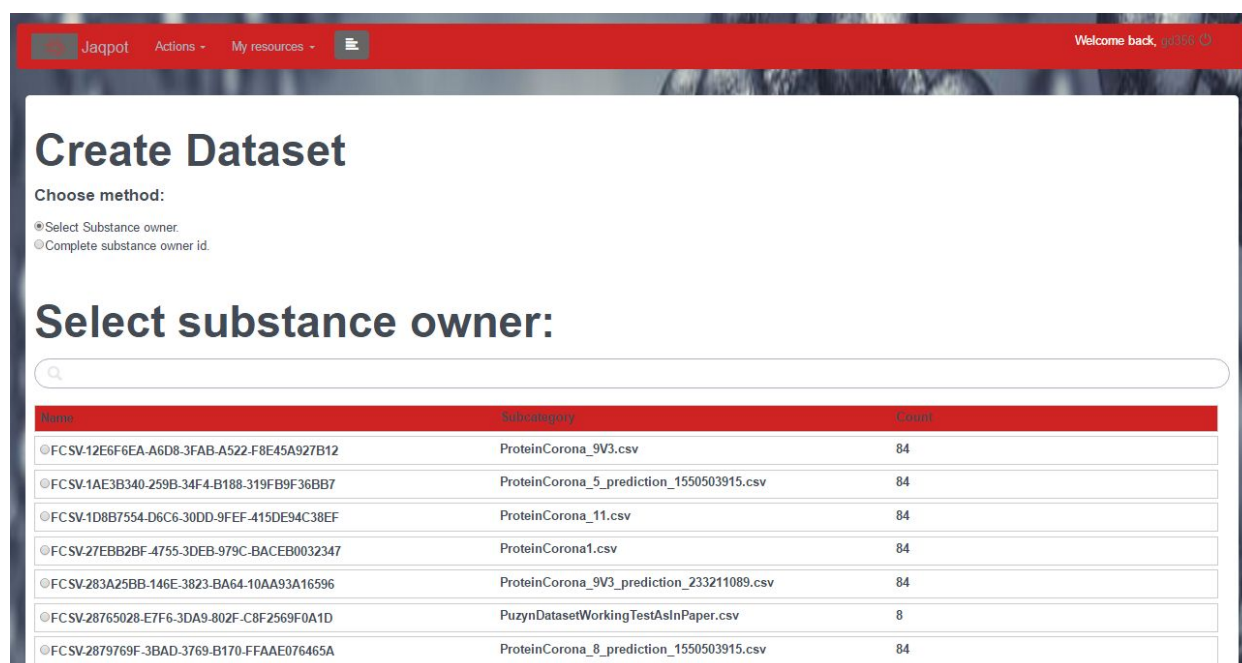


Figure 6: Create dataset by selecting Substance Owner from list

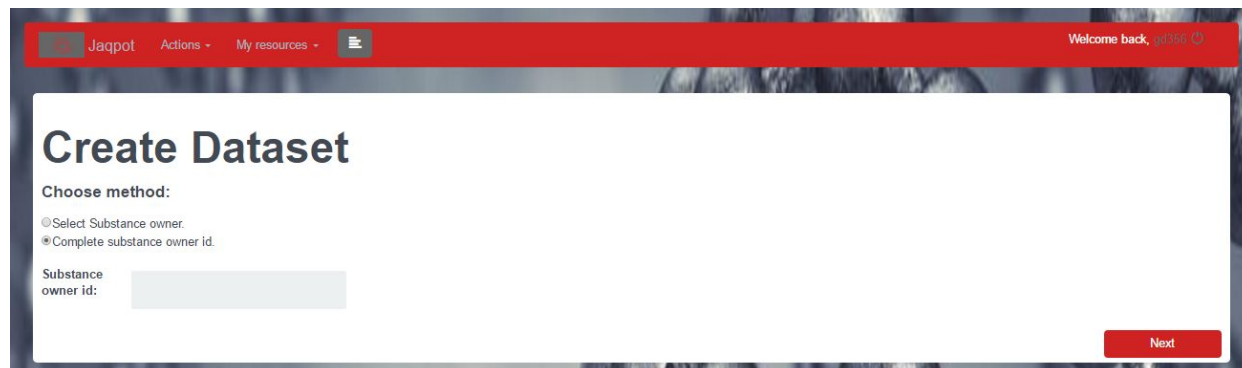


Figure 7: Create dataset by manually completing Substance Owner field

The next pages allow the user to select particular nanoparticles (“substances”) to include in their dataset, their experimental properties (single or by category) and descriptors to be calculated. The current version of JQ supports calculation of MOPAC descriptors should a PDB file be available and image analysis descriptors should an image file be available. For the next release, we plan to add functionalities for calculating CDK descriptors and GO descriptors should proteomics data are available. These screens are shown in Figures 8, 9, and 10. At the last screen, users must provide dataset name and description.

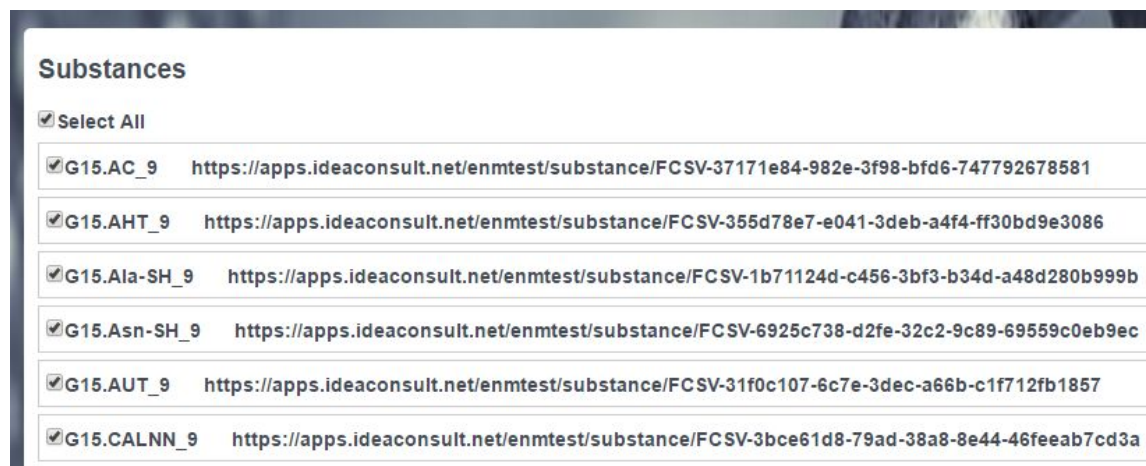
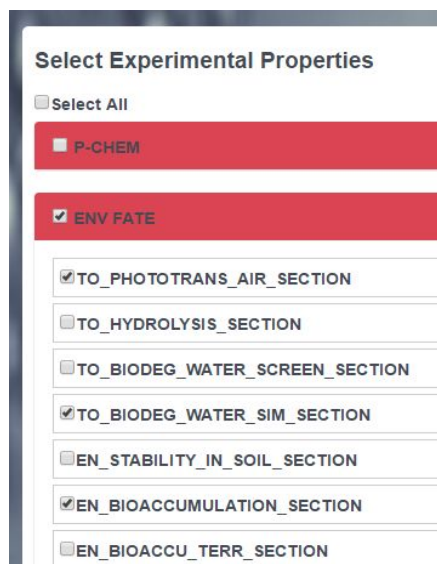


Figure 8: Screenshot of nanoparticle selection from Substance owner



Select Experimental Properties

Select All

P-CHEM

ENV FATE

TO_PHOTOTRANS_AIR_SECTION

TO_HYDROLYSIS_SECTION

TO_BIODEG_WATER_SCREEN_SECTION

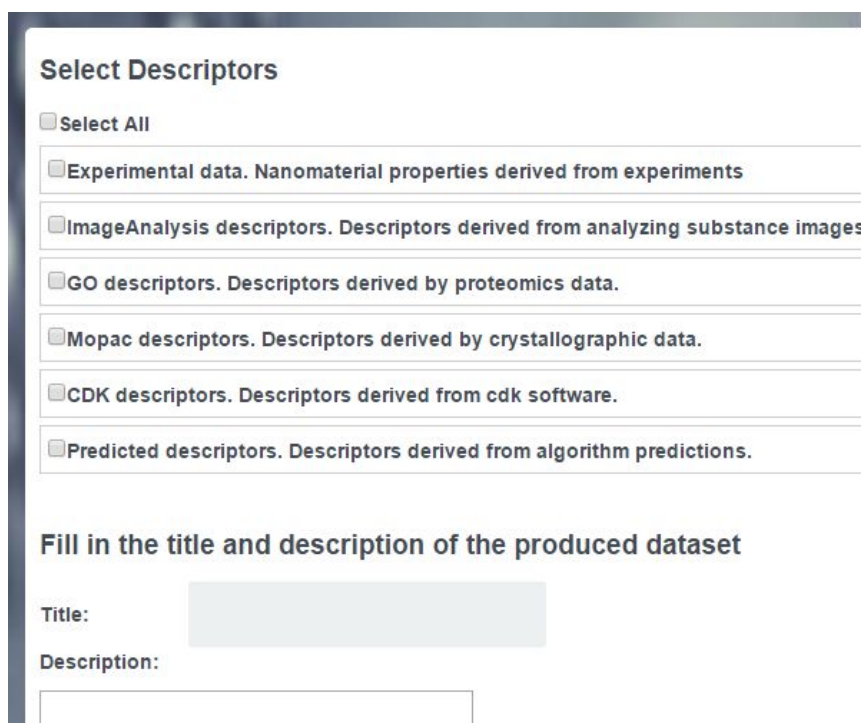
TO_BIODEG_WATER_SIM_SECTION

EN_STABILITY_IN_SOIL_SECTION

EN_BIOACCUMULATION_SECTION

EN_BIOACCU_TERR_SECTION

Figure 9: Screenshot of Experimental Property selection to include in dataset creation



Select Descriptors

Select All

Experimental data. Nanomaterial properties derived from experiments

ImageAnalysis descriptors. Descriptors derived from analyzing substance images

GO descriptors. Descriptors derived by proteomics data.

Mopac descriptors. Descriptors derived by crystallographic data.

CDK descriptors. Descriptors derived from cdk software.

Predicted descriptors. Descriptors derived from algorithm predictions.

Fill in the title and description of the produced dataset

Title:

Description:

Figure 10: Screenshot of additional calculated descriptors selection process available in JQ and the fields for dataset title and description.

Users are then provided with a task which they can monitor. Once the status changes to “Completed” users can view their dataset by clicking on “See Result”. The unique identifier assigned to the dataset will also be provided (seen in Figure 11).

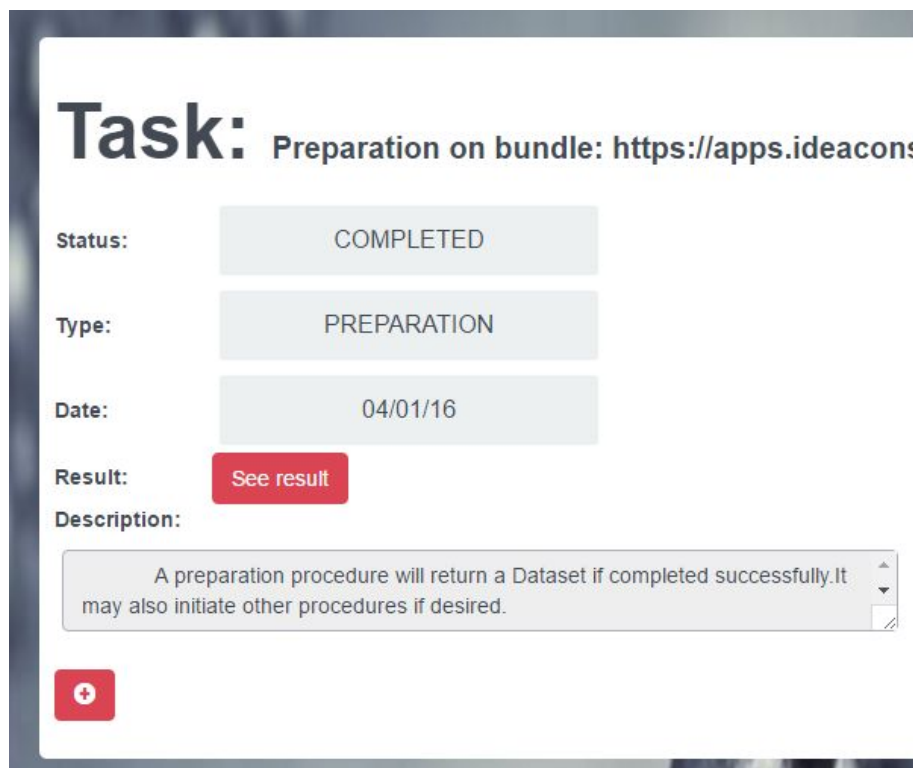


Figure 11: Task monitoring for dataset creation.

4. TRAIN A MODEL

Once users have at least one dataset created or wish to use one of the example datasets provided by JQ, they may use it to train a model. The “Train” process starts with the selection of a dataset (Figure 12), proceeds to the algorithm selection (Figure 13) and finally allows parameterization (Figure 14), which ranges from setting the available algorithm parameters all the way to variable selection, normalization of the dataset values and setting an algorithm for defining the domain of applicability. The current JQ release contains R, Python or WEKA implementations of all major statistical and machine learning regression and classification algorithms, but additional algorithms will be available in future releases.

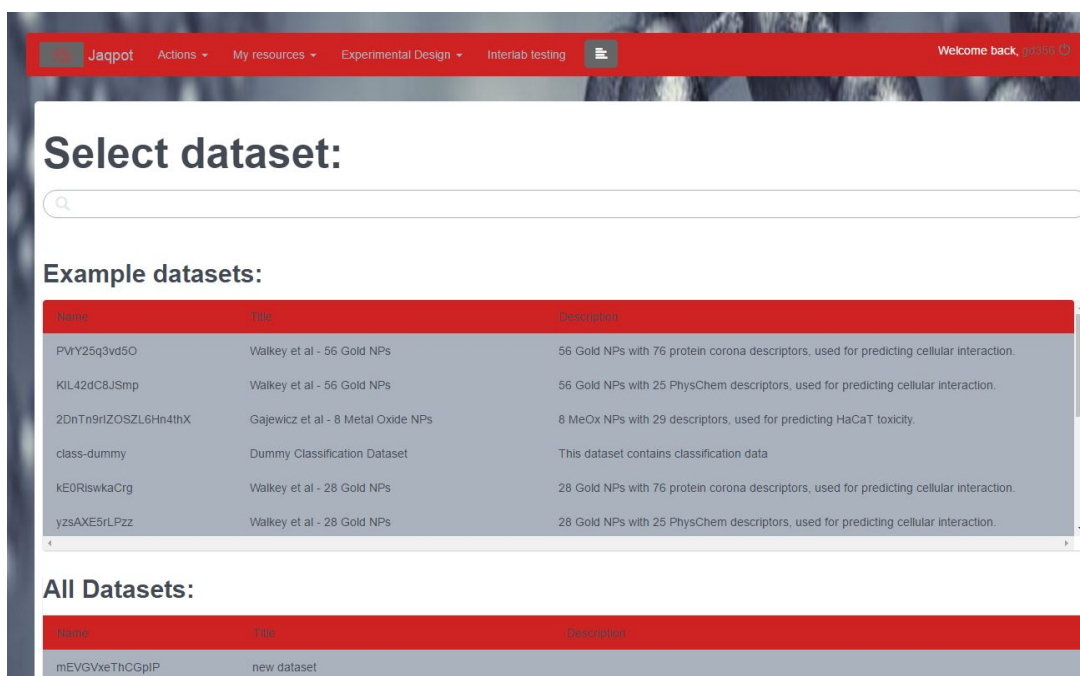


Figure 12: Screenshot of dataset selection for training a model.

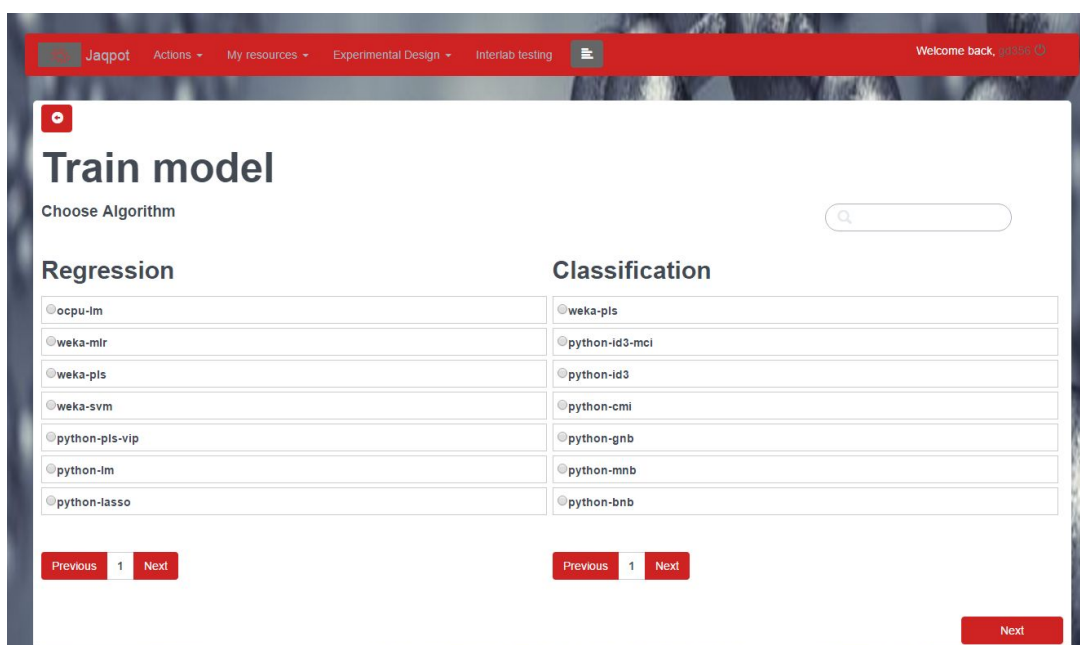
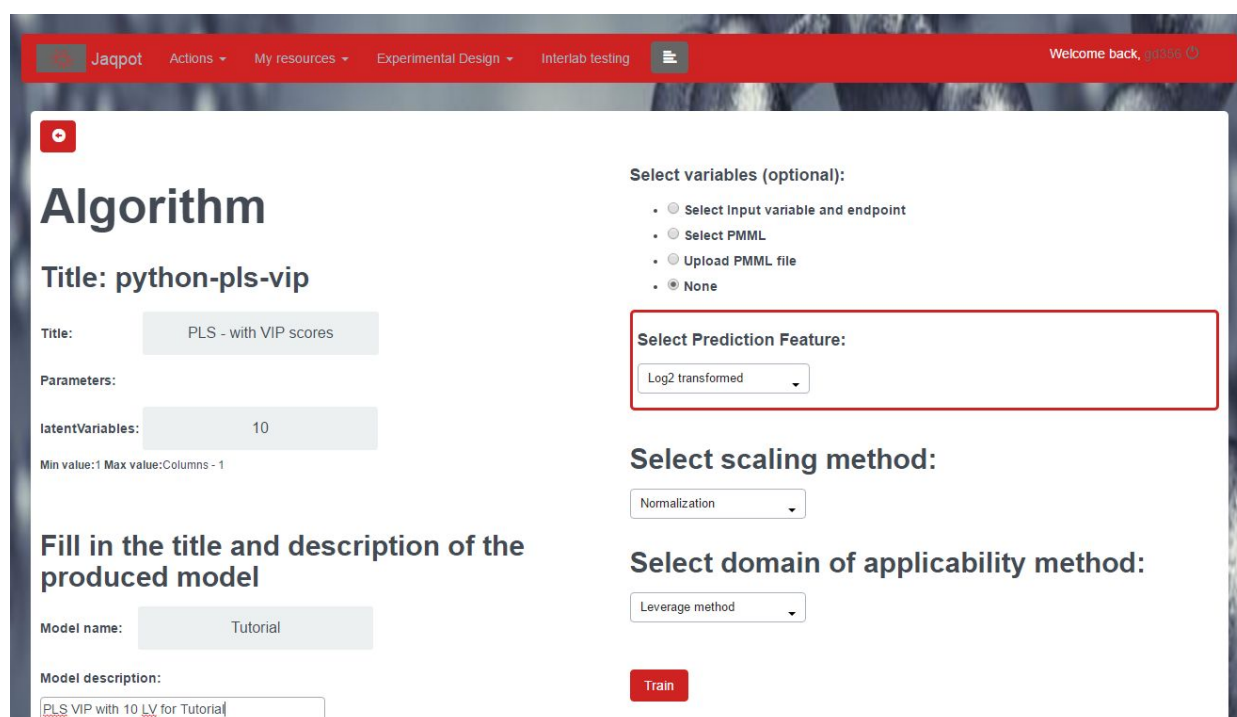


Figure 13: Screenshot of algorithm selection for training a model

The final page before training the model shows all necessary details for pre-processing and parameterizing the algorithm. More specifically, it can be seen in Figure 14 on the left that under the algorithm title (“PLS with VIP scores”), users may set the parameter “latentVariables” for the algorithm. A help note under the parameter provides the user with the correct range/type depending on the parameter. Users are also prompted to provide a model name and description. On the right-hand side, users may select variables and perform PMML transformations either by selecting an existing file or providing their own. PMML files are XML files which can represent different types of

machine learning models. They contain a data dictionary and transformation dictionary which comprise the input. The latter can be used to define transformations such as addition, subtraction, derive absolute values, whilst the former can be used to perform variable selection. A PMML example is shown in Figure 15, which defines the subtraction, division and absolute value transformations for two different measurements of Zeta Potential. In order to create such files users must know the URI of the features on which they want to apply transformations, which can be derived from viewing the dataset. The prediction feature must be provided separately from a drop-down menu. By clicking on the “Select Input variable(s) and endpoint” choice, all the variables appear as input and endpoint candidates. The user may select some or all of them as input variables and only one as the endpoint. When an end-point is selected, it is automatically deselected from the input variables. If a user chooses the “None” option, they are prompted to select only the endpoint, whilst all other variables are considered as input.

Finally, users have the option to scale/normalize their data and select an algorithm for defining the domain of applicability (DOA) of their model (currently the leverage method is available). As before, users can see when the training has been completed by following the task progress (Figure 16).



The screenshot shows a web interface for training a model. The main heading is "Algorithm". Below it, the title is "python-pls-vip". There are several input fields and dropdown menus:

- Title:** PLS - with VIP scores
- Parameters:** (empty field)
- Latent Variables:** 10
- Min value:** 1 **Max value:** Columns - 1
- Model name:** Tutorial
- Model description:** PLS VIP with 10 LV for Tutorial
- Select variables (optional):**
 - Select Input variable and endpoint
 - Select PMML
 - Upload PMML file
 - None
- Select Prediction Feature:** Log2 transformed
- Select scaling method:** Normalization
- Select domain of applicability method:** Leverage method
- Train** button

Figure 14: Screenshot of pre-processing and parameterization for training a model.

```

<PMML version="4.0"
  xsi:schemaLocation="http://www.dmg.org/PMML-4_0
    http://www.dmg.org/v4-0/pmml-4-0.xsd"
  xmlns="http://www.dmg.org/PMML-4_0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<DataDictionary numberOfFields="4" >
  <DataField
name="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/7F8B3FB820
19B1CCF8A8C3FD2B5A2DACBDDDB832/3ed642f9-1b42-387a-9966-dea5b91e5f8a" optype="continuous" dataType="double"
/>

  <DataField
name="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/06399AE160
9F65589E8D7C6DECF4A7E8565336CA/3ed642f9-1b42-387a-9966-dea5b91e5f8a" optype="continuous" dataType="double" />

</DataDictionary>
<TransformationDictionary>
  <DerivedField dataType="double" name="zp_ch" optype="categorical">
    <Apply function="-">
      <FieldRef
field="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/7F8B3FB8201
9B1CCF8A8C3FD2B5A2DACBDDDB832/3ed642f9-1b42-387a-9966-dea5b91e5f8a"/>
      <FieldRef
field="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/06399AE1609
F65589E8D7C6DECF4A7E8565336CA/3ed642f9-1b42-387a-9966-dea5b91e5f8a"/>
    </Apply>
  </DerivedField>
  <DerivedField dataType="double" name="zp_rel" optype="categorical">
    <Apply function="/">
      <FieldRef
field="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/7F8B3FB8201
9B1CCF8A8C3FD2B5A2DACBDDDB832/3ed642f9-1b42-387a-9966-dea5b91e5f8a"/>
      <FieldRef
field="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/06399AE1609
F65589E8D7C6DECF4A7E8565336CA/3ed642f9-1b42-387a-9966-dea5b91e5f8a"/>
    </Apply>
  </DerivedField>
  <DerivedField dataType="double" name="zp_synth_mag" optype="categorical">
    <Apply function="abs">
      <FieldRef
field="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/7F8B3FB8201
9B1CCF8A8C3FD2B5A2DACBDDDB832/3ed642f9-1b42-387a-9966-dea5b91e5f8a"/>
    </Apply>
  </DerivedField>
  <DerivedField dataType="double" name="zp_serum_mag" optype="categorical">
    <Apply function="abs">
      <FieldRef
field="https://apps.ideaconsult.net/enmtest/property/P-CHEM/ZETA_POTENTIAL_SECTION/ZETA+POTENTIAL/06399AE1609
F65589E8D7C6DECF4A7E8565336CA/3ed642f9-1b42-387a-9966-dea5b91e5f8a"/>
    </Apply>
  </DerivedField>
</TransformationDictionary>
</PMML>

```

Figure 15: Example of valid PMML file containing transformations

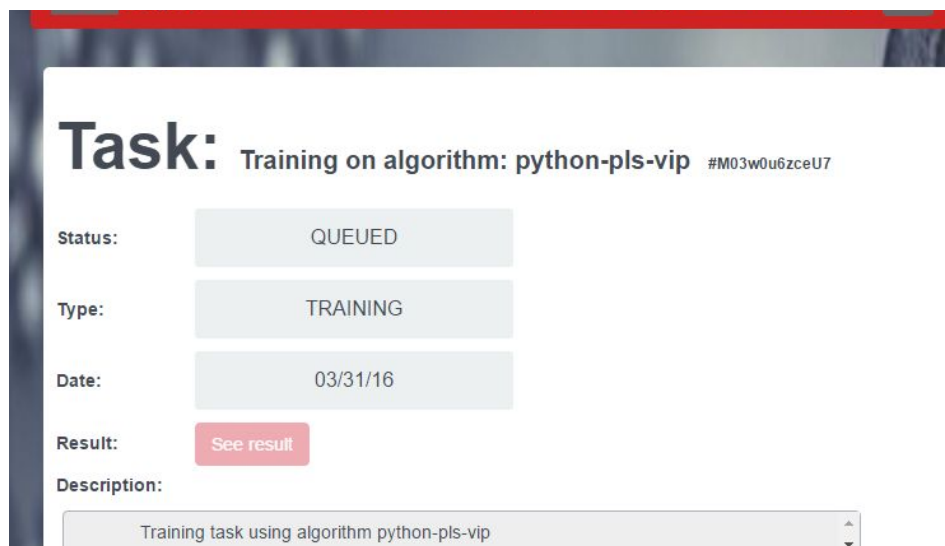


Figure 16: Monitoring the training task progress

Once the task is complete, by clicking on “See result” the user is shown the model screen (screenshot in Figure 17), where the title, description and algorithm used to create are shown, along with additional clickable features such as required features etc. These are more useful to the advanced user, as they are information mostly for the algorithm, for example, variables left after variable elimination. By clicking on the PMML button, the user may see the PMML representation of the model, if the type of model is supported by the PMML format Furthermore, the model unique ID is shown along with options to Validate, Predict and Delete on the top right of the screen. The prediction and validation options will be described in the next sections of this document.

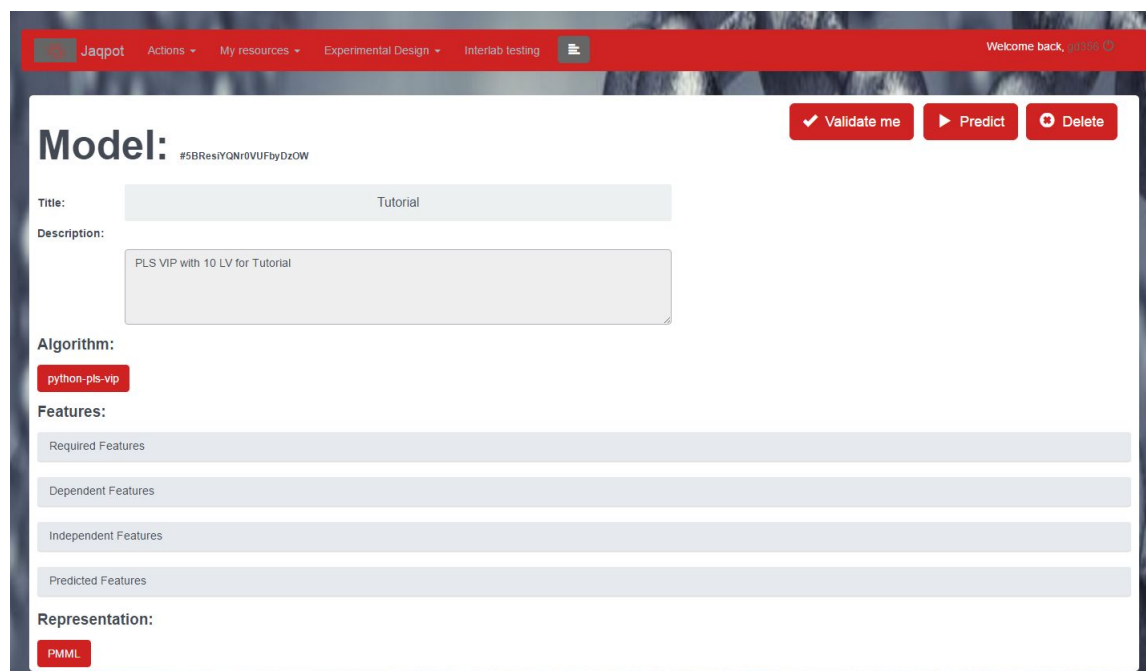
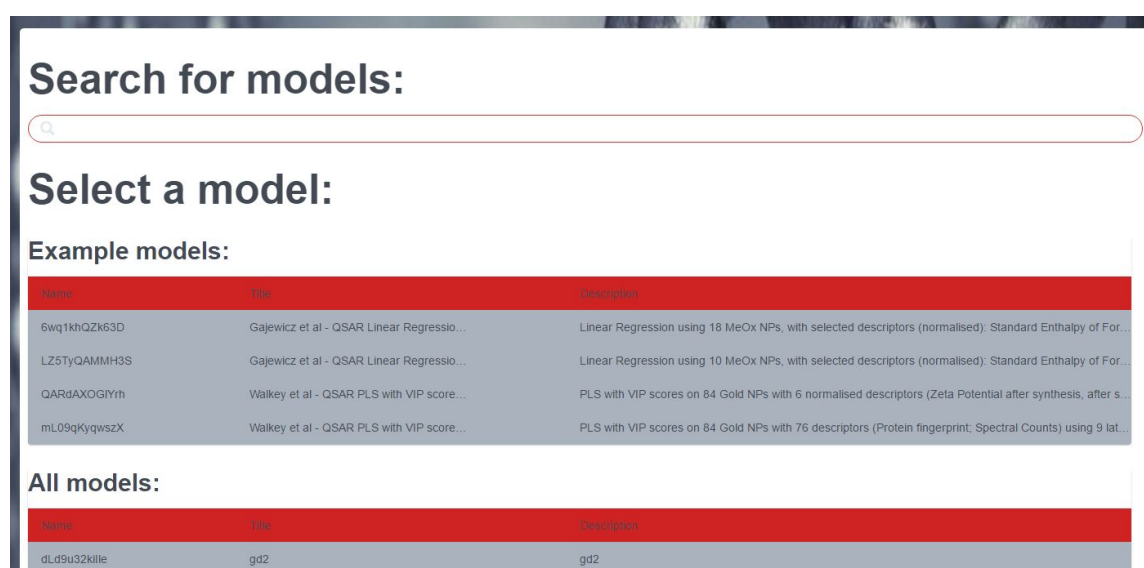


Figure 17: Screenshot of model screen options after training.

5. MAKE A PREDICTION

By clicking on the “Predict” button, a user may make a prediction on a dataset, given that a model has been created, or that the user wishes to use one of the example models found in the literature made available by Jaqpot. Therefore, the user must select a model (Figure 18) and then a dataset on which to perform the prediction. Here, the user may select an existing dataset as shown in Figure 19. An additional option “Insert Values” is provided which can also be seen on the screenshot shown in Figure 19, where users can upload their own dataset or manually input values for each of the variables used by the model in order to make a prediction. The option is shown in Figure 20.



Search for models:

Select a model:

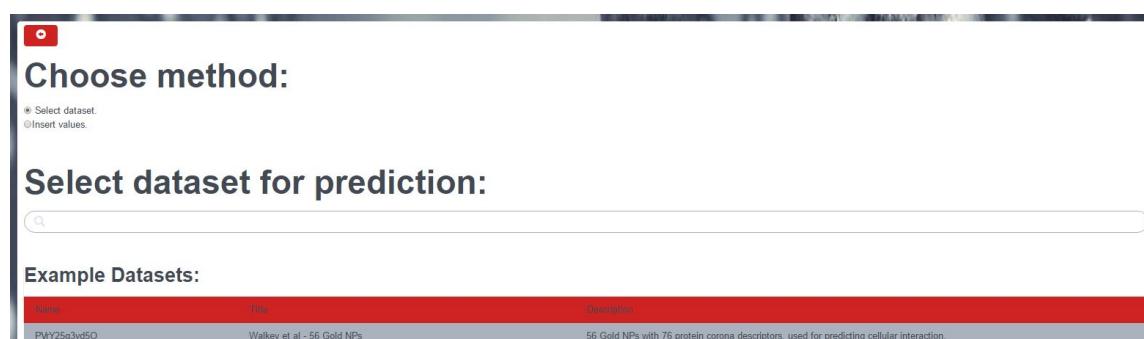
Example models:

Name	Title	Description
6wq1khQZk63D	Gajewicz et al - QSAR Linear Regressio...	Linear Regression using 18 MeOx NPs, with selected descriptors (normalised): Standard Enthalpy of For...
LZ5TyQAMMH3S	Gajewicz et al - QSAR Linear Regressio...	Linear Regression using 10 MeOx NPs, with selected descriptors (normalised): Standard Enthalpy of For...
QARdAXOGiYrh	Walkey et al - QSAR PLS with VIP score...	PLS with VIP scores on 84 Gold NPs with 6 normalised descriptors (Zeta Potential after synthesis, after s...
mL09gKyqwszX	Walkey et al - QSAR PLS with VIP score...	PLS with VIP scores on 84 Gold NPs with 76 descriptors (Protein fingerprint, Spectral Counts) using 9 lat...

All models:

Name	Title	Description
dLd9u32kille	gd2	gd2

Figure 18: Select model for prediction



Choose method:


Select dataset.
 Insert values.

Select dataset for prediction:

Example Datasets:

Name	Title	Description
PWY25q3vd5O	Walkey et al - 56 Gold NPs	56 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.

Figure 19: Select dataset for prediction



Choose method:

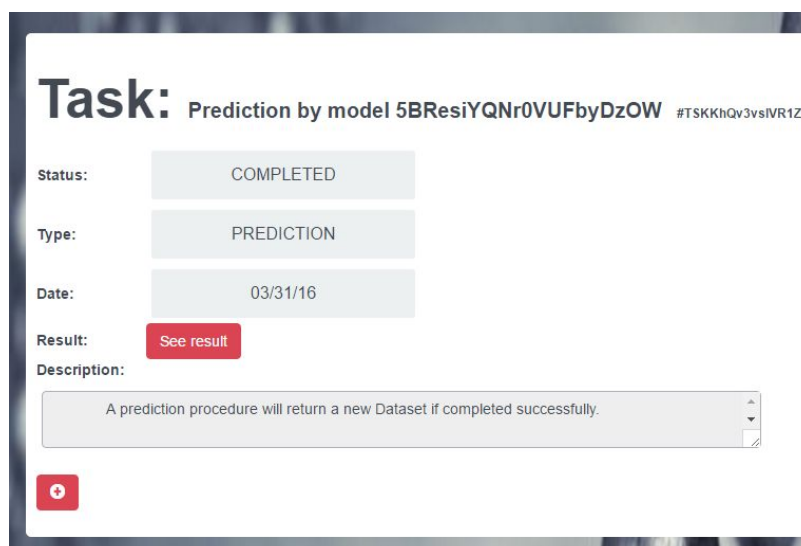
Select dataset
 Upload dataset.

	Mopac	P08567	P10720	P07996	O95445	P03952	P03951	P03950	P02741	P04004	P01011	P04003	POC0L5	P05452	P01857	P02748	P20851	P01019	P02788
1	Choose file	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Predict

Figure 20: Upload dataset or manually input variable values for prediction

As above, a task is created and the user can monitor its progress (Figure 21). By clicking on “See Result” the user can see the predicted values (Figure 22), as well as whether the instances are within the DOA of the model (assuming that the option was checked when the model was created). For the leverage method which is currently the only available option, a value of 0 means that the prediction is outside the DOA of the model, while any positive value means the prediction is reliable as the query instance is within the DOA of the model.



Task: Prediction by model 5BResiYQNr0VUFbyDzOW #TSKKhQv3vsIVR1Z

Status: COMPLETED
 Type: PREDICTION
 Date: 03/31/16
 Result: **See result**
 Description: A prediction procedure will return a new Dataset if completed successfully.

Figure 21: Monitoring of prediction task

Predicted values of dataset

#AmFv1W83ZGCrceLDbog

Search:

Compounds	https://apps.chemsonline.net/enanmapper/property_TOXUNKNOWN_TOXDTY_SECTION/qspectra/stored/406842F64928A0F4004A528C4731813044A4833a4a4316-1042-037a-3885-0a3b391a3933_predicted	Leverage
G15.Ala-SH_1	-5.67655971805	0.355720031225
G15.CALNN_1	-6.42881857352	0.0
G15.DDT@BDHDA_1	-5.87267855933	0.0
G15.DDT@ODA_1	-4.9199332973	0.0
G15.DTNB_1	-7.29504611551	0.0
G15.HDA_1	-0.830575861201	0.0
G15.MBA_1	-3.40646452808	0.13463168034
G15.MHA_1	-4.9237630666	0.0
G15.MSA_1	-2.8829083008	0.0
G15.NT@OCA_1	-7.04739320396	0.0
G15.NT@PSMA-EDA_1	-4.89746803624	0.0
G15.PAH-SH_1	-1.40573896685	0.0
G15.Phe-SH_1	-6.16834018413	0.498858361758

Figure 22: Prediction outcome screen

Once the prediction has been retrieved, users may click on the plus sign under “Predicted Values of dataset”, which will show the full dataset, including the original variables, or those created by scaling/normalization and PMML transformations. This is shown in Figure 23. By clicking on the additional plus sign (mouse-over reads “dataset info”) users may see the full dataset used for prediction. This information includes the publication DOI, the contributors (authors) the journal information and keywords, annotated as “subjects”. This is shown in Figure 24.

Dataset: Walkey et al - 28 Gold NPs #AmFv1W83ZGCrceLDbog

Search:

Compounds	qspectra https://apps.chemsonline.net/enanmapper/property_TOXPROTEOMIC_SECTION/qspectra/stored/031448A38133405C100CEB73220AA40AB160B8783a8f4279-1042-3873-9395-0a3b391a3933	Standardized https://apps.chemsonline.net/enanmapper/property_TOXPROTEOMIC_SECTION/qspectra/stored/031448A38133405C100CEB73220AA40AB160B8783a8f4279-1042-3873-9395-0a3b391a3933
G15.Ala-SH_1	-0.803848949051	0.650799787818
G15.CALNN_1	-0.467912074821	-0.64082584471
G15.DDT@BDHDA_1	0.53989854787	1.41879340716
G15.DDT@ODA_1	2.89145666748	-0.64082584471
G15.DTNB_1	-0.803848949051	-0.920096251743
G15.HDA_1	-0.467912074821	-0.291737835918
G15.MBA_1	-0.467912074821	-1.30405306141
G15.MHA_1	0.53989854787	-1.12954905702
G15.MSA_1	-0.131975200591	-0.920096251743
G15.NT@DCA_1	-0.131975200591	0.860252593093

Figure 23: Full dataset (attribute values, transformations etc.) with predictions

Dataset: Walkey et al - 28 Gold NPs #AmFvx1W83ZGCrceLDbog

Doi: 10.1021/nn406018q

Contributors: Yoram Cohen, Hongbo Guo, Rong Liu, Carl D. Walkey, Fayi Song, Andrew Emili, Warren C. W. Chan, Jonathan B. Olsen, D. Wesley H. Olsen

Publishers: <http://informahealthcare.com/nan>, Nanotoxicology, informa healthcare

Subjects: liquid chromatography tandem mass spectrometry, protein corona, cell uptake, nanomedicine, nanobiotechnology, structure-activity model, quantitative proteomics

Compounds	Standardized
	https://apps.ideaconsult.net/enmtest/property/TOX/PROTEOMICS_SECTION/Spectral+counts/03184EA3833640EC100CE871b42-387a-9368-dca36b1e5f9a/B9A084
G15.Ala-SH_1	-0.803848949051
G15.CALNN_1	-0.467912074821
G15.DDT@BDHDA_1	0.53989854787
G15.DDT@ODA_1	2.89145666748
G15.DTNB_1	-0.803848949051
G15.HDA_1	-0.467912074821

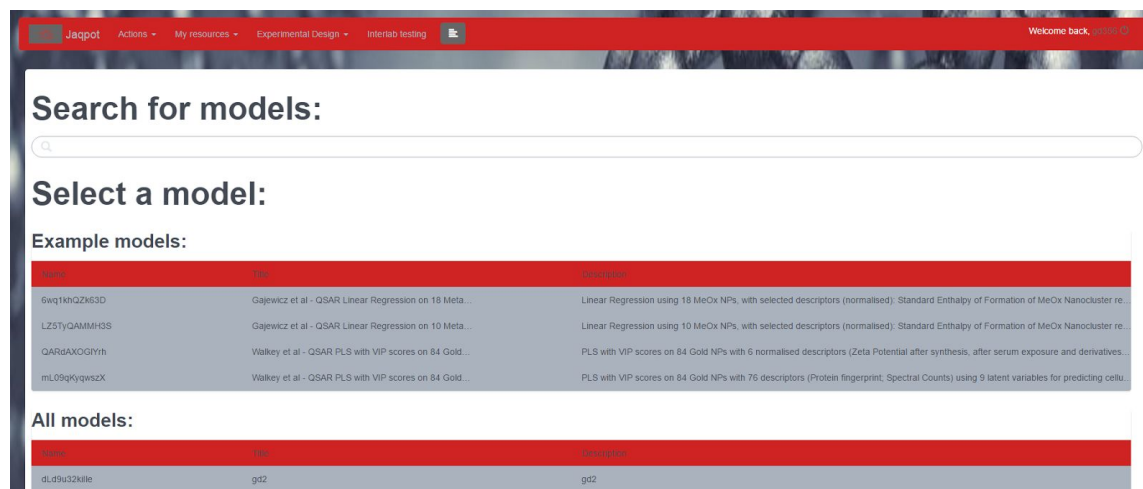
Figure 24: Additional information on dataset used for prediction

6. VALIDATION

Three types of validation are available in JQ, namely external validation (where an external test set is provided), cross-validation (where dataset is split into equal folds and results are aggregated) and training set split, for which a subset of the training set is held-out and used for prediction.

6.1 EXTERNAL VALIDATION

Users first select a model created (Figure 25) and then an external dataset (Figure 26) on which to perform the prediction and subsequent validation.



Select a model:

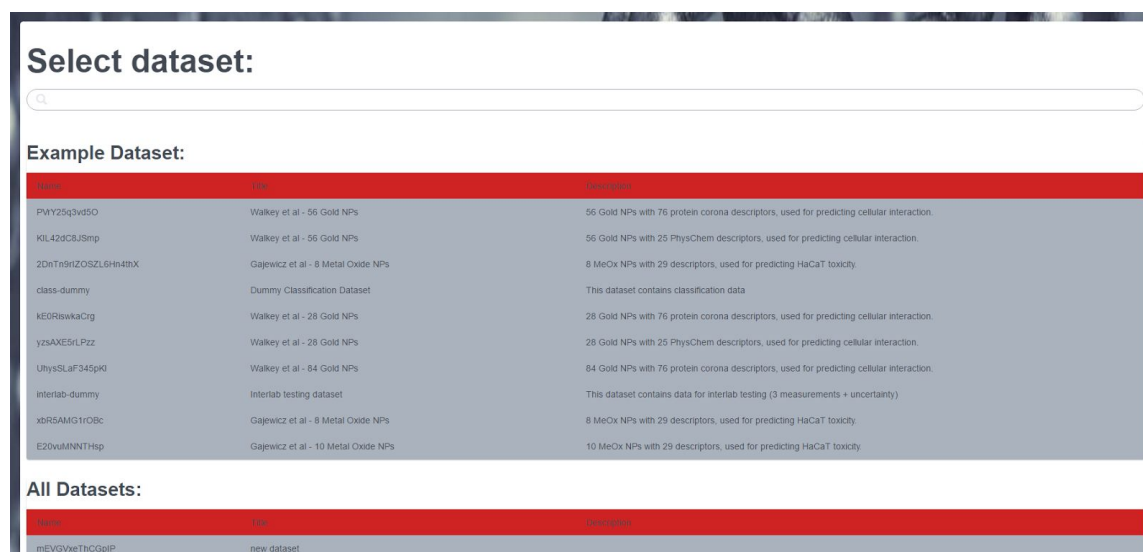
Example models:

Name	Url	Description
swq1khQZk63D	Gajewicz et al - QSAR Linear Regression on 18 Meta...	Linear Regression using 18 MeOx NPs, with selected descriptors (normalised). Standard Enthalpy of Formation of MeOx Nanocluster re...
LZSTyQAMMH3S	Gajewicz et al - QSAR Linear Regression on 10 Meta...	Linear Regression using 10 MeOx NPs, with selected descriptors (normalised). Standard Enthalpy of Formation of MeOx Nanocluster re...
QARdAXOGIYh	Walkey et al - QSAR PLS with VIP scores on 84 Gold...	PLS with VIP scores on 84 Gold NPs with 6 normalised descriptors (Zeta Potential after synthesis, after serum exposure and derivatives...
mL09qiyqwszX	Walkey et al - QSAR PLS with VIP scores on 84 Gold...	PLS with VIP scores on 84 Gold NPs with 76 descriptors (Protein fingerprint, Spectral Counts) using 9 latent variables for predicting cellu...

All models:

Name	Url	Description
dl_d9u32kIIE	gdz	gdz

Figure 25: Select model for external validation



Select dataset:

Example Dataset:

Name	Url	Description
PMY2sq3vd5O	Walkey et al - 56 Gold NPs	56 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.
KIL42dCSJsrnp	Walkey et al - 56 Gold NPs	56 Gold NPs with 25 PhysChem descriptors, used for predicting cellular interaction.
2DnTr9rIZOSZLGHn4Bx	Gajewicz et al - 8 Metal Oxide NPs	8 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.
class-dummy	Dummy Classification Dataset	This dataset contains classification data
KEORswkaCrg	Walkey et al - 28 Gold NPs	28 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.
yzsAXE5kLPzz	Walkey et al - 28 Gold NPs	28 Gold NPs with 25 PhysChem descriptors, used for predicting cellular interaction.
UhySSLAF345pkq	Walkey et al - 84 Gold NPs	84 Gold NPs with 76 protein corona descriptors, used for predicting cellular interaction.
interlab-dummy	Interlab testing dataset	This dataset contains data for interlab testing (3 measurements + uncertainty)
xBR5AMG1rORc	Gajewicz et al - 8 Metal Oxide NPs	8 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.
E20valMNNTHsp	Gajewicz et al - 10 Metal Oxide NPs	10 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.

All Datasets:

Name	Url	Description
mEVGVveThCGpIP	new dataset	

Figure 26: Select dataset for external validation

Once the task is completed, the user may view the results in the form of a report, which contains various performance metrics and the predictions shown in Figure 27, as well as plots such as QQ Plot shown in Figure 28.

Report: #W0Co9Im4AVGmEoD

All Data

Number of predictor variables:	2
RMSD:	0.12
StdError:	0.14
R ² :	0.93
F-Value:	47.49
Algorithm Type:	REGRESSION
R ² Adjusted (if applicable):	0.91

	Real	Predicted
Row_10	2.02	2.2469
Row_4	2.64	2.5709
Row_5	2.31	2.3771
Row_6	2.12	1.9833
Row_7	1.76	1.7457
Row_1	2.5	2.6228
Row_2	2.83	2.762
Row_3	2.92	3.0257
Row_8	2.24	2.1215
Row_9	3.32	3.1709

Figure 27: Validation Performance metrics and predictions

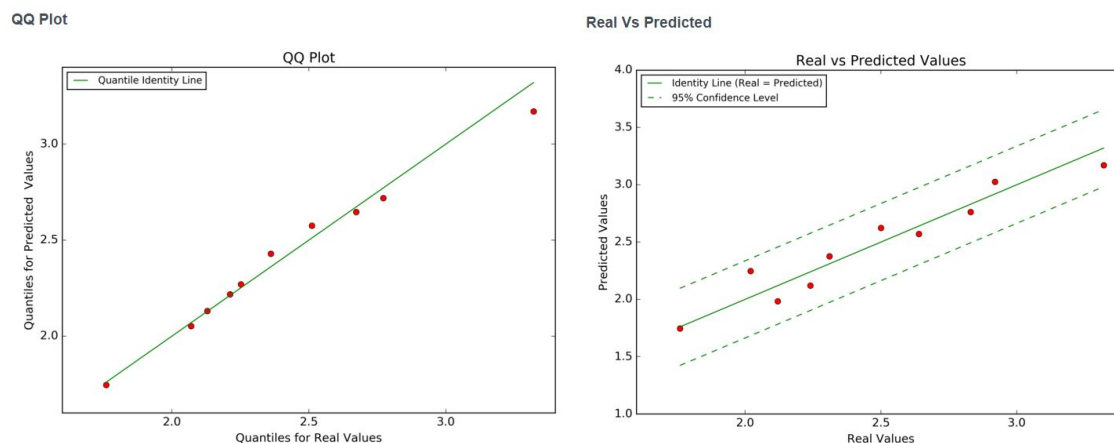


Figure 28: Plots contained within validation report

6.2 CROSS VALIDATION

For cross-validation, after the dataset is selected, the user is prompted to select an algorithm on which to perform the training. This screen is almost identical to the “Train” screen (Figure 29). For the algorithm, the user now has two additional options to specify, the number of folds and whether the dataset should be stratified (normal), random (user provides a seed for random function) or splits based on the observed order of data instances (None). A screenshot of this selection is shown in Figure 30. As with each of the options above, users may monitor a task (Figure 31).

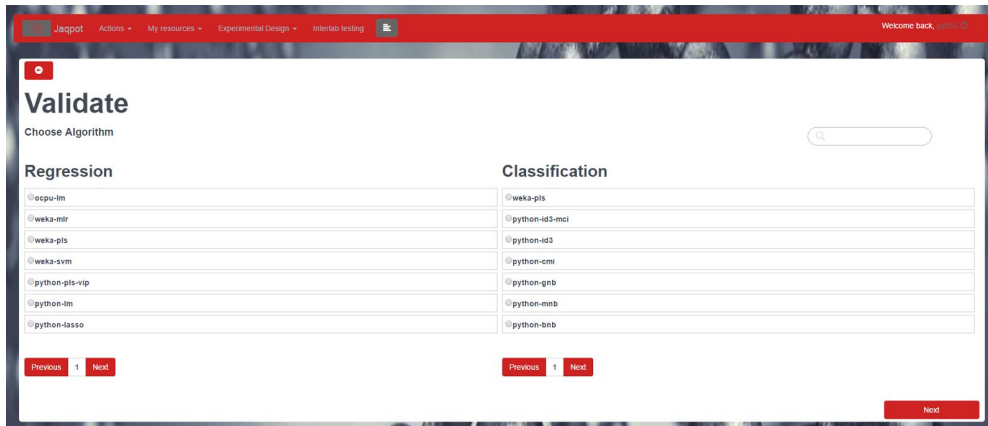


Figure 29: Select algorithm for cross-validation

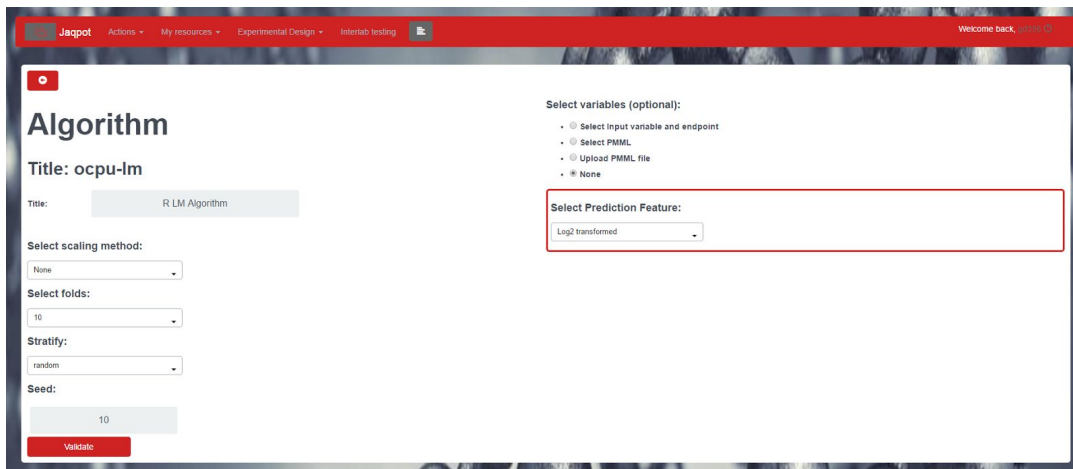


Figure 30: Screenshot of additional algorithm parameters (folds, stratify and seed) for cross-validation.

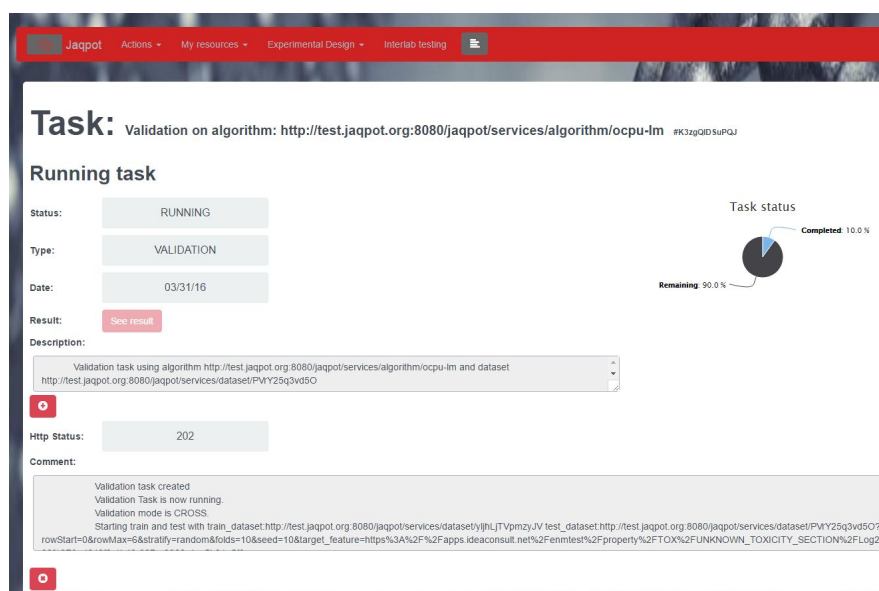
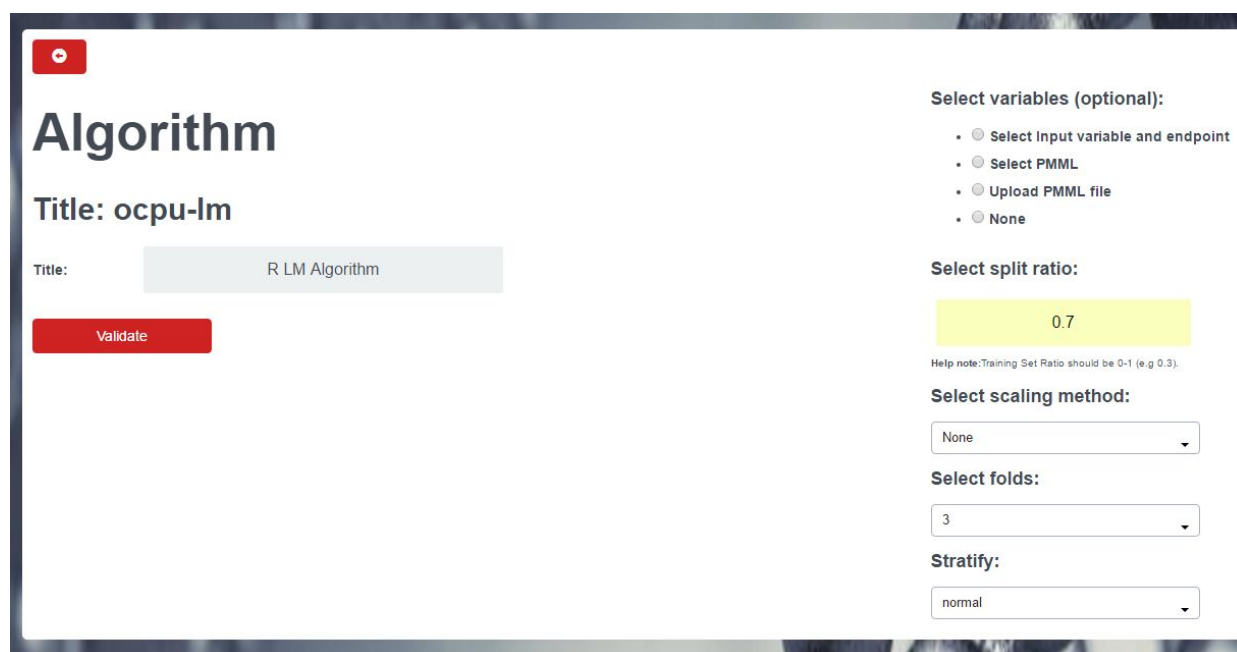


Figure 31: Full screen shot for cross-validation

6.3 TRAINING SET SPLIT VALIDATION

For training set split validation, users must select a dataset and algorithm as described above in detail. The only additional option for the algorithm is the “Select Split Ratio” field, which should be a value larger than zero and smaller than one, which describes the training-to-test-set-ratio to the algorithm. Therefore 0.7 would suggest that 70% of instances will be used for training and 30% will be held out for testing/validation. This is shown in Figure 32. The “Stratify” drop down menu is still available for the different instance selection options. The resulting model can be viewed once the task is completed.



The screenshot shows a web interface for configuring an algorithm. On the left, the title is 'ocpu-lm' and the description is 'R LM Algorithm'. A red 'Validate' button is visible. On the right, there are several configuration options:

- Select variables (optional):**
 - Select Input variable and endpoint
 - Select PMML
 - Upload PMML file
 - None
- Select split ratio:** A text input field containing '0.7'.
- Help note:** Training Set Ratio should be 0-1 (e.g 0.3).
- Select scaling method:** A dropdown menu with 'None' selected.
- Select folds:** A dropdown menu with '3' selected.
- Stratify:** A dropdown menu with 'normal' selected.

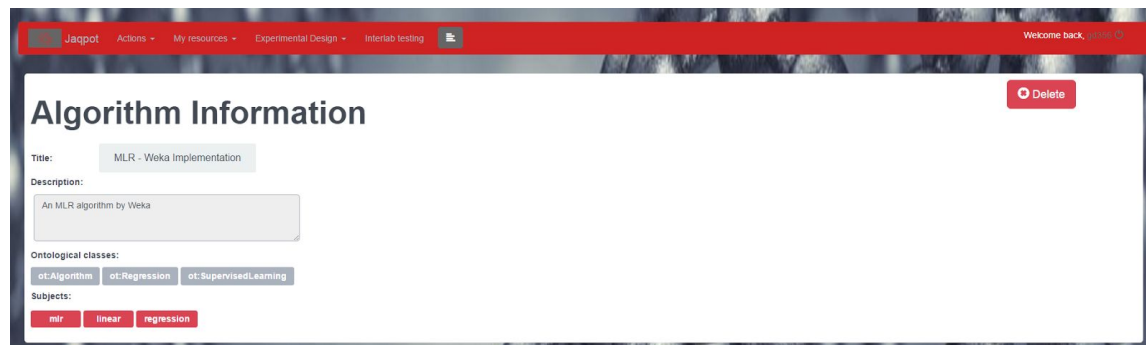
Figure 32: Algorithm options for training set split validation

7. MY RESOURCES

Under the “My resources” tab users may view algorithms, models, datasets and reports available to them, either because they are the creators or because they have been made available by Jaqpot from publications.

For algorithms, ontological classes and categories are available in addition to title/description (Figure 33). Datasets can be viewed in full including their unique ID assigned (Figure 34). Moreover, under models, users can view more information than just title/description. This includes contributors (could be authors of the publication from which the model is derived), publishers, DOI, keywords, algorithm used and attribute information (required features, dependent features, etc.). A screenshot is shown on Figure 35. Reports are listed in the same manner as datasets and models (Figure 36) and contain information from model validations (inter-lab testing or read-across methods will also create reports when these functionalities will be available). In the case of each resource type, users may delete an entry which they have created in two different ways. Firstly, each list of resources has a red “X” to the right of each entry which deletes it. This can be seen in Figure 36. Alternatively, users may go to the

individual entry and click on “Delete”, which is located at the top right of each resource entry. This allows users to not exceed their quota, which is 20 datasets, 20 models, 20 algorithms and 20 reports.



Algorithm Information Delete

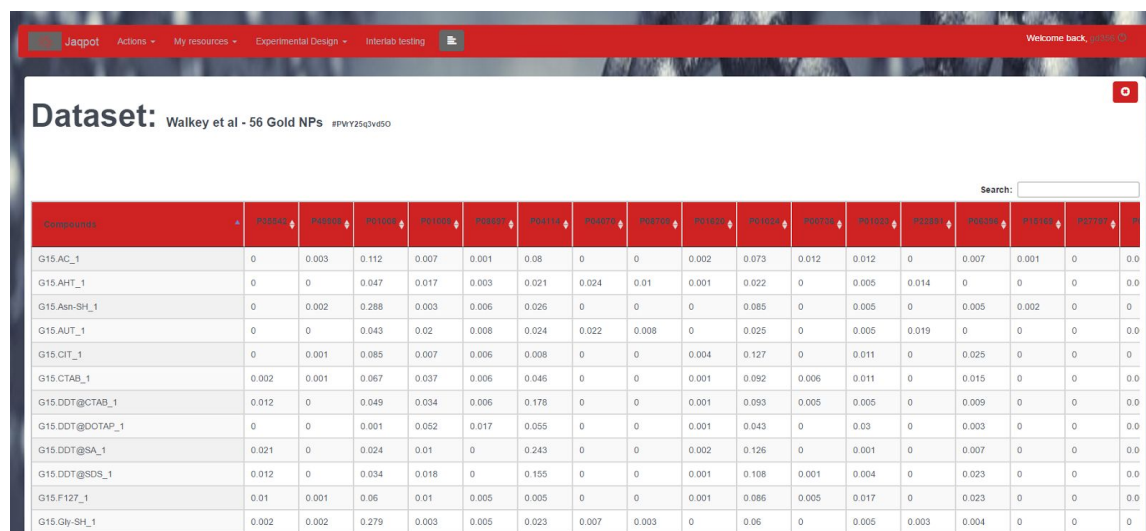
Title: MLR - Weka Implementation

Description: An MLR algorithm by Weka

Ontological classes: ot:Algorithm ot:Regression ot:SupervisedLearning

Subjects: mlr linear regression

Figure 33: Algorithm information under “My resources”



Dataset: Walkey et al - 56 Gold NPs #FWY25qVv650

Search:

Compounds	P15246	P15247	P15248	P15249	P15250	P15251	P15252	P15253	P15254	P15255	P15256	P15257	P15258	P15259	P15260	P15261	P15262
G15.AC_1	0	0.003	0.112	0.007	0.001	0.08	0	0	0.002	0.073	0.012	0.012	0	0.007	0.001	0	0.0
G15.AHT_1	0	0	0.047	0.017	0.003	0.021	0.024	0.01	0.001	0.022	0	0.005	0.014	0	0	0	0.0
G15.Asn-SH_1	0	0.002	0.288	0.003	0.006	0.026	0	0	0	0.085	0	0.005	0	0.005	0.002	0	0
G15.AUT_1	0	0	0.043	0.02	0.008	0.024	0.022	0.008	0	0.025	0	0.005	0.019	0	0	0	0.0
G15.CIT_1	0	0.001	0.085	0.007	0.006	0.008	0	0	0.004	0.127	0	0.011	0	0.025	0	0	0
G15.CTAB_1	0.002	0.001	0.067	0.037	0.006	0.046	0	0	0.001	0.092	0.006	0.011	0	0.015	0	0	0.0
G15.DDT@CTAB_1	0.012	0	0.049	0.034	0.006	0.178	0	0	0.001	0.093	0.005	0.005	0	0.009	0	0	0.0
G15.DDT@DOTAP_1	0	0	0.001	0.052	0.017	0.055	0	0	0.001	0.043	0	0.03	0	0.003	0	0	0.0
G15.DDT@SA_1	0.021	0	0.024	0.01	0	0.243	0	0	0.002	0.126	0	0.001	0	0.007	0	0	0.0
G15.DDT@SDS_1	0.012	0	0.034	0.018	0	0.155	0	0	0.001	0.108	0.001	0.004	0	0.023	0	0	0.0
G15.F127_1	0.01	0.001	0.06	0.01	0.005	0.005	0	0	0.001	0.086	0.005	0.017	0	0.023	0	0	0.0
G15.Gly-SH_1	0.002	0.002	0.279	0.003	0.005	0.023	0.007	0.003	0	0.06	0	0.005	0.003	0.004	0	0	0

Figure 34: Dataset information under “My resources”

Model:

#mL09qKyqwszX

Validate me Predict Delete

Title: Walkey et al - QSAR PLS with VIP scores on 84 Gold NPs

Doi: 10.1021/nn406018q

Description:

PLS with VIP scores on 84 Gold NPs with 76 descriptors (Protein fingerprint, Spectral Counts) using 9 latent variables for predicting cellular interaction.

Contributors:

Yoram Cohen Hongbo Guo Rong Liu Carl D. Walkey Fayi Song Andrew Emili Warren C. W. Chan Jonathan B. Olsen D. Wesley H. Olsen

Publishers:

http://informahealthcare.com/nan Nanotoxicology informa healthcare

Subjects:

liquid chromatography tandem mass spectrometry protein corona cell uptake nanomedicine nanobiotechnology structures activity model quantitative proteomics

Algorithm:

python-pls-vip

Features:

Required Features
Dependent Features
Independent Features
Predicted Features

Representation:

PMML

Figure 35: Example model information contains title, description, keywords (subjects), publication information (contributors, publishers and DOI), algorithm used, algorithm features and model PMML if available.

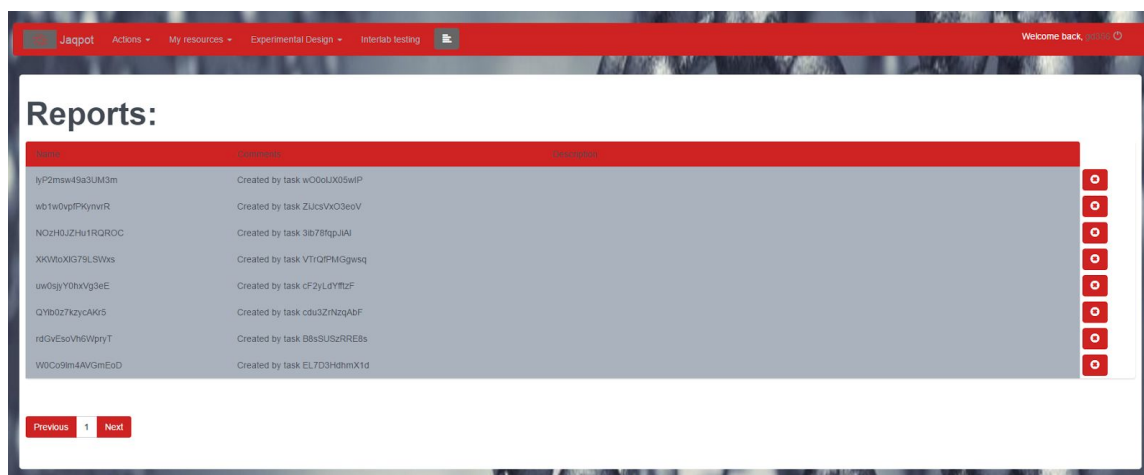


Figure 36: Report selection menu under “My resources”

8. FUTURE FUNCTIONALITIES

The next release of the JQ application will include functionalities for performing read-across predictions, interlaboratory proficiency testing and comparison and optimal experimental design when input information is available or not.

9. ACKNOWLEDGMENTS

The eNanoMapper project is funded by the European Union's Seventh Framework Program for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134.

10. REFERENCES

- Walkey et al., Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles, *ACS Nano* **8** (3), 2439-2455 (2014)

11. KEYWORDS

NanoQSAR modelling, data mining, model validation