USER'S GUIDE TO MaxTox

Developed under OpenTox – An Open Source Predictive Toxicology Framework

(2008-2011)

Table of Contents

Preface

User's guide to MaxTox is an attempt to list simplified instructions so that an end-user who is unaware of the detailed logic behind MaxTox can also go through few simple steps and learn to use the services offered by MaxTox.

MaxTox documentation is periodically updated and the latest versions are available at the site:

http://www.opentox.org/tutorials/maxtox

Any queries, bug-reports and feature-requests should be addressed to indirag@mail.jnu.ac.in or sunil@seascapelearning.com .

MaxTox is a service developed under OpenTox to provide a set of tools to build toxicity models against a specific endpoint using toxicity data. OpenTox is developed to provide an interoperable predictive toxicology framework to build and validate in-silico models and algorithms, and to provide a quality source of structures and toxicity data in order to develop a support of alternative methods to in-vitro assays for toxicity testing.

Note: OpenTox covers all products or services described in this manual. This document is meant for testing of the described services, as they exist on the current date of publication. The services are continuously developing so the manual and features described herein are subject to change without notice.

Acknowledgement: OpenTox - An Open Source Predictive Toxicology Framework, www.opentox.org, is funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011).

Audience

User's guide to MaxTox is intended for use of researchers, chemists and drug designers who wish to build in-silico models for predicting toxicity. MaxTox web application is available online and users may submit tasks from anywhere, which will run on the MaxTox server, and the links to results are sent to user's email address given at the time of job submission. MaxTox has a user friendly GUI and this manual guides a user how to use it with the help of properly illustrated screenshots. The manual also describes how to use MaxTox through command line calls.

Organization of the manual

The **User's guide to MaxTox** is organized as follows

Introduction: Gives the introduction about the MaxTox.

Algorithms: Description of the algorithms used.

Services: Services provided by the MaxTox.

Glossary: Description of terms used in MaxTox.

Introduction to MaxTox

MaxTox is a suite of tools to make models to predict toxicity of novel compounds based on their structural similarity to compounds with known toxicities. Compounds forming a model have the same Endpoint Toxicity.

For a training set of compounds, substructures occurring (in more than 1 compound and which consist of more than 2 atoms) are extracted and dictionaries built. New compounds are compared to this dictionary and a fingerprint denoting the presence/absence (of dictionary fragments in the test molecule) is generated. These fingerprints are binary i.e. they are represented using 0/1 for denoting absence/presence of a substructure in the compound. The fingerprints are then used to pass through a statistical Random Forest Model or SVM model to generate a predictive model. This predictive model is then used to predict the toxicity of the unknown set against the particular endpoint.

MaxTox is delivered as an open source application. It uses open source packages like CDK, Restlet, Jena & R to generate models and predict toxicities. It is part of the OpenTox initiative to create freely accessible resources to predict toxicity.

SVM used in MaxTox is SVMlight which is an implementation of Support Vector Machines (SVMs) in C (http://svmlight.joachims.org).

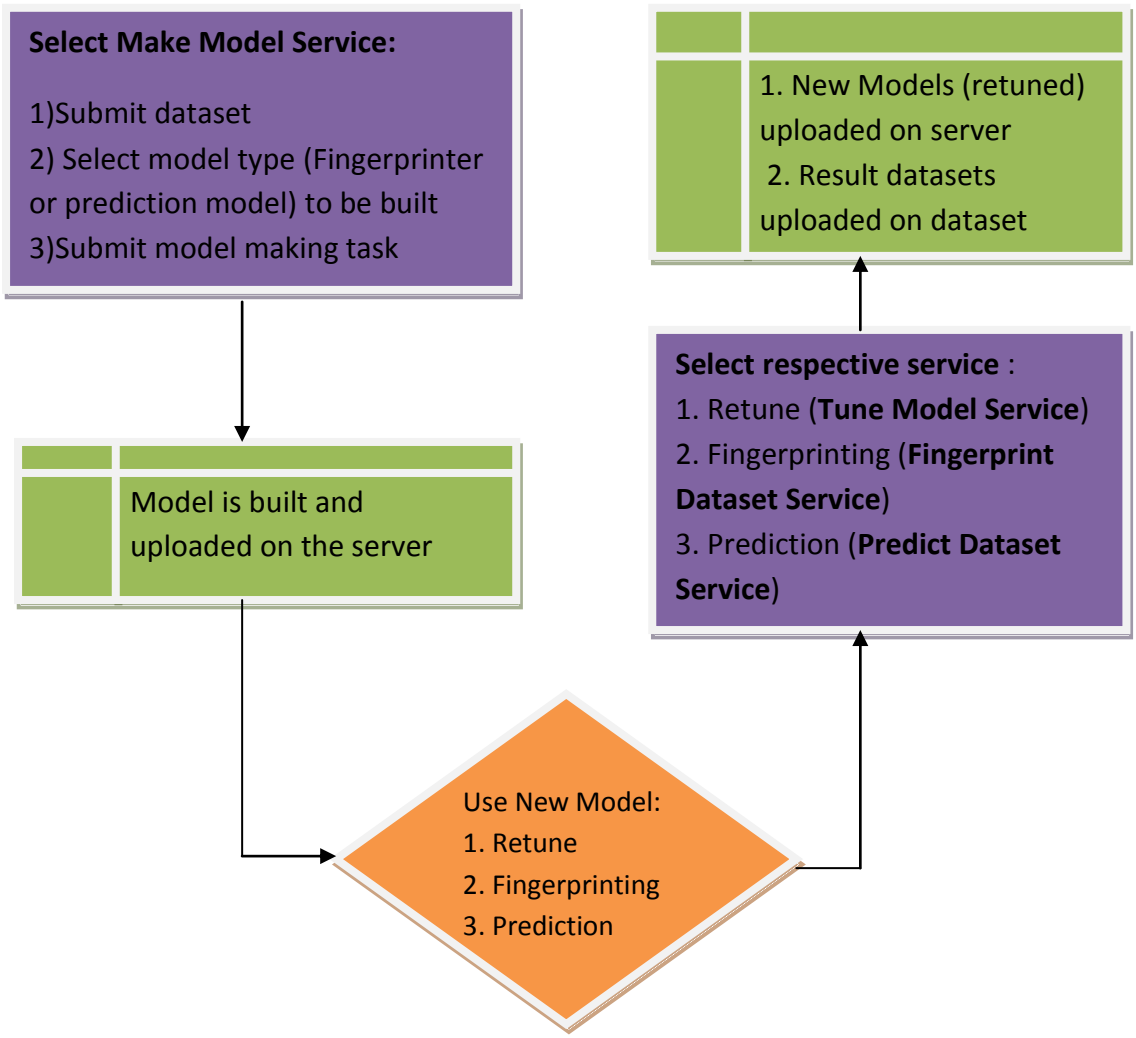
MaxTox can consume and respond in RDF datatypes and works with the OpenTox API specifiction. The MaxTox application can be used as a component of other prediciton use-cases hosted from other servers, as long as the data transactions are performed according to the OpenTox API.

The link to MaxTox web application is http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS .

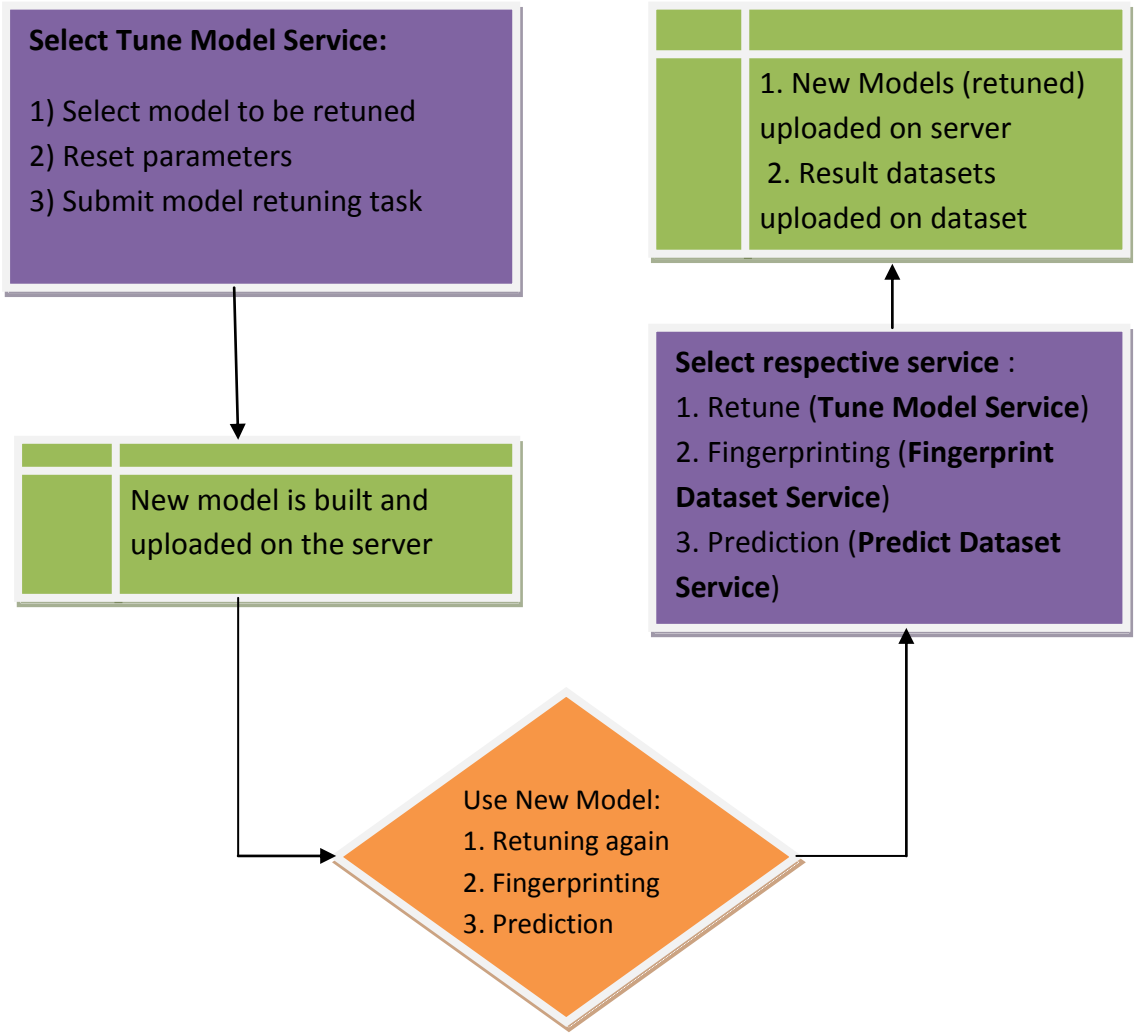
Use Cases

MaxTox offers a number of services, but mainly its utility can be summarised under following use cases. Following flowcharts explain how to use MaxTox for various use cases. Click on the respective services to go onto details.

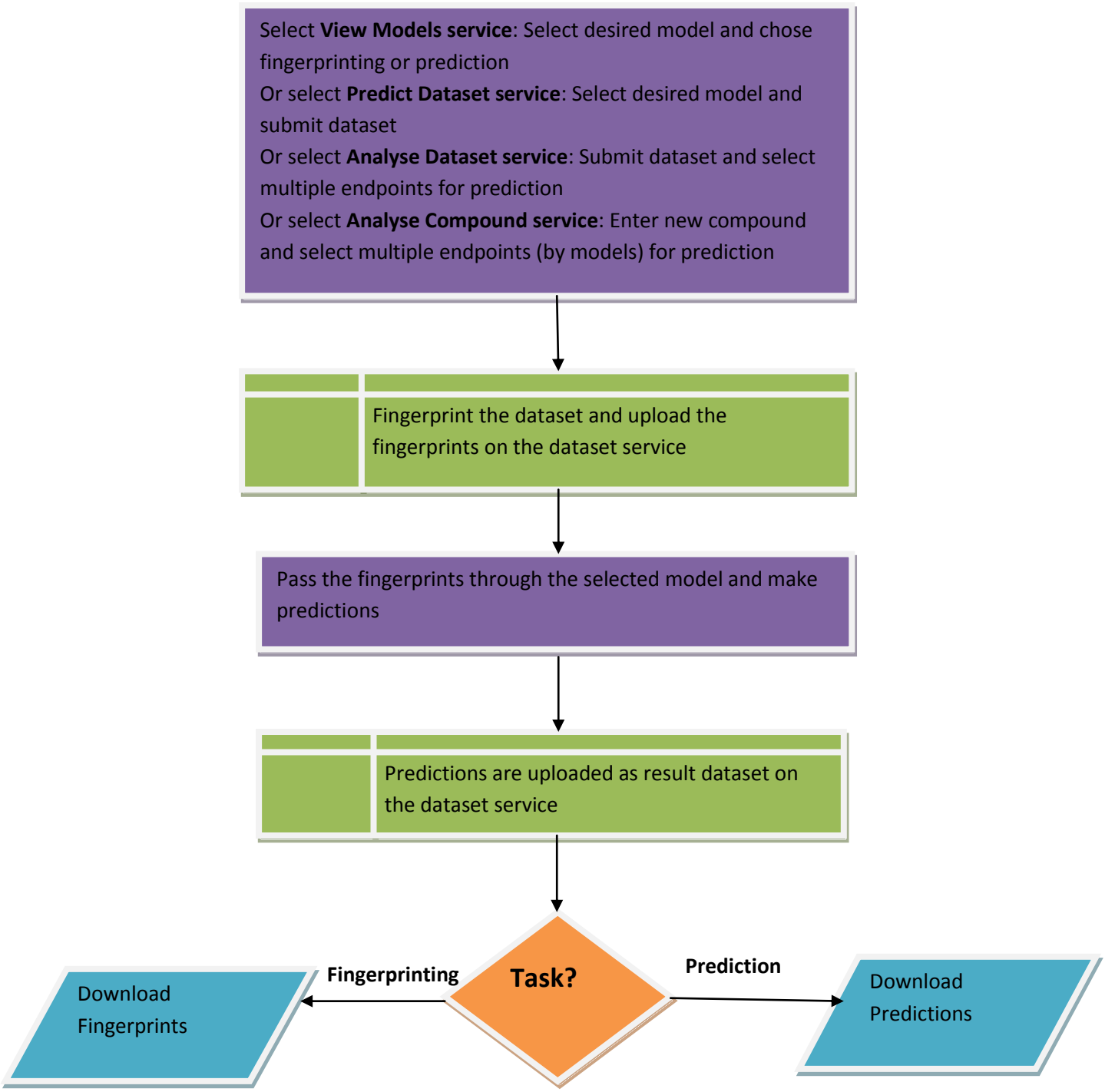
Making new model



Retuning existing model



Prediction/Fingerprinting using an existing model



Algorithms and Services developed under MaxTox

ALGORITHMS

Algorithms developed under MaxTox are listed below and are available at the address:

curl http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/algorithm

- **MCSSFinder:** Extracts Substructures from dataset and removes redundants to create a dictionary. Fingerprints the dataset and optionally creates an SVM model. Details of the algorithm are available at address :

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/algorithm/MCSSFinder

- **SVMRemodeller:** Remodels an existing SVM model using new parameters. Details of the algorithm are available at address :

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/algorithm/SVMRemodeller

- **MultiModelAnalyser:** An algorithm to use multiple models for predictions.

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/algorithm/MultiModelAnalyser

SERVICES

MaxTox offers a number of services for building models and generating fingerprints based on three basic algorithms developed under MaxTox, which are mentioned above. The services offered by MaxTox are:

- View Models
- Make Model
- Tune Model
- Fingerprint dataset
- Predict dataset
- Analyse dataset
- Analyse compound

The above-mentioned services are detailed in the later sections of the manual along with suitable examples to understand.

MaxTox GUI

MaxTox homepage shows the various services offered by MaxTox as shown in the following figure.

Fig 1 Home page

View Models

Address for the View Models service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/model

Step 1: Click on the View Models link to navigate to the page displaying the list of all the models available at MaxTox.

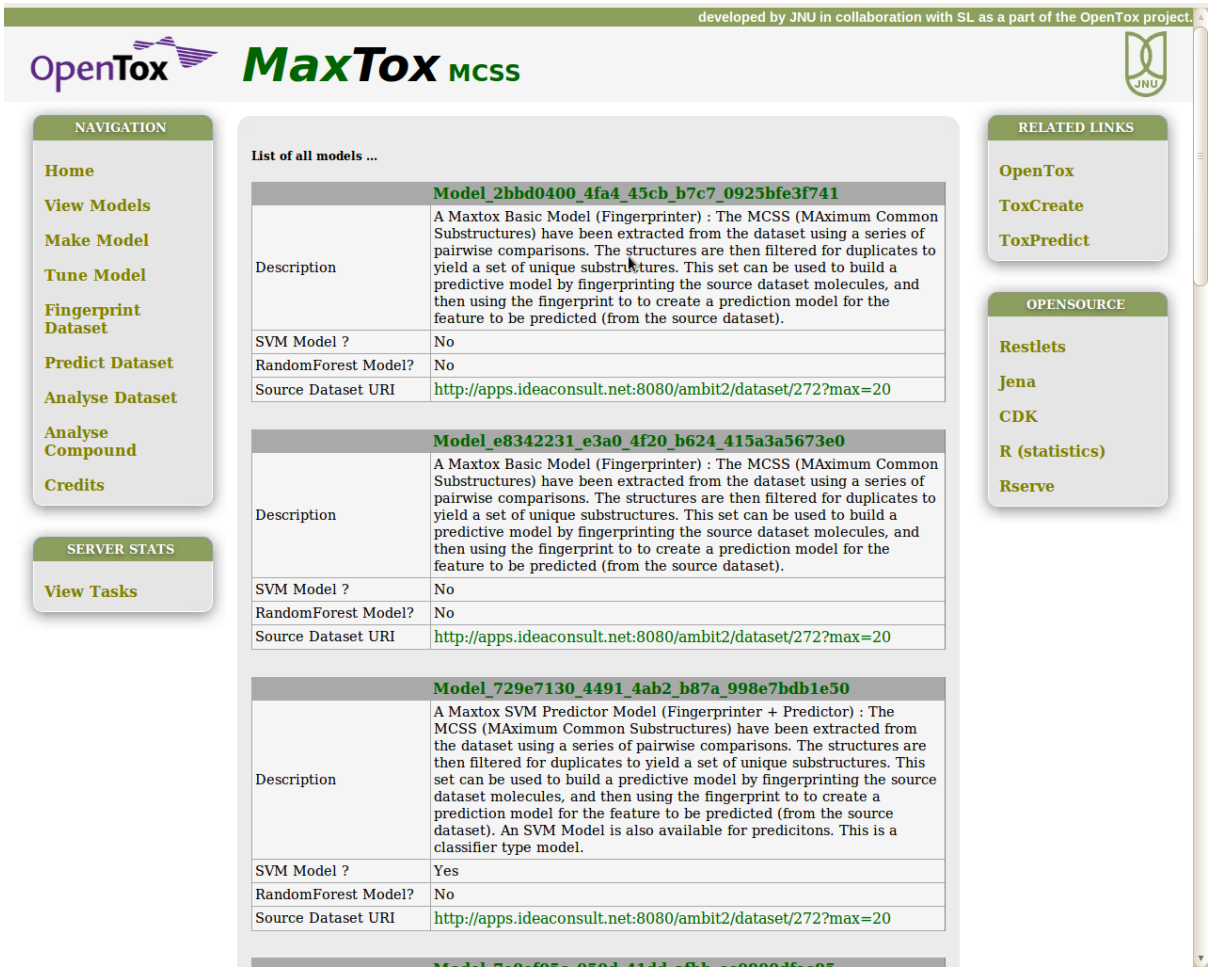


Fig 2: List of all models

Step 2: Click on model id to view the details of a particular model, for example, when you click on model id Model_729e7130_4491_4ab2_b87a_998e7bdb1e50, it takes you to the details about that model (Fig 3).

Fig 3a : Details of the model available at MaxTox

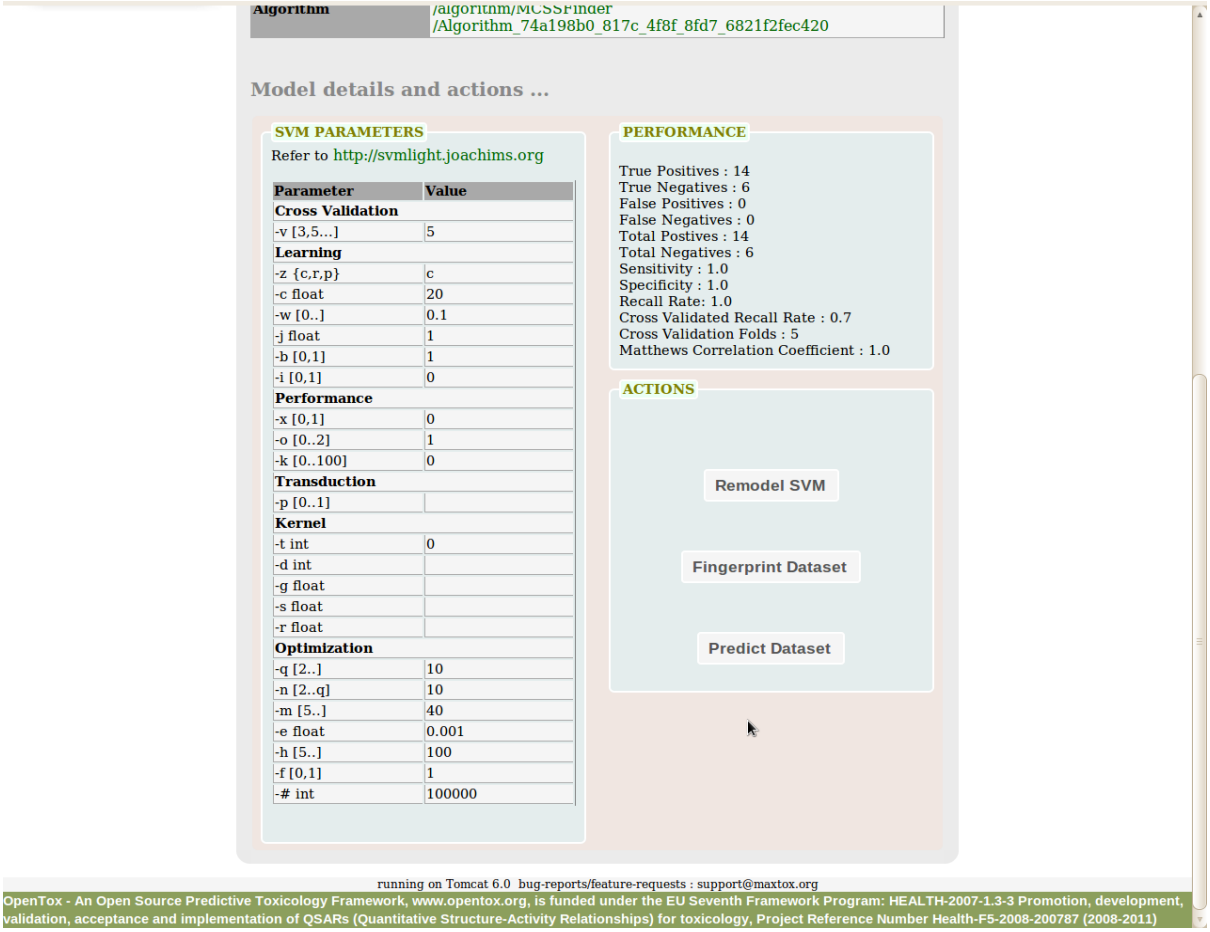


Fig 3b : Details of the model available at MaxTox

Now you can use this model to

Step 2a: Remodel SVM- If you wish to change the parameters used for SVM model building, click on the option Remodel SVM. This will take you to the Tune Model service of MaxTox. For details go to section **Tune Model.**

Step 2b: Fingerprint Dataset- If you wish to use this model to generate fingerprints using MaxTox algorithm, click on the option Fingerprint Dataset. This will take you to the Fingerprint Dataset service of MaxTox. For details go to section **Fingerprint Dataset.**

Step 2c: Predict Dataset- If you wish to use this model to predict a dataset, click on the option Predict Dataset. This will take you to the Predict Dataset service of MaxTox. For details go to section **Predict Dataset**.

User's Discussion :

The following main features are described for each enlisted model:

Description : It states that whether the model built is only a Fingerprinter model or a (Fingerprinter + predictor) model. A Fingerprinter only creates a dictionary of unique substructures, extracted from the dataset using MaxTox MCSSFinder algorithm. This dictionary can then be used for fingerprinting the source dataset compounds and the fingerprints can be used to create a prediction model either using the SVM, a machine-learning algorithm available at MaxTox or some other model building service compliant with OpenTox API version. While a (Fingerprinter + Predictor) is a classifier type model where the dictionary of MCSS is built , fingerprints are generated, the prediction model for the feature (like active-inactive, mutagen-nonmutagen etc) is built using SVM algorithm and the training dataset is predicted at one go.

SVM Model /RandomForest Model : This feature is valid for (Fingerprinter + Predictor) model. It says 'Yes' against SVM Model if the prediction model is built using SVM algorithm. Note: Currently RF model building option is not provided by MaxTox web application.

Source dataset URI: This gives the address of the source dataset used for making MCSS dictionary or the prediction models. If you want to look at the structure of compounds used for building the model, click on this link.

Fragments: Fragments link you to the page where all the substructure fragments are listed in detail along with their SMILES, structures and their occurrences in the dataset. These unique substructures are extracted by pairwise comparison of compounds of training set.

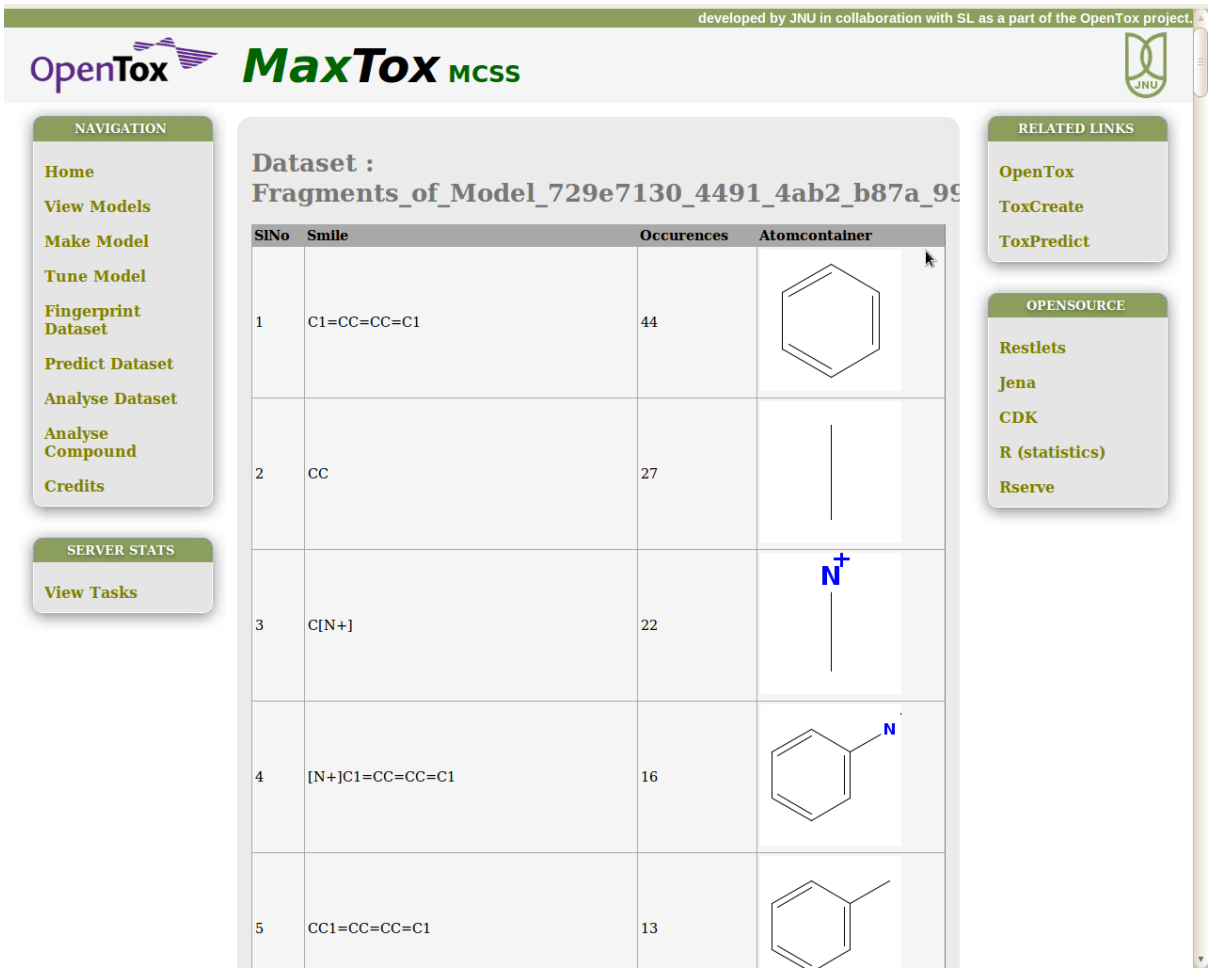


Fig 4: Fragments

SVM Predicted Training dataset URI : This is the address for the training set after being predicted using this model and contains original class of compound(as in source dataset), the predicted class of compound (using MaxTox model) an fragments present or absent in each compound. These datasets are uploaded on an external dataset service like ambit, for example in Fig 3, the predicted training dataset URI is http://apps.ideaconsult.net:8080/ambit2/dataset/601421.

Make Model

Address for the Make Model service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/makeMaxtoxModel

MaxTox can be used to generate Fingerprinter model or Prediction (or Fingerprinter + Predictor) model depending upon the user's requirements. (See user's discussion in View Models section for details)

If you wish to build a new model for your own dataset rather than using one of the existing models, Make Models service at MaxTox enables you to do so by following few simple steps.

Fig 5: Making MaxTox Fingerprinter and Prediction models.

Step 1: Uploading datasets

First, you have to upload your datasets to a dataset service. Since MaxTox works on the lines of OpenTox API framework, a dataset service is mandatory to upload source dataset and where the result dataset will be uploaded once MaxTox completes the task. The default dataset service that we are using for example use case is http://apps.ideaconsult.net:8080/ambit2/dataset.

The example dataset that we are using is http://apps.ideaconsult.net:8080/ambit2/dataset/272?max=10. Here dataset id is 272. max=10 simply means that we are selecting first ten compounds from an otherwise large dataset for example use case because fingerprint generation and model building may take very long for a large number of compounds. For example, full prediction run with existing models for fifteen compounds of Bursi mutagenicity dataset (http://apps.ideaconsult.net:8080/ambit2/dataset/272) took whole 1 minute and

15 seconds to complete.

Step 2: Submit a dataset URI

Enter the example dataset in this field. This will be the source dataset that is used to make the model.

Step 3: Submit a dataset URI for uploading of results

Enter a dataset service address that is OpenTox Dataset API compatible or if you want to use the default dataset service that is, AMBIT dataset service, leave this field blank.

Step 4: Submit an email address for notifications

The notification about successful task completion or any unexpected failure caused during the job processing will be sent to the email submitted. Users are requested to give a valid email in this field to free them from repeatedly checking for the completion of the tedious task of model building.

Step 5: Submit a Fingerprinter maker task

If you intend to make only a Fingerprinter model that is, only MCSS dictionary is to be generated, click on the button 'Submit a Fingerprinter-Model maker task'. However if you also want a prediction model to be built, proceed to next step.

Step 6: Submit a Prediction-Model maker task

If you want to create a Prediction Model using SVM algorithm, select radio button for 'Yes' against the text '**Would you like to predict features using SVM?'** This would unveil a set of modelling parameters. Each field describes clearly, what that parameter means (see Fig 6). Fill up these fields and then click on the button '**Submit a Fingerprinter-Model maker task**' to submit the task.

Fig 6: Modelling parameters to make SVM prediction model

Step 7: Submission of task

When the task is submitted, it will provide you with a unique task id, for example, for our example use case, we have **Task_dbe8b6e5_9ef8_4f37_a1ff_848bcc8ff2d2.**

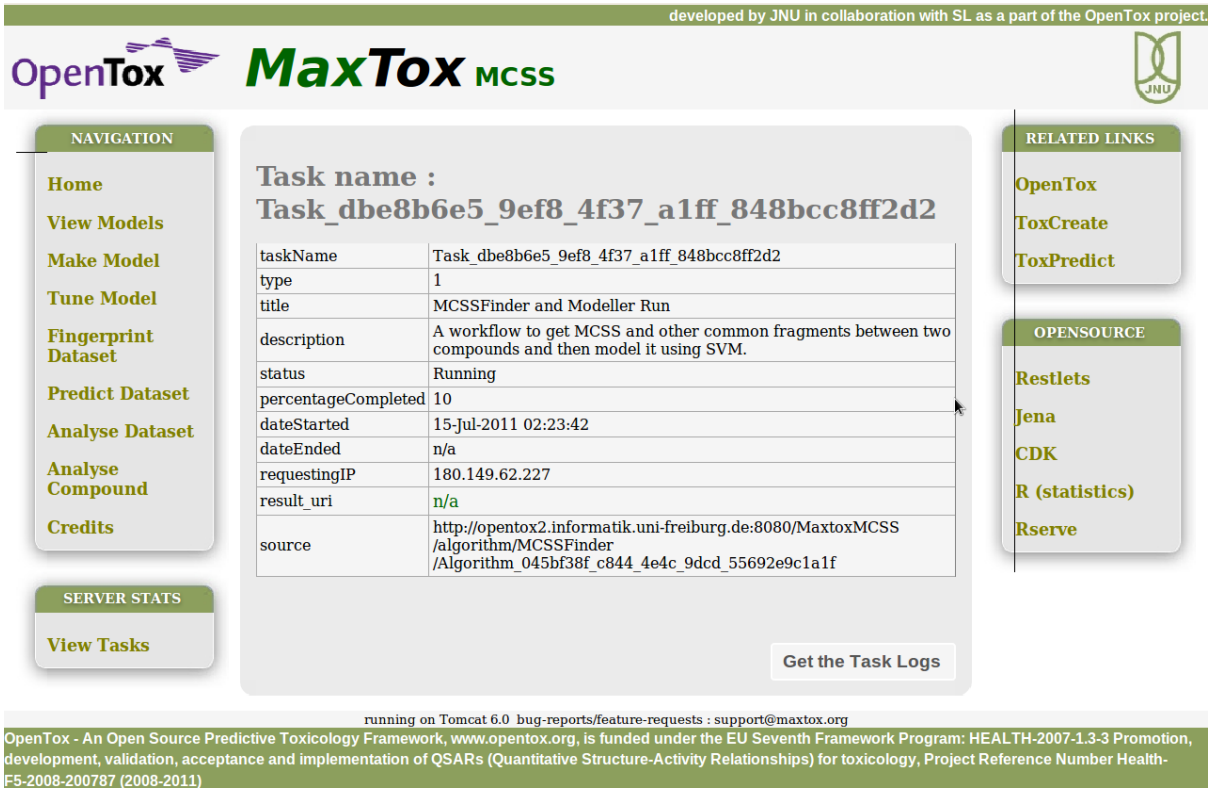


Fig 7: Task details

The details of the task like description, start time, end time, percentage completion etc is given at the address http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/task/Task_dbe8b6e5_9ef8_4f37_a1ff_848bcc8ff2d2. The model building may take time, so when the task is completed an email notification is sent to you.

Step 8: See the results

Refresh the task details page. If the task is completed, percentage completed is shown 100 and a result URI is provided which navigates to the detailed page for model built using the submitted dataset (same as Fig 3). Model details include SVM parameters used for model building and model performance measures like sensitivity, specificity, cross validation accuracy, correlation coefficient etc. To proceed from here refer to step 2 of View Models section.

User's Discussion:

When a source dataset is submitted for model buiding, first the MCSS substructures are extracted using pairwise comparisons of compounds done to obtain the largest overlaps. The fragments or MCSS dictionary is prepared for the training dataset by removing the duplicates. This MCSS dictionary is then used to generate fingerprints. These fingerprints are then used as input for SVM algorithm and a classifier type prediction model is built using the SVM parameters submitted by the user at the time of task submission. N fold cross validation during SVM modeling, gives a cross validation accuracy (n may be equal to 5 or 10, default is 5). In addition, sensitivity and specificity of the model give an idea about the model performance. This model can be further used for making predictions on unknown compounds against the same end-point.

View Tasks: MaxTox has a feature that enables a user to see the list of task histories submitted at MaxTox server and their status at the address: http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/task. Click on the task id, whose details you want to see.

Fig 8: Task history

Get Task Logs: Also, you can get the step-by-step log of your task specially if there is some unexpected error during the task processing. This helps the user as well as the developer in debugging.

Tune Model

Address for the Tune Model service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/remakeModel

Suppose you want to use an already existing model. The source dataset, the feature, and models everything is desirable but you are not satisfied with the model performance and you want to play around with the parameters used in SVM model building. Here the Tune Model service (Fig 9) comes in handy.

Fig 9: Submit a model for retuning

Step 1: Select the model

First, select the model whose parameters you want to tune and then click on the button '**Submit selected model for re-tuning'.** This will navigate you to a page where you can set new parameters for SVM model building (see Fig 10) like different kernel function or different number of folds of cross-validation.

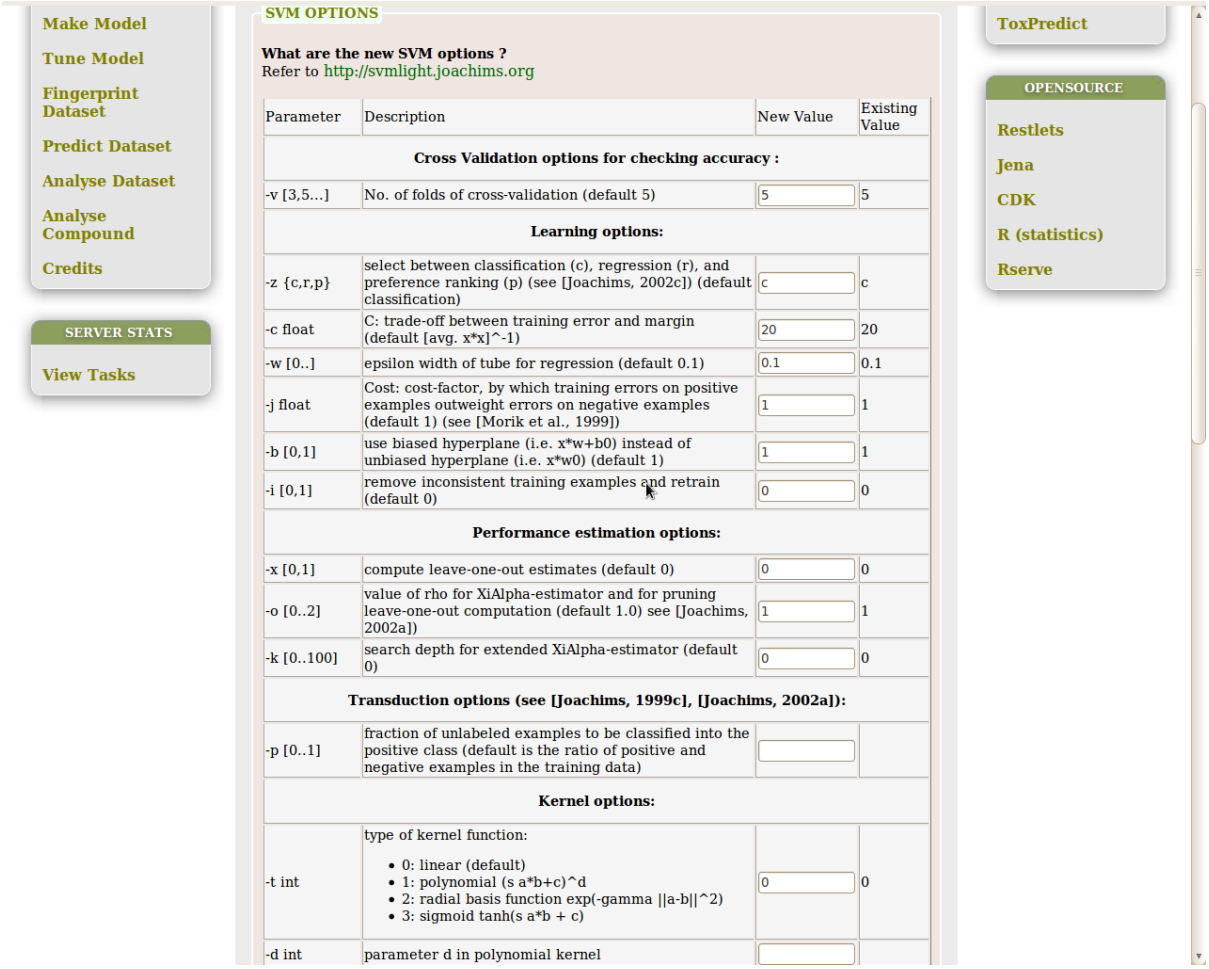


Fig 10: Tuning parameters of an existing SVM model

Step 2: Retune the parameters

After submitting the new SVM parameters, click on the button '**Submit model for re-tuning**' and submit the task. This will return a task id leading you to a page showing task details like the percentage completed, start time, end time etc. This may take a while.

Step 3: Check for re-tuned model

When you refresh the task page, the percentage completed is 100 and a result URI is provided which leads you to the re-tuned model that you desired. Now you can use your new model for future predictions.

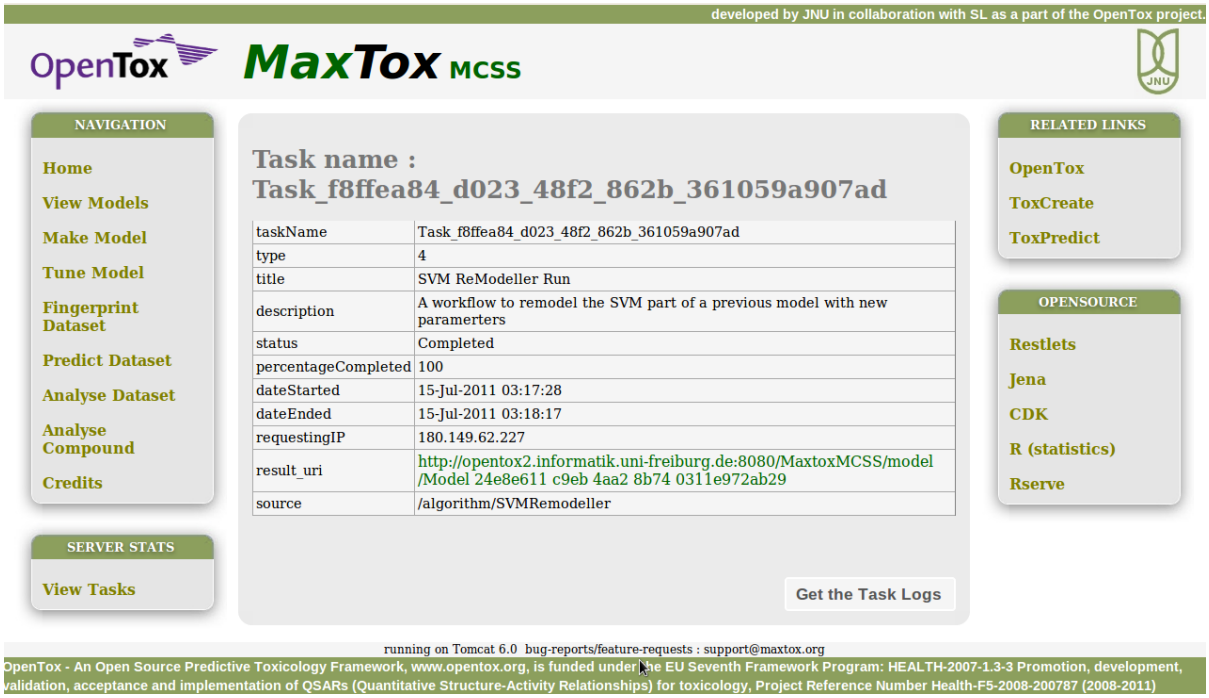


Fig 11: Getting a retuned model

User's Discussion:

SVM models performance that is, sensitivity and specificity can be improved by optimising the SVM parameters. Therefore, Tune Model service here provides user an option to improve model's performance. The algorithm used behind this service is SVMRemodeller instead of the MCSSFinder algorithm that was operating behind the Make Model service.

Fingerprint Dataset

Address for the Fingerprint Dataset service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/fingerprint

Step 1: Select a model for fingerprinting

To fingerprint a dataset using the models available at MaxTox, first select a model as shown in Fig 12 and then click on the button '**Submit selected model to use for fingerprinting'.** This will take you to the next page (see Fig 13).

Fig 12: Submit a fingerprinting task

Step 2: Submit a dataset for fingerprinting

Now first enter the address of the dataset whose fingerprints has to be generated using the model that was selected in step 1, for example, http://apps.ideaconsult.net:8080/ambit2/dataset/272?max=10 . Then enter a dataset service URI for uploading of results or leave the field blank (see Fig 13) to use the default dataset service at AMBIT.

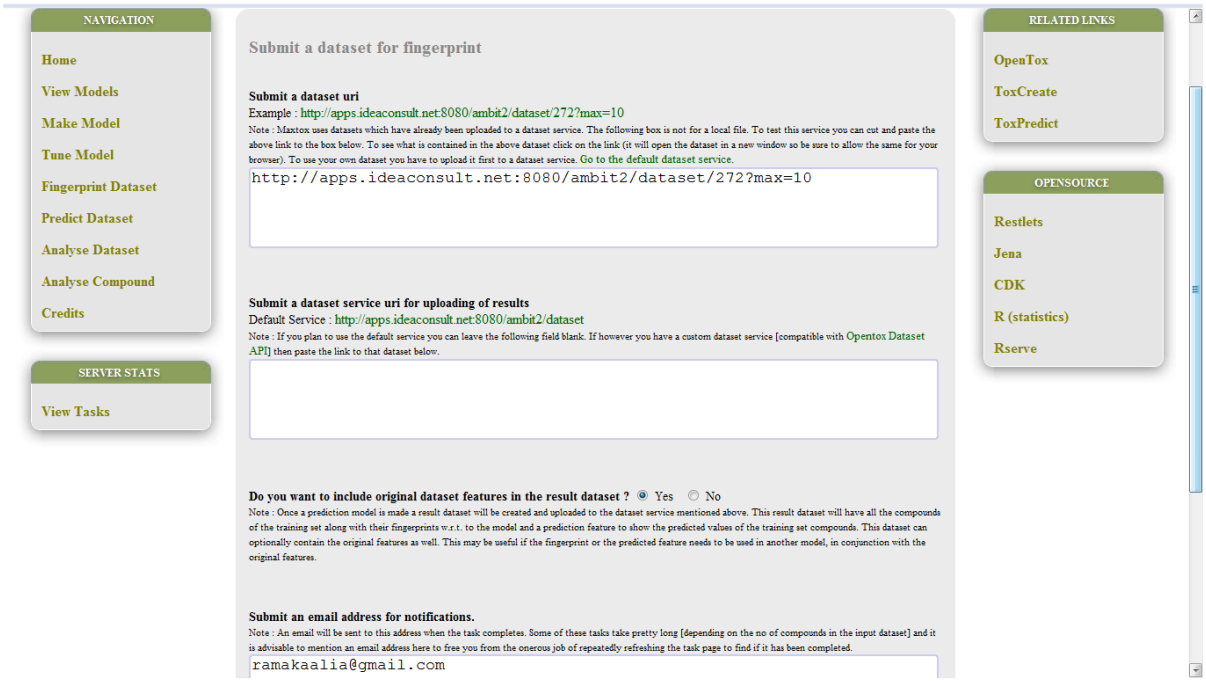


Fig 13: Submit a dataset for fingerprinting

After entering the respective fields, click on the button '**Submit dataset for fingerprinting**'. This will return a unique id for the task you have submitted and you will see a page showing task details like percentage completed. Completion of task may take a while.

Step 3: Task completion and obtaining the results

You will get an email notification on completion of the task. Refresh the page, the percentage completed will be 100 and the result URI will be provided which will eventually point at the dataset uploaded at the dataset service.

User's Discussion:

When a model is used for fingerprinting an unknown dataset, each of the fragments from the MCSS dictionary of the model is checked for its presence in each compound of the dataset. The absence/presence of these fragments is denoted by 0/1, thus generating the binary fingerprints for each compound that are uploaded at the result URI at the dataset service. These fingerprints can be downloaded from the result URI in a desired format (csv, text, pdf etc).

Predict Dataset

Address for the Predict Dataset service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/predict

Step 1: Select a model to use for prediction task

After selecting the desired model, click on the submit button provided on the bottom of the

page.

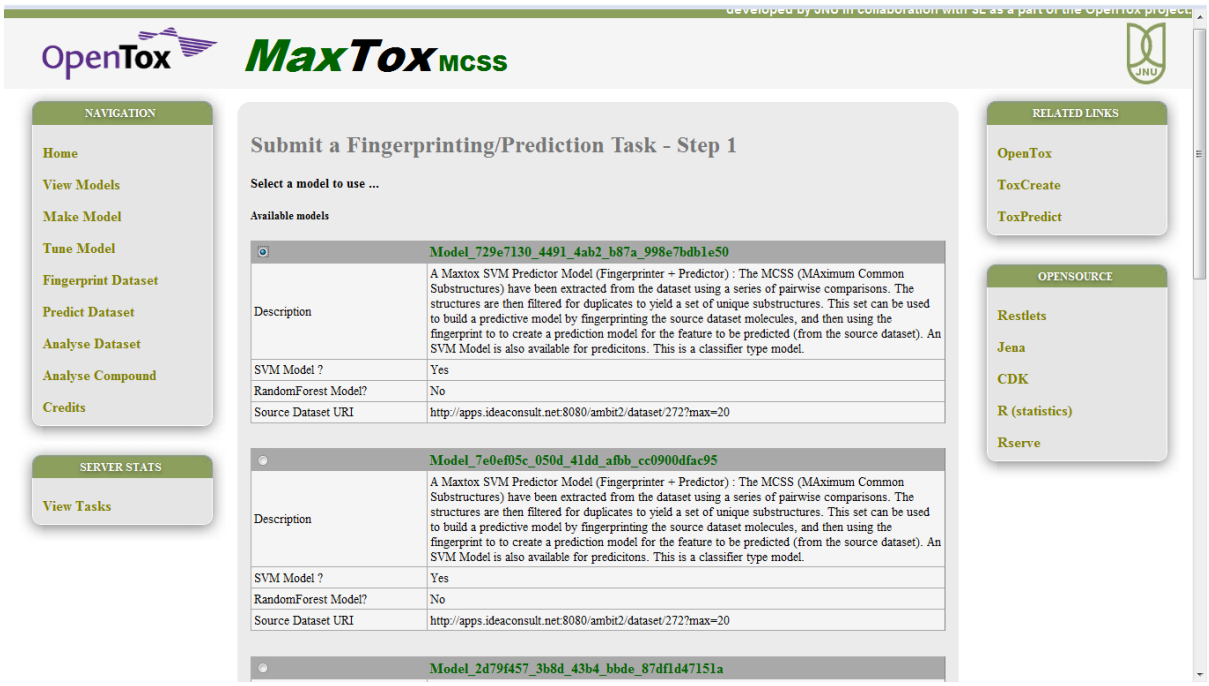


Fig 14: Select a model for prediction task

Step 2: Submit dataset for predictions

Step 1 will lead you to the next page (see Fig 15) showing a set of fields to enter the dataset for prediction. Enter appropriate values in all the required fields and click on the submit button provided on the bottom of the page '**Submit dataset for fingerprinring/predictions'** and you will be directed to the task detail page showing the percentage completion of the task.

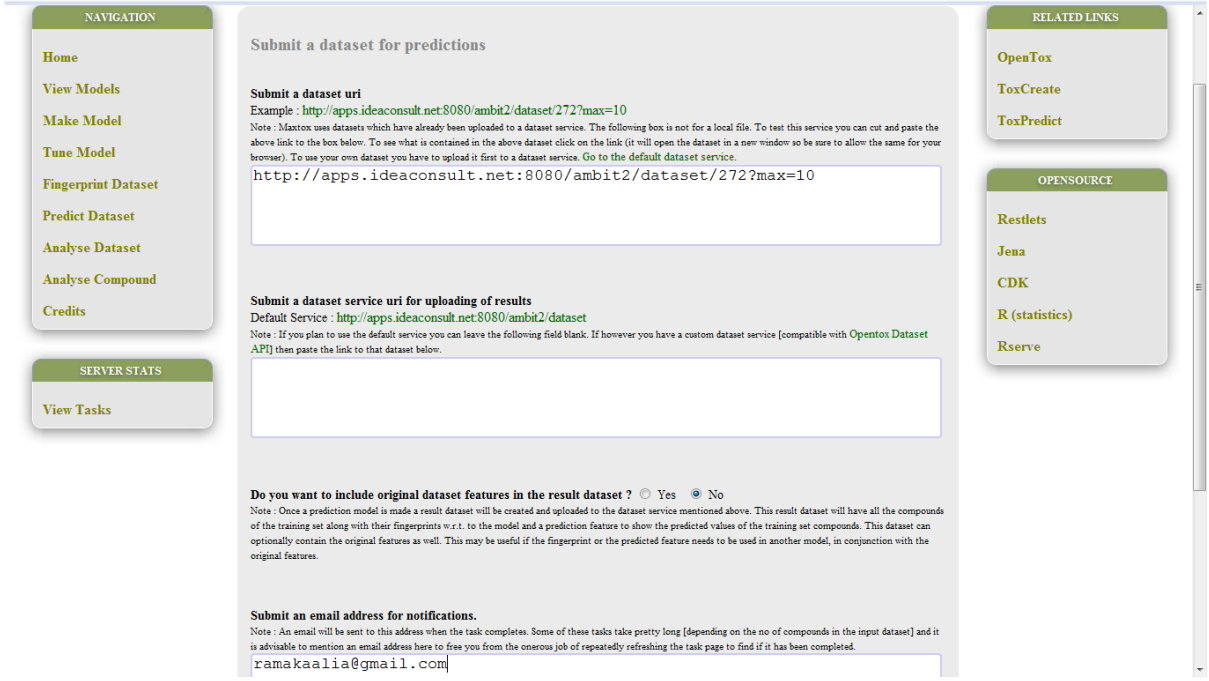


Fig 15: Submit dataset for predictions

Step 3: Task completion and prediction results

Once you are notified about the task completion by email, refresh the page and check for 100 percent completion of the task. Click on the result URI that is uploaded on the dataset service to see the predictions made by the model selected in the step 1.

User's Discussion:

When a model is selected for predicting an unknown dataset, one important thing to consider is that a model built for a dataset having aquatic toxicity, as the endpoint should not be used to predict the mutagenicity of a dataset.

Prediction task involves fingerprinting of the unknown dataset using the MCSS dictionary of the model and then using these fingerprints to pass through the prediction model (classifier type) to decide whether a particular compound belongs to one class or the other. The prediction results can always be downloaded from the dataset service using the result URI provided on task completion.

Note that MaxTox models presently work only for binary endpoint data, that is, dataset having two classes.

Analyse Dataset

Address for the Analyse Dataset service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/predictDataset

Step 1: Submit a dataset for multi endpoint analysis

Enter the dataset URI which has to be analysed using more than one endpoint in the field given at the above address and click on the submit button '**Submit dataset for analysis**' given at the bottom of the page.

Step 2: Select endpoints to predict

Select as many endpoints or models as you want to analyse your dataset for as shown in Fig 16, and then click on the submit button on the bottom of the page to submit the task.

Fig 16: Select endpoints to predict

Step 3: Task completion and analysis results

When you refresh the task detail page on completion, it will provide you a result URI. Click on the result URI to see the multi-endpoint analysis results for the given dataset (see Fig 17).

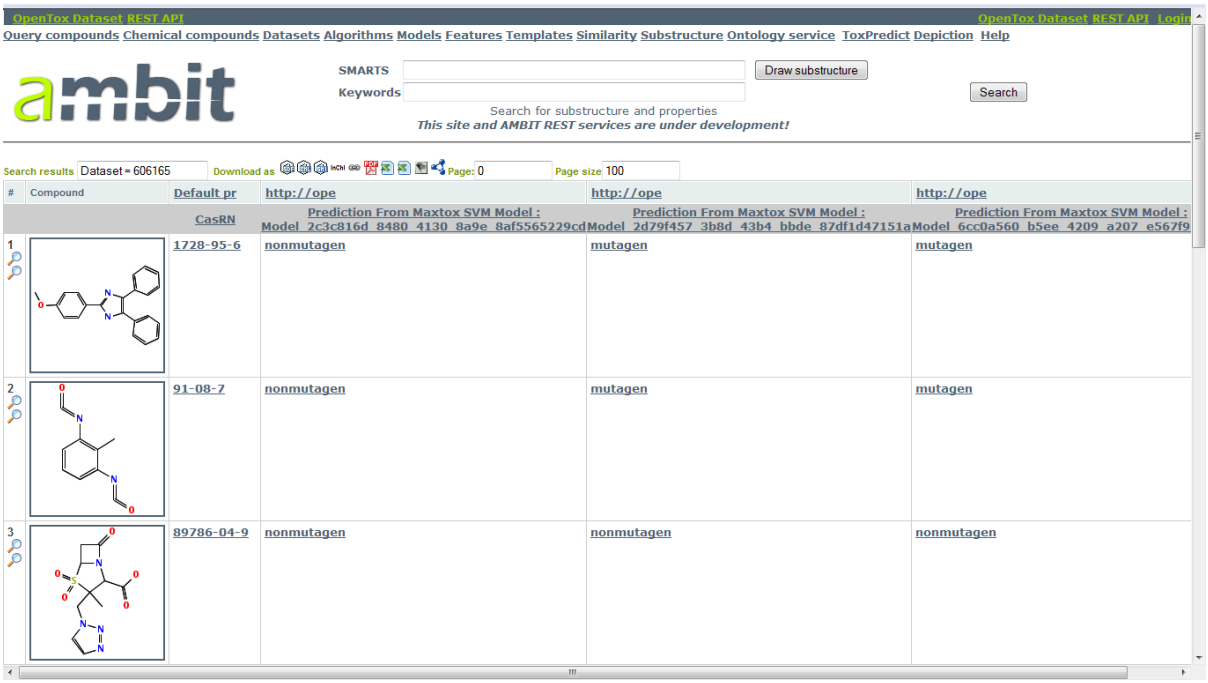


Fig 17: Multi-endpoint analysis results uploaded on the dataset service

User's Discussion:

In the multi endpoint analysis of a dataset using MaxTox Analyse Dataset service, you can predict a dataset using more than one model (and endpoints). The MaxTox algorithm running

behind this service is the MultiModelAnalyser. It enables you to predict your dataset for multiple endpoints, for example, Salmonella mutagenicity, Aquatic toxicity or carcinogenicity simultaneously by selecting the respective endpoint models at the same time.

Analyse Compound

Address for the Analyse Compound service:

http://opentox2.informatik.uni-freiburg.de:8080/MaxtoxMCSS/enterNewCompound

If a user wants to use MaxTox services for a private compound instead of the whole dataset, he can use Analyse Compound utility of MaxTox as follows.

Step 1: Enter a new compound

You can draw your compound or submit SMILES directly and then submit the compound for analysis by clicking on the submit button at the bottom of the page. For example, we have submitted benzene for testing.

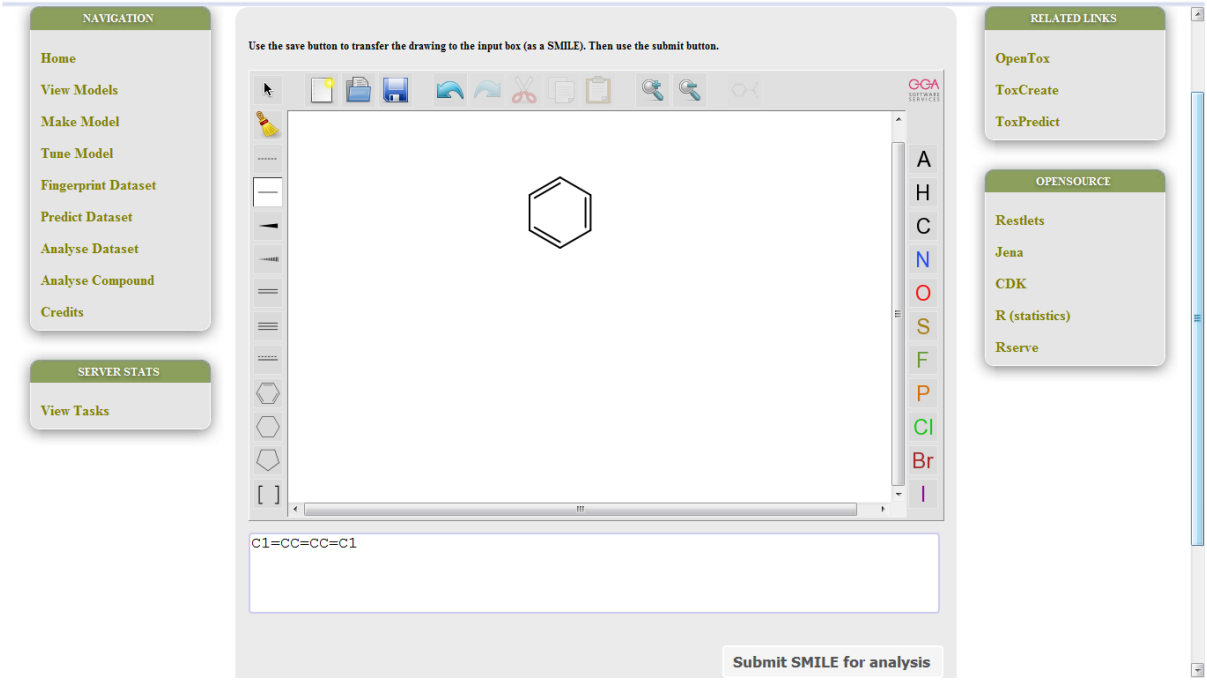


Fig 18: Enter a new compound for analysis

Step 2: Make new compound or select an existing one

When you submit a compound for analysis, first it is searched through the already available database. If the compound is already available in the database, select your compound and if not, there is a **Make new compound** option at the bottom of the page (see Fig 19).

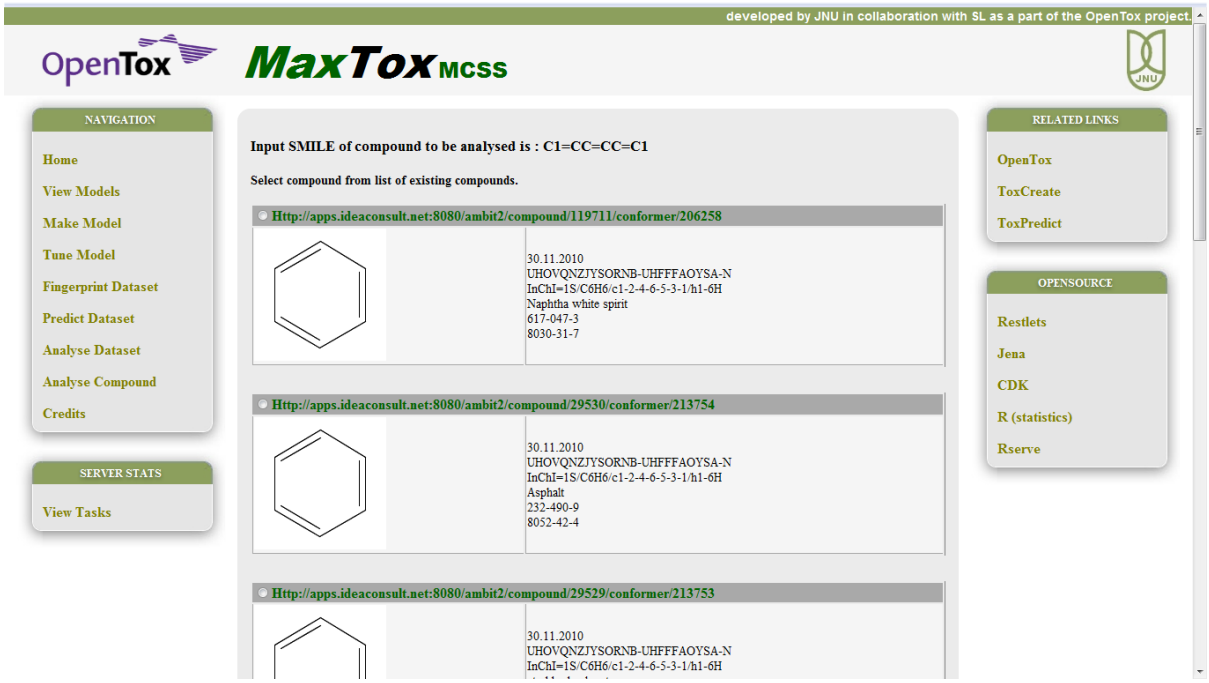


Fig 19a: Make new compound or select an existing one

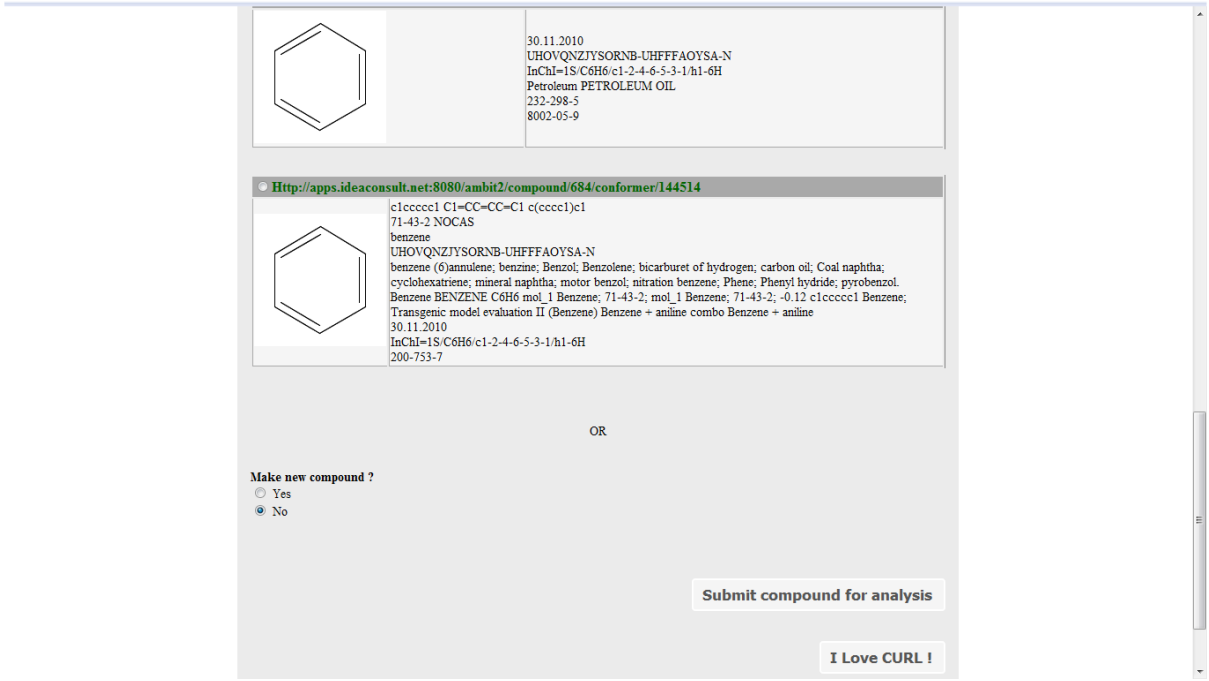


Fig 19b: Make new compound or select an existing one

Click on the submit button on the bottom either to select an existing compound or create a new compound and upload it on the dataset service.

Step 3: Select endpoints of models to predict

When you create a new compound, a new compound URI is obtained. For example, despite existence of benzene we opted for **Make new compound** option, and a new compound is formed at the URI:

http://apps.ideaconsult.net:8080/ambit2/compound/23575/conformer/709842

Now select the endpoints (or Models) for which you want to analyse your compound (see Fig 20) just like in case of multi-endpoint analysis of the dataset described in section **Analyse Dataset.**

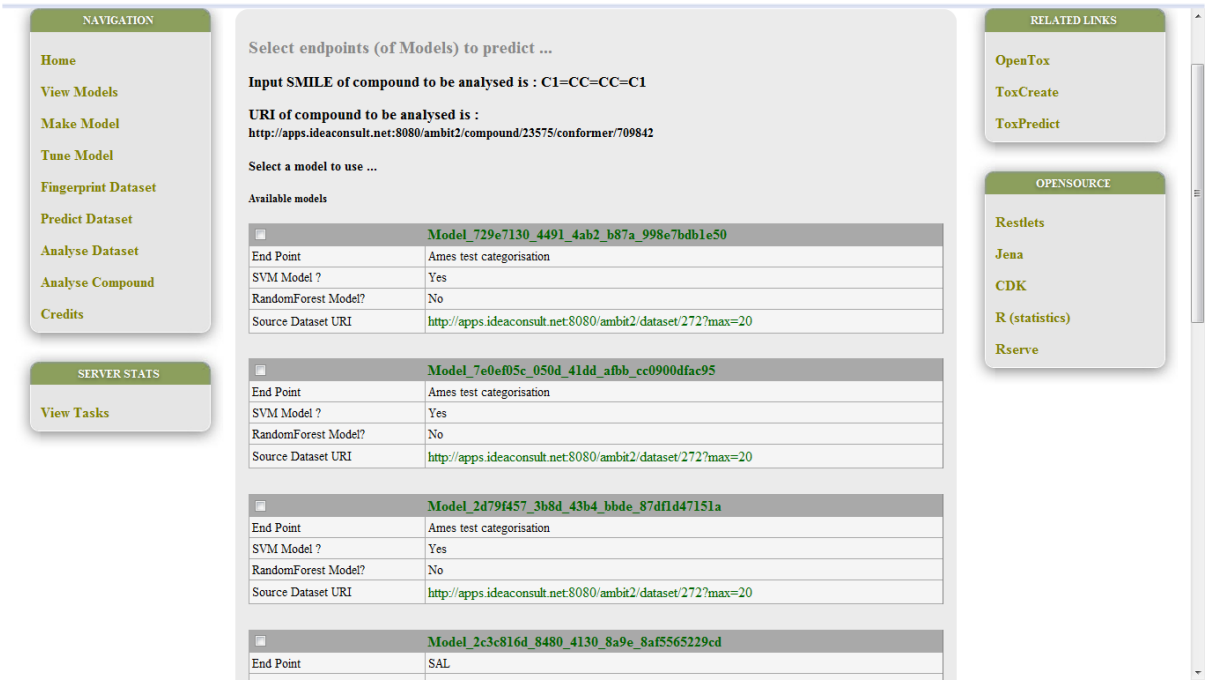


Fig 20: Select endpoints to predict

After selecting, click on the submit button to submit the task.

Step 4: Task completion and compound analysis results

Refresh the task detail page on completion and click on the result URI to see the analysis result for the compound you entered against the multiple endpoints.

User's Discussion:

Analysis Compound service is same as the Analyse Dataset service except the fact that in compound analysis you can provide a single compound instead of the whole dataset. Even for a single compound analysis, first, it has to be uploaded on the dataset service, only then it can be used by a MaxTox service. **Make new compound** option thus uploads the compound on the dataset service and provides the user with the compound URI.

Prediction results can be downloaded from the dataset service using the result URI for predictions where the given compound is predicted against the multiple endpoints selected in step 3.

Note: For using command line calls, one can click on the "I Love CURL" button provided at the bottom of each page and use the cURL command displayed to perform the same task as the respective GUI will do.

Glossary

End-point: End-point for toxicity denotes the result of the study conducted to determine how toxic or harmful a substance is. For example, Carcinogen/Noncarcinogen, Mutagen/Nonmutagen etc

Fingerprints: Fingerprint is a binary descriptor for a molecule denoting the absence/presence of a substructure fragment as 0/1.

MCSS: Maximum Common Substructure Search (MCSS) is an approach where Maximum Common Substructure (MCS) is the largest substructure that appears in structures of both compounds. MCS can be used as a measure of chemical similarity to compute chemical descriptors, which can be further used for activity predictions.

QSAR: Quantitative Structure Activity Relationship is the process by which the chemical structure of a compound is quantitatively correlated with biological or chemical activity of that compound.

SVM: Support Vector Machines is a machine-learning algorithm that is based on supervised learning mechanism and is used for classification and regression analysis of data. When a training set of data belonging to one of the two classes is given as input, SVM builds a prediction model that can classify new data entries to either of the two input classes.

RF: Random Forest is an ensemble classifier that makes its decision based on outputs from multiple decision trees so that the final class predicted is mode of the class's output by individual trees.

URI: Uniform Resource Identifier represents the name or address of a resource on the internet. URIs are of two types: URN (Uniform Resource Name) and URL (Uniform Resource Locator).

SMILES: Simplified Molecular Input Line Entry Specification is a chemical language to represent chemicals and reactions using ASCII characters.

Dataset service: Dataset service refers to a service, which contains data in the format specified by OpenTox API guidelines and where a service like MaxTox can upload more datasets and retrieve the existing ones from the service.

References

Bron, C.; Kerbosch, Finding All Cliques of an Undirected Graph. J. Commun. ACM 1973, 16, 575-577.

Prakash, O.; Ghosh, I, Developing an Antituberculosis Compounds Database and Data Mining in the Search of a Motif Responsible for the Activity of a Diverse Class of Antituberculosis Agents, J. Chem. Inf. Model. 2006, 46, 17-23.

John W Raymond , Eleanor J. G. Ardiner, Peter Willet, RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs, The Computer Journal. 2002, 45(6), 631-644.