

Table B – Drug Discovery Predictive Toxicology Application

An example of a predictive toxicology application in drug discovery is given using the data on antimalarial compounds made available at the ChEMBL Neglected Tropical Disease (NTD) archive (<http://www.ebi.ac.uk/chemblntd/>).

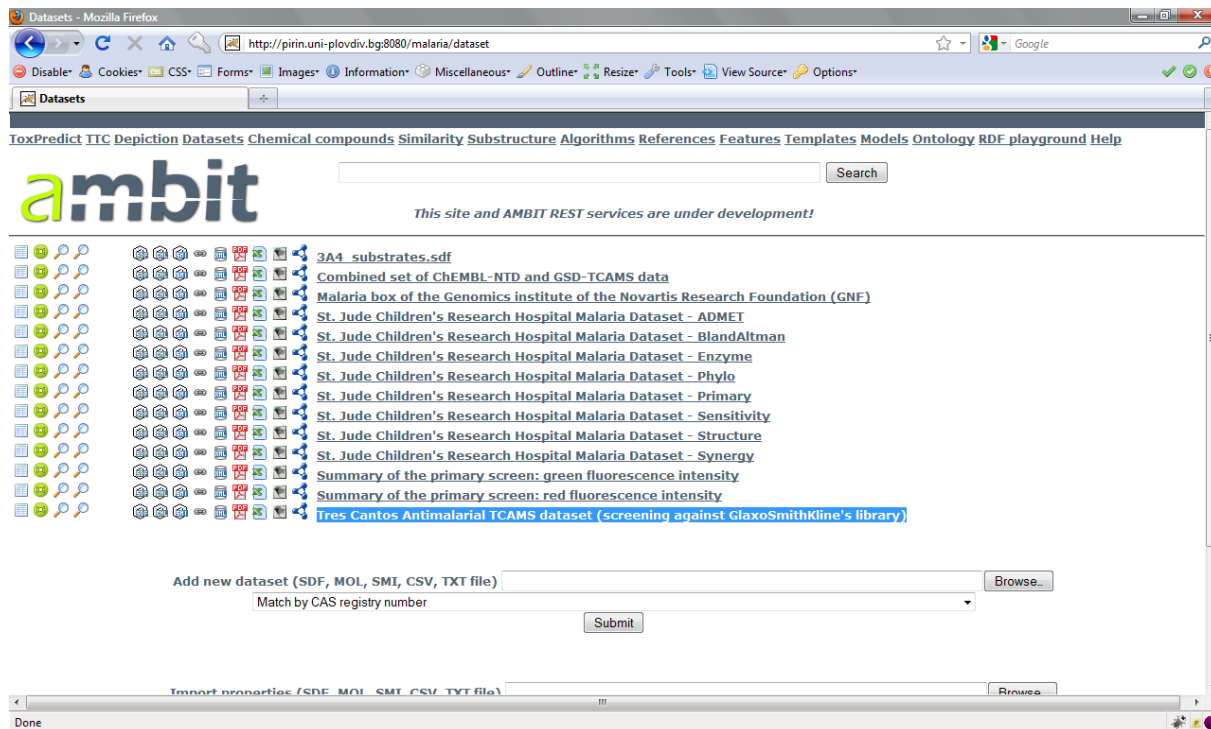
Activity B: Selecting a subset to create a model with ToxCreate

As a second exercise, subsets of the antimalarials are extracted to be used in a model building exercise via the OpenTox demo application ToxCreate.

857 of the 13'519 compounds contained in the TCAMS dataset are annotated with a protein (class) target. Of these 857 compounds, 233 are annotated as Ser/Thr kinase inhibitors. In this exercise we'll use this information to create a dataset that can be used to build a model that predicts whether or not a given compound is likely to be a kinase inhibitor. The dataset for the model building therefore needs to consist of two columns: the SMILES string of the compound and its classification (Ser/Thr kinase inhibitor = 1, otherwise 0).

To create this dataset go to <http://pirin.uni-plovdiv.bg:8080/malaria/dataset>

Click on “[Tres Cantos Antimalarial TCAMS dataset \(screening against GlaxoSmithKline's library\)](#)”

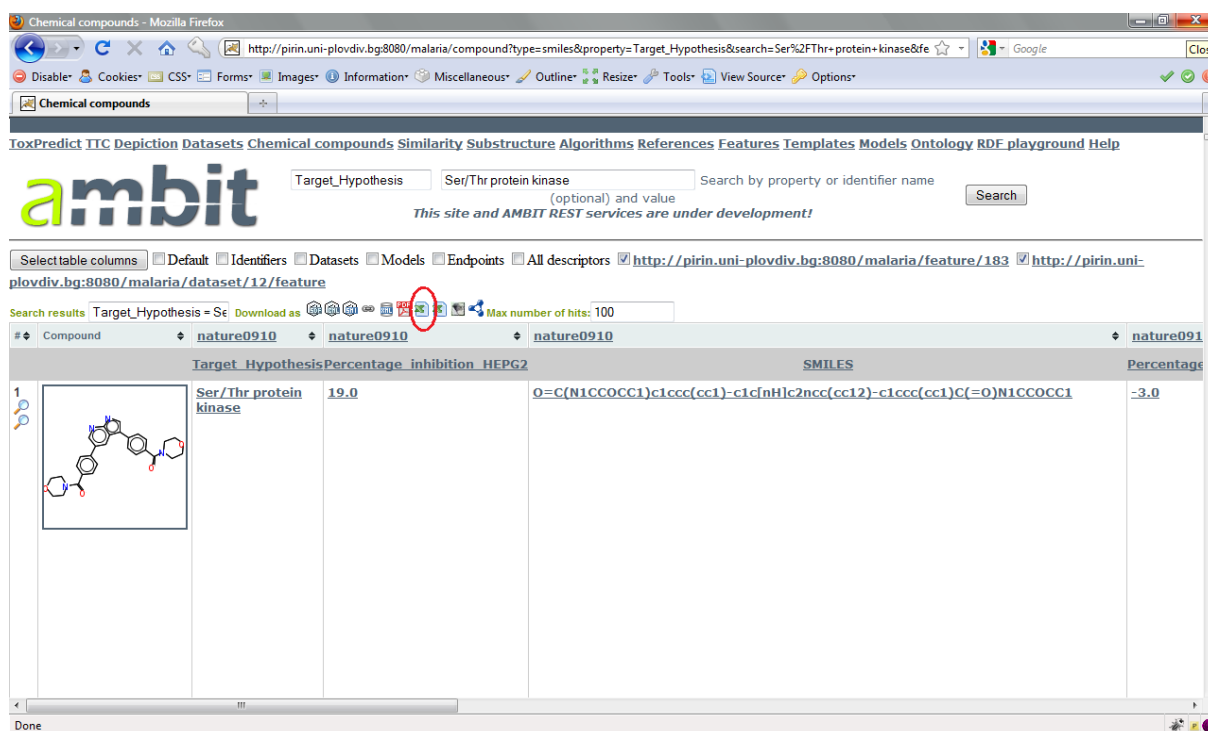


The screenshot shows the Ambit web application interface. At the top, there is a search bar and a navigation menu with links like 'ToxPredict', 'TTC Depiction Datasets', 'Chemical compounds', 'Similarity', 'Substructure', 'Algorithms', 'References', 'Features', 'Templates', 'Models', 'Ontology', 'RDF playground', and 'Help'. The main content area displays a list of datasets, each with a small icon and a title. The dataset 'Tres Cantos Antimalarial TCAMS dataset (screening against GlaxoSmithKline's library)' is highlighted in blue. Below the list, there is a form to 'Add new dataset (SDF, MOL, SMI, CSV, TXT file)' with a 'Browse...' button and a 'Submit' button. At the bottom, there is another form to 'Import properties (SDF, MOL, SMI, CSV, TXT file)' with a 'Browse...' button.

Browse the dataset, find the column “Target hypothesis”. You'll not that most entries are empty (only ~6% of the compounds have a target hypothesis annotated). In the 100 compounds displayed by default when following the link to the TCAMS data, you will only find one entry with value “[Adrenergic](#)”

[receptor antagonist](#)". You could click on the link, which would filter out only compounds with this potential target.

For our purpose, we want the list of compounds annotated to be kinase inhibitors. You could try to increase the number of displayed compounds until you find one, or you could enter "Ser/Thr protein kinase" in the searching text box at the top of the page and click the "Search" button. The results will be displayed as below.



The screenshot shows the OpenTox web interface in a Mozilla Firefox browser. The search bar contains "Ser/Thr protein kinase" and the search button is highlighted. Below the search bar, there are several tabs for "nature0910". A table of search results is displayed with columns for "Target_Hypothesis", "Percentage_inhibition_HEPG2", "SMILES", and "Percentage". The first row shows "Ser/Thr protein kinase" with a percentage of 19.0 and a SMILES string: O=C(N1CCOCC1)c1ccc(cc1)-c1cfnH1c2ncc(cc12)-c1ccc(cc1)C(=O)N1CCOCC1. A chemical structure of this compound is shown in a small window on the left. A red circle highlights a download icon in the search results area.

#	Compound	Target_Hypothesis	Percentage_inhibition_HEPG2	SMILES	Percentage
1	nature0910	Ser/Thr protein kinase	19.0	<chem>O=C(N1CCOCC1)c1ccc(cc1)-c1cfnH1c2ncc(cc12)-c1ccc(cc1)C(=O)N1CCOCC1</chem>	-3.0

To build a model, it is not enough to have a list of Ser/Thr kinase inhibitors. We also need some "negatives". Although strictly speaking we don't have any true negatives, we will use the compounds that do have a target hypothesis annotation – but one that is not "Ser/Thr kinase" – as negatives. So, we extract the whole list of compounds with non-empty target hypothesis, and replace "Ser/Thr kinase" with a "1", and all the other target hypotheses with "0".

To extract the list of compounds with non-empty target hypotheses, use the following URL:

[http://pirin.uni-plovdiv.bg:8080/malaria/compound?type=smiles&property=Target_Hypothesis&search=+&feature_uris\[\]=http://pirin.uni-plovdiv.bg:8080/malaria/feature/183&feature_uris\[\]=http://pirin.uni-plovdiv.bg:8080/malaria/dataset/12/feature&max=1000&condition=!%3D](http://pirin.uni-plovdiv.bg:8080/malaria/compound?type=smiles&property=Target_Hypothesis&search=+&feature_uris[]=http://pirin.uni-plovdiv.bg:8080/malaria/feature/183&feature_uris[]=http://pirin.uni-plovdiv.bg:8080/malaria/dataset/12/feature&max=1000&condition=!%3D)

This operation is not (yet) possible via the "Search" text field (it does not allow negation, e.g. something like Target_Hypothesis !=""), but only via the URL: briefly, the search for non-empty Target Hypothesis is done in the above URL, first with **&search=+** (the "+" stands for empty) – thus searching for all the empties – and then negating the search by **&condition=!%3D** (%3D stands for the "=" sign, thus !%3D stands for !=, or "not equal").

When following the above URL you'll get a table with compounds that have a non-empty Target_Hypothesis. The next step will be to export data. Click on the left one of the two little Excel

icons (when moving the mouse pointer on top of it, a small text box “text/csv” should appear) to save the selected data as CSV.

For the model building, we will use the OpenTox application ToxCreate (<http://toxcreate3.in-silico.ch/toxcreate>). Thus, first we need to format the data as explained at <http://toxcreate3.in-silico.ch/toxcreate/help>. That is, we leave only the SMILES column and the Target_Hypothesis column.

Now you should have the Target_Hypothesis in column 1 (or A), and the SMILES in column 2 (or B). If you are using Excel, go to the cell C2. Type

```
=IF(A2="Ser/Thr protein kinase"; 1; 0)
```

and hit “Enter”. Again click on cell C2 to activate it. Now double-click on the little black square at the bottom-right corner of the cell’s border to fill the column with this formula.

Now, copy the whole column C, and paste it (at the same place) using Excel’s “Paste Special” function, pasting only the values. Once that’s done, delete column A (holding the text entries for the Target_Hypothesis). Delete as well row 1 and save the resulting table as text CSV file to TCAMS-kinase_full.csv.

To calculate the model we will use another OpenTox prototype application, ToxCreate. In your web browser, navigate to <http://toxcreate3.in-silico.ch/toxcreate>. Read the instructions, and try to create a model using your dataset.

ToxCreate being a prototype (as other OpenTox applications), there are still some limitations. You might get an error in the model building, in which case you could try to reduce the number of compounds used to build the model to about 600. Just delete some rows until that table contains 600 rows or less. Save the resulting table to TCAMS-kinase-subset.csv.

Unfortunately, most likely the validation of the model will fail. This is due to memory restrictions on the server where the validation service is hosted.