AMBIT WEB SERVICES

CHEMICAL DATA AND MODELS VIA OPENTOX API

NINA JELIAZKOVA
IdeaConsult Ltd.
Sofia, Bulgaria
www.ideaconsult.net

IDEACONSULT LTD.

Based in Sofia, Bulgaria

Provides consultancy, technical services and software development in the areas of chemoinformatics, QSAR and data mining since 2004;

Develops and maintains several open source software products, designed and developed in close co-operation with academia, regulatory bodies and industry. Examples:

- Toxtree (toxicity prediction, threshold of toxicological concern estimation);
- Toxmatch (for encoding and applying chemical similarity indices);
- Ambit (a QSAR decision support system, including generic database management, structure conversions and searching, as well as applicability domain assessment); Ambit REST web services;
- (Q)SAR model reporting format (QMRF) inventory;

Partner in several EU FP7 funded projects:

CADASTER (2009-2012), OPENTOX (2008-2011), TOXBANK (2010-2015) (SEURAT-1 cluster);
Associated partner in the EU FP7 project OpenPHACTS;

http://www.ideaconsult.net

AMBIT

http://ambit.sf.net

- **AMBIT, AMBIT XT**

CEFIC Long Range Research Initiative project *"Building blocks for a future (Q)SAR decision support system: databases, applicability domain, similarity assessment and structure conversions"*; Ambit XT (Standalone GUI, workflow)

- **AMBIT REST web services**
  - OpenTox Application Programming Interface (API)
  - **Web Applications** using AMBIT REST web services

Software

Highly accessed

Open Access

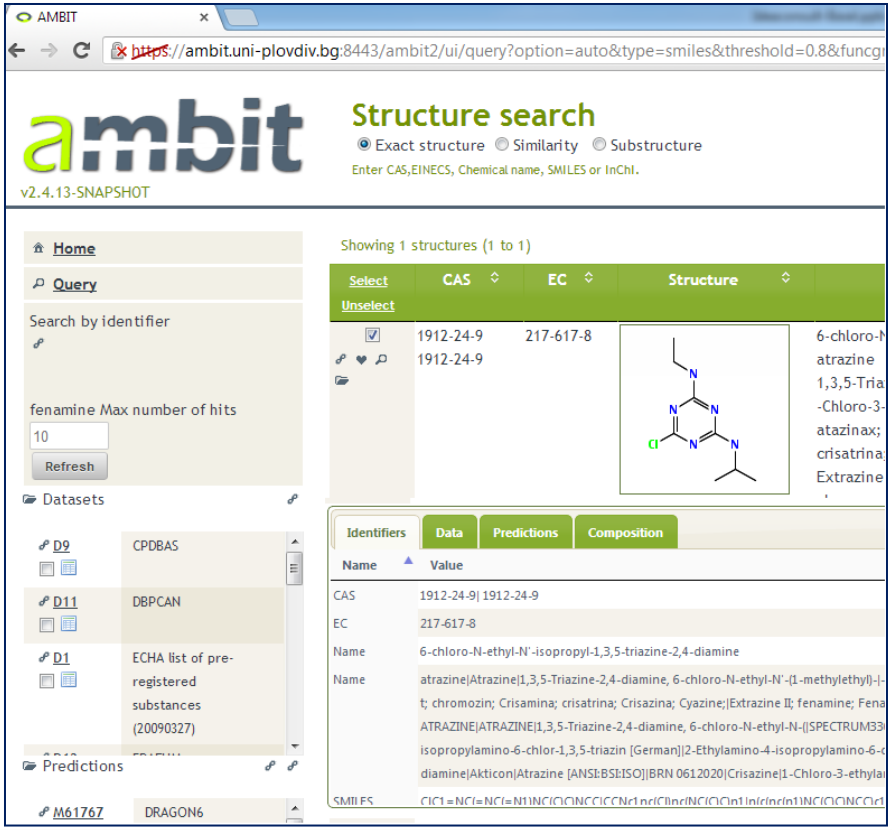
AMBIT RESTful web services: an implementation of the OpenTox application programming interface

Journal of Cheminformatics

AMBIT WEB SERVICES

DATASET WEB SERVICE

- Chemical structure query
- Upload chemical compounds and properties
- Data collation
- Built-in heuristics for automatic discovery of 2D chemical structure inconsistencies;
- Uploading a file with chemical structures and properties makes it automatically available in several formats; structure and similarity searchable;



Available as online services, as a downloadable archive or Virtual Machine appliance

DATASETS WEB SERVICES

- AMBIT uses MySQL with a flexible database schema.
- There are no predefined field names for chemical properties.
- The properties are annotated with ontology entries
- The OpenTox API is database engine and database schema agnostic
- Supports common chemical formats for data import

Default database content :

Aggregation of data from multiple sources

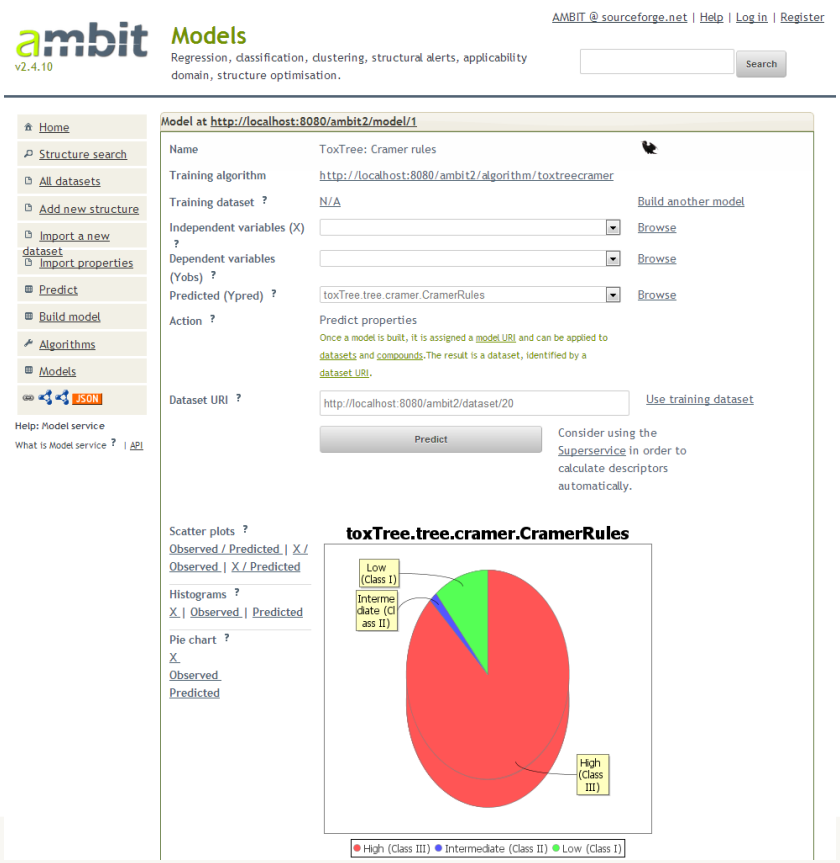
ECHA list of pre-registered substances, EPA DSSTox , ECETOC skin irritation , LLNA skin sensitisation , Bioconcentration factor (BCF) Gold Standard Database , ToxCast , Benchmark Data Set for pKa Prediction, Benchmark Data Set for In Silico Prediction of Ames Mutagenicity , Bursi AMES Toxicity Dataset , EpiSuite data , PubChem subset, confidential data, etc.

Download: The resources are serialized in various formats (RDF, JSON, SDF, etc.)

AMBIT WEB SERVICES

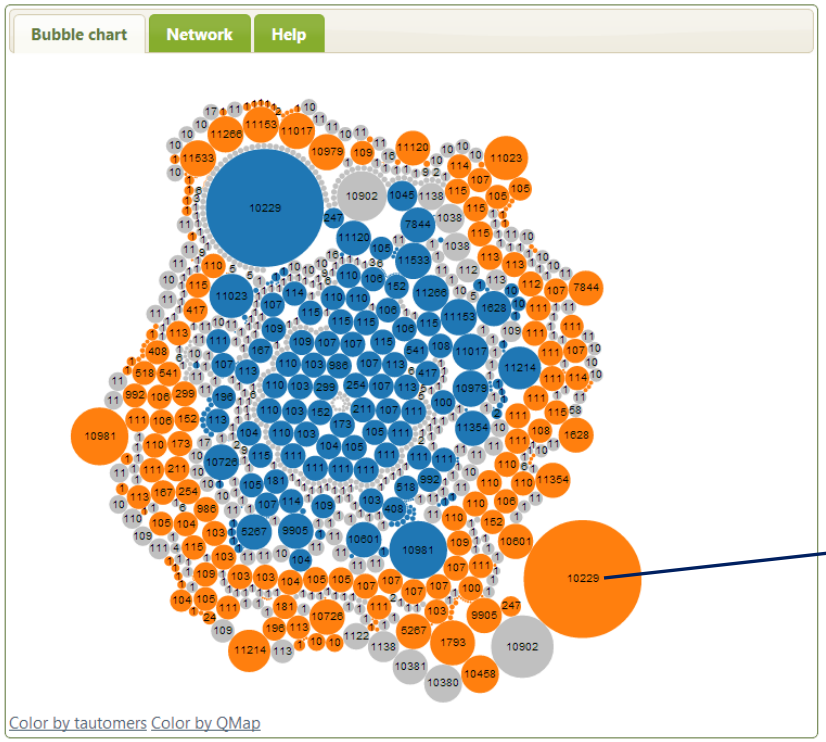
COMPUTING SERVICES

- Descriptor calculation (CDK, MOPAC, Dragon, etc.);
- Model building methods (classification, regression, clustering);
- Expert rules (Toxtree);
- Applicability domain assessment;
- Chemical landscape analysis
- Tautomer generation
- Wrappers for third party tools and web services.

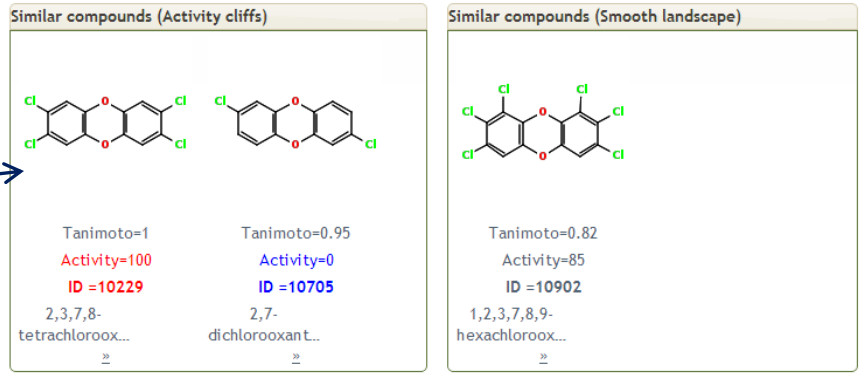


Calculations launched in in an uniform way (REST API)

CHEMICAL LANDCAPES ANALYSIS



An efficient method for identifying activity cliffs, implemented in Ambit and available via OpenTox API



Curr Top Med Chem. 2012;12(18):1987-2001.

Chemical landscape analysis with the OpenTox framework.

Jeliazkova N, Jeliazkov V.
Ideaconsult Ltd., 4 A.Kanchev str., Sofia 1000, Bulgaria. jeliazkova.nina@gmail.com

AMBIT IT TECH

REQUIREMENTS

Server:

- Java 1.6, 1.7
- Tomcat 5.5, 6.x, 7.x
- MySQL 5.5

Clients:

- Modern JavaScript enabled browser
- Desktop applications
- Workflow engines

MAIN LIBRARIES

- The Chemistry Development Kit
- AMBIT SMARTS
- AMBIT SMIRKS
- AMBIT Tautomers
- InChI / JNI-INCHI
- Restlet (REST)
- Machine learning (Weka , etc.)

The clients communicate via
OpenTox API (Application programming Interface)

AMBIT WEB SERVICES AS BUILDING BLOCKS

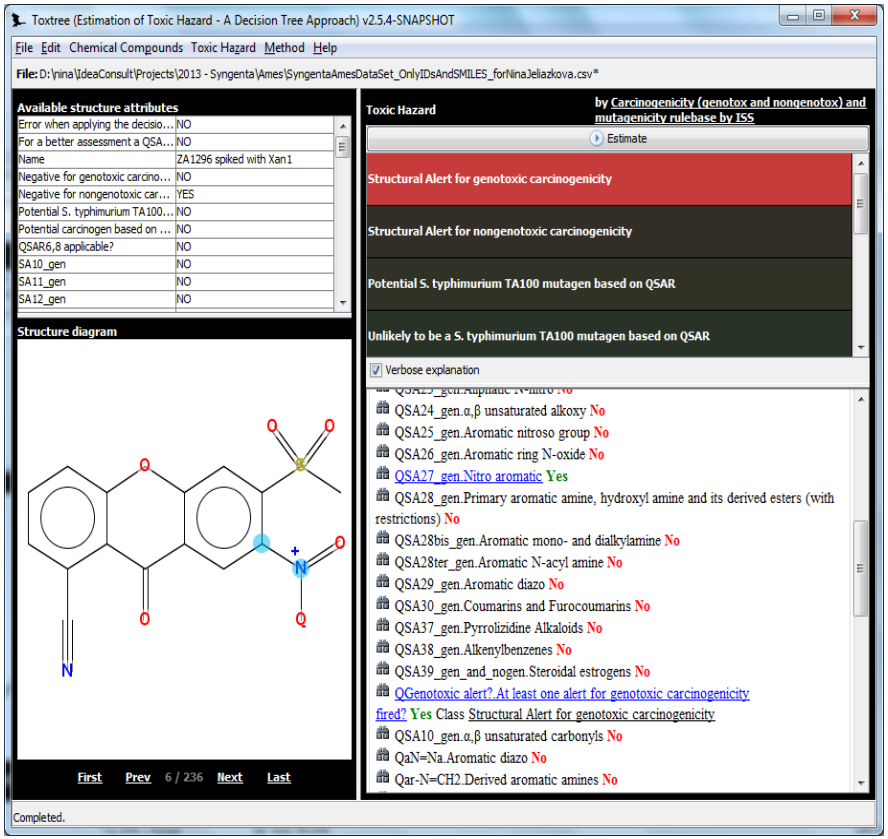
- Web applications
- http://ToxPredict.org – aggregates remote predictions
- OpenTox API Wrapper for SOAP web services
- QMRF database (server & client)
- Applicability domain used by CADASTER web site (client)
- REPDOSE structure & similarity search www.repdose.de
- Xenobiotics Metabolism Database XMetDB (server & client)
- Bioclipse – OpenTox (client)
- CheS-Mapper (client)
- Toxbank structure search (server & client)
- Toxtree web edition (AJAX client)

TOXTREE 2.6.0

Estimates toxic hazard by applying expert defined rules, arranged as a decision tree. Commissioned 2005 by JRC.

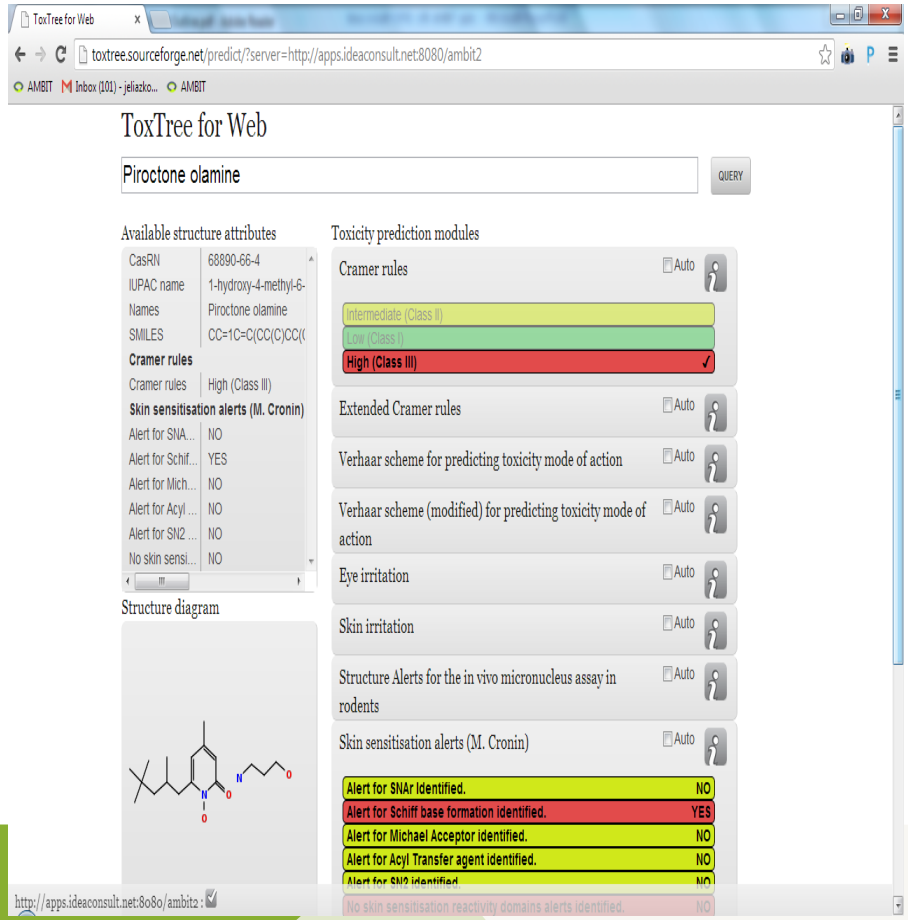
17 plugins: Cramer rules , Verhaar scheme , Skin irritation, Eye irritation, Mutagenicity and carcinogenicity, Biodegradation, Reactivity domains, Protein binding, TTC Decision tree, SMARTCyp , in vivo micronucleus assay in rodents , Structural Alerts for Functional Groups, etc.

New decision trees can be built with the GUI or by developing new plug-ins in Java code



Download at
http://toxtree.sourceforge.net

HTTP://TOXTREE.SF.NET/PREDICT



AJAX interface for Toxtree

Frontend (JavaScript only)

- anywhere (copy JS and html to your site)

Backend (Ambit web services)

- anywhere (uri parameter)

Code at github

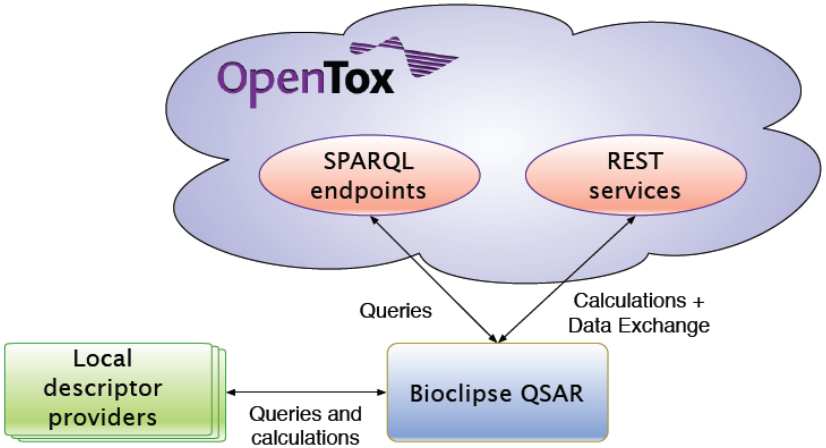
https://github.com/ideaconsult/Toxtree.js

Web interface for Toxtree
Ambit web services serve as a backend database and calculation

AMBIT WEB SERVICES AS BUILDING BLOCKS

Desktop applications

- Bioclipse (www.bioclipse.net) dynamically discovers computational algorithms exposed via the OpenTox servers
- There is a (SPARQL) registry of available data and computational services http://apps.ideaconsult.net:8080/ontology
- When a new descriptor algorithm or model is registered on the OpenTox ontology service, it will automatically be picked up by Bioclipse.

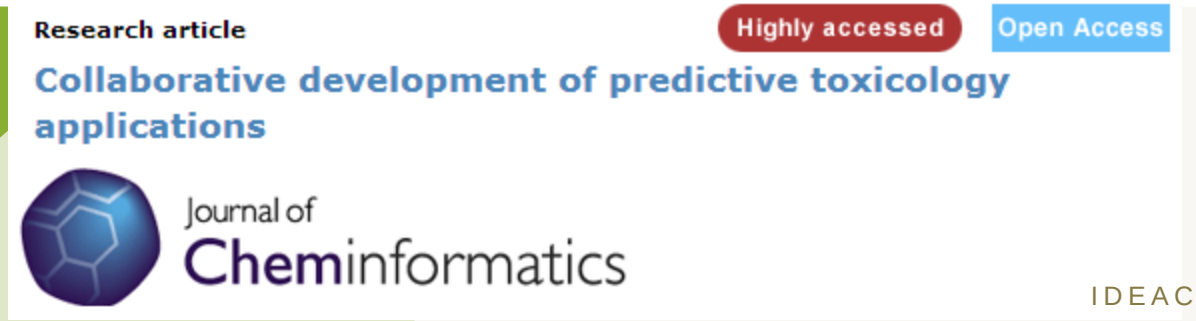


Willighagen E., Jeliazkova N., Hardy B., Grafstrom R., Spjuth O., Computational toxicology using the OpenTox application programming interface and Bioclipse, BMC Research Notes 2011 4 (1), 487

EU FP7 PROJECT OPENTOX

An Open Source Predictive Toxicology Framework

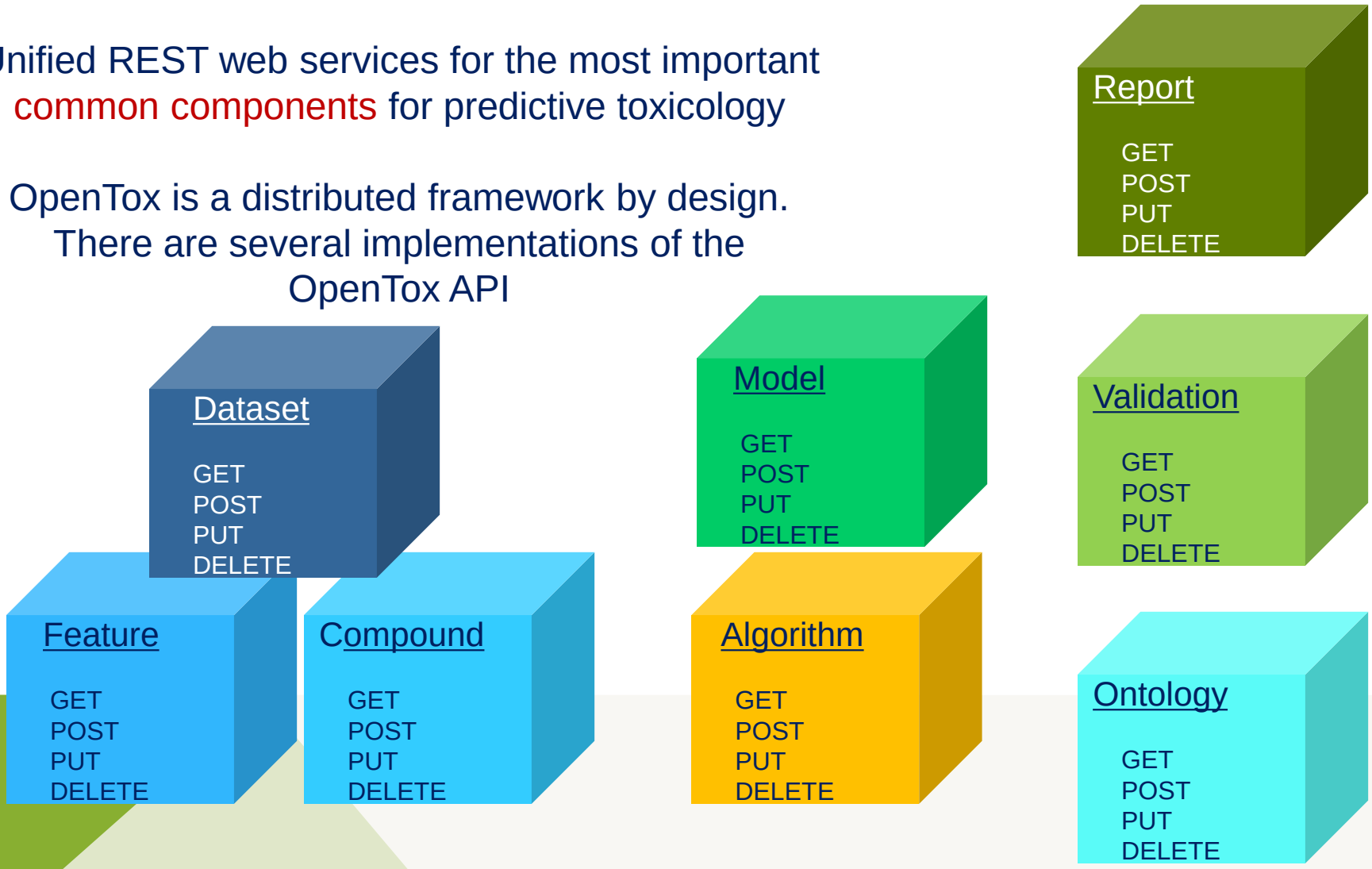
- European Commission Framework Program 7 funded project (2008-2011)
- Objective: develop a distributed framework for predictive toxicology.
- The building blocks considered are : data, chemical structures, algorithms and models.
- The framework allows to build models, apply models, validate models, access and query data in various ways.
- Technologies used are REST style web services and W3C Resource Description Framework for description of services.
- 11 partners



OPENTOX API

Unified REST web services for the most important common components for predictive toxicology

OpenTox is a distributed framework by design. There are several implementations of the OpenTox API



OPENTOX API CRASH COURSE *

1) Find a compound by an identifier, structure, similarity, substructure

curl –X GET **http://host/query/compound/search/all?search=caffeine**

- Returns the URI of the compound http://host/compound/328

curl –X GET http://host/query/smarts?search=c1cccnc1-c2ncccc2

- Returns URIs of the hits http://host/compound/456

2) Find a predictive model

curl –X GET http://host/model

Returns URI of the available models, e.g. http://host/model/8

3) Apply the model to the compound

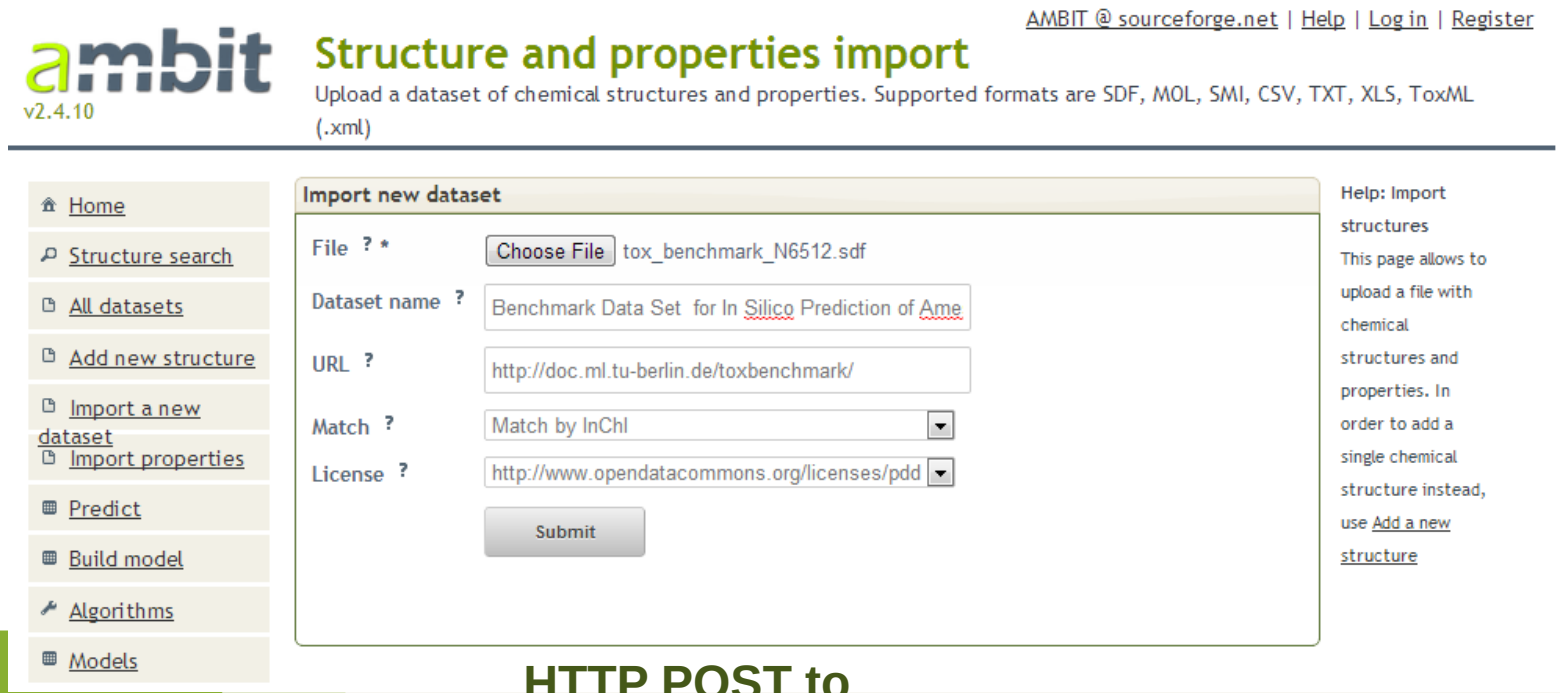
curl –X POST http://host/model/8 -d "dataset_uri=http://host/compound/328"

- Returns URI of the results, e.g. **http://host/dataset/999**
- The results can be retrieved in all chemical MIME formats, as well RDF/XML, N3, CSV, ARFF, JSON, JSONP

* with the help of cURL http://curl.haxx.se/

AMBIT: IMPORT A NEW DATASET

Uploads datasets in common chemical formats; No predefined fields; Unique URI assigned



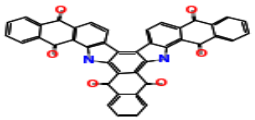
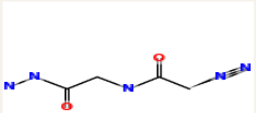
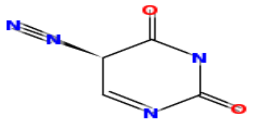
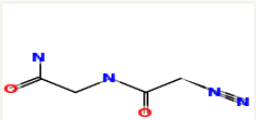
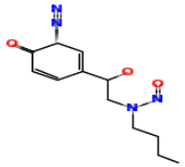
HTTP POST to
http://host:port/ambit2/dataset
http://host:port/ambit2/compound
Uniform interface: (OpenTox web services API)

THE DATASET UPLOADED

Download as JSON |

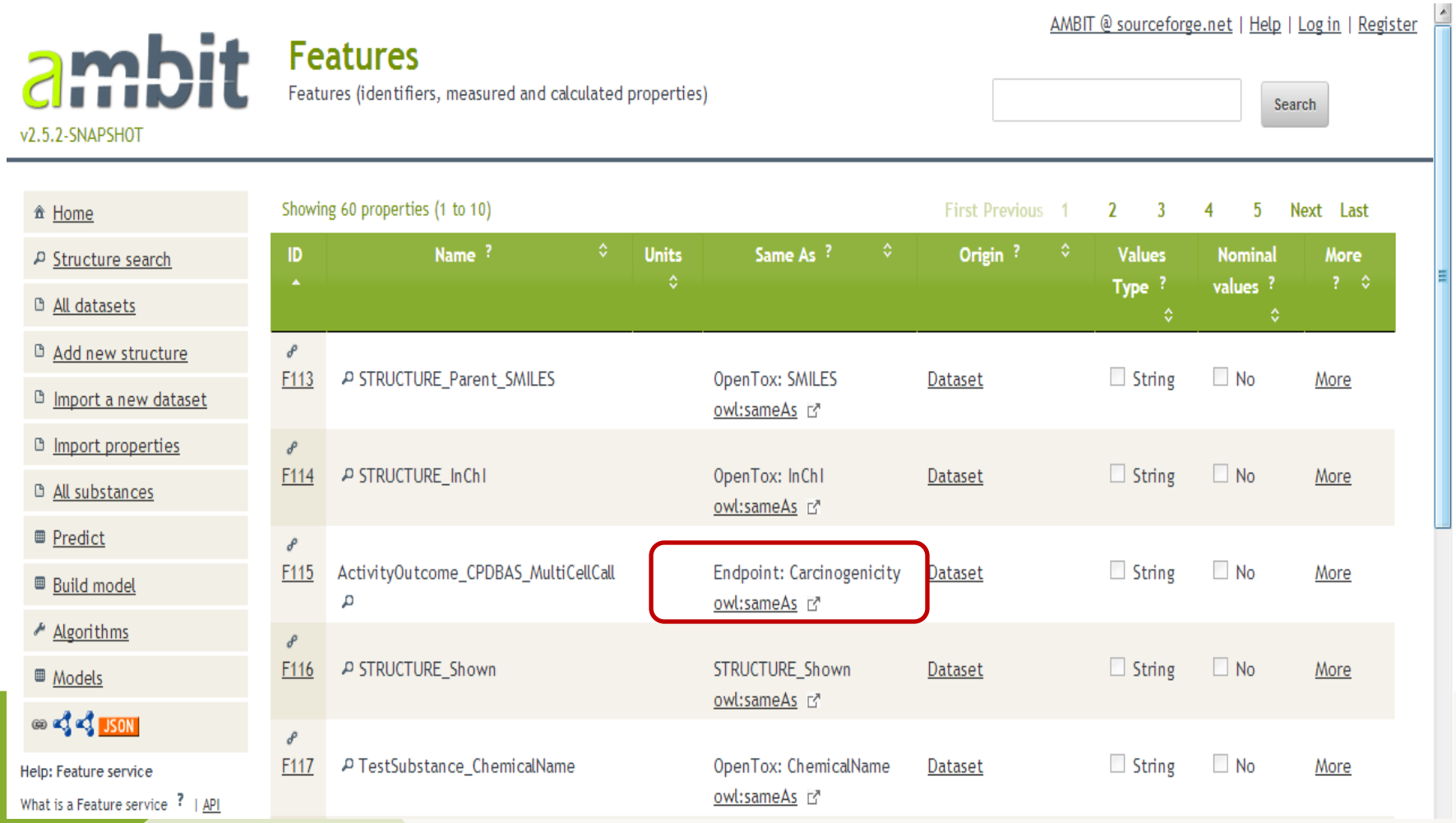
Page 0 Page size 100

Show 10 entries Search:

			CAS	Name	Molecular Weight	MC Pred	DEREK Example	MC Example	Source	REFERE
	1.		2475-33-4	.	646.602	0	0	0	VITIC	JUDSON COOKE, DOERRE
	2.		820-75-7	.	157.131	0	0	0	CCRIS	MCCANI CHOI,E, YAMASA
	3.		2435-76-9	.	138.084	1	0	0	CCRIS	HIRAMC KATO,T KIKUGA
	4.		817-99-2	.	142.116	0	0	0	CCRIS	MCCANI CHOI,E, YAMASA
	5.		116539-70-9	.	264.28	1	0	0	CCRIS	KIKUGA KATO,T TAKEDA

HTTP GET to http://host:port/ambit2/dataset/{id}

VIEW DATASET COLUMNS



HTTP GET to
http://host:port/ambit2/dataset/{id}/feature

ambit
v2.5.2-SNAPSHOT

Structure search

Exact structure | Similarity | Substructure

Enter SMARTS or draw a structure.

AMBIT @ sourceforge.net | Help | Log in | Register

Nitriles — Functional groups

N#C — Draw (sub)structure

Search

Home

Query

Substructure search

N#C Max number of hits

10

Refresh

Datasets

D4 — CPDBAS: Carcinogenic Potency Database Summary Tables - All Species Database

D1 — ECHA Preregistration list

Predictions

M61 — 130326084909.e0...eccc-4893-9150-dc4893cc6a97.w...

M62 — 130326085806.d7...8cc5-4289-9ff5-bdbdf9dcb2c2....

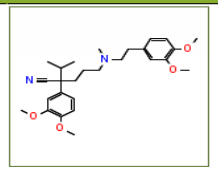
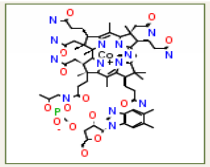
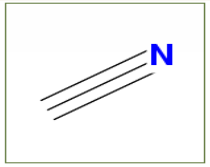
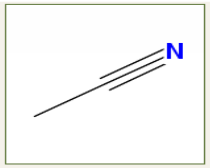
M63 — 130326105220.3d...d2b7-45bd-a05f-03972fe79a7e.w...

Download

JSON

Showing 10 structures (1 to 10)

First Previous 1 Next Last

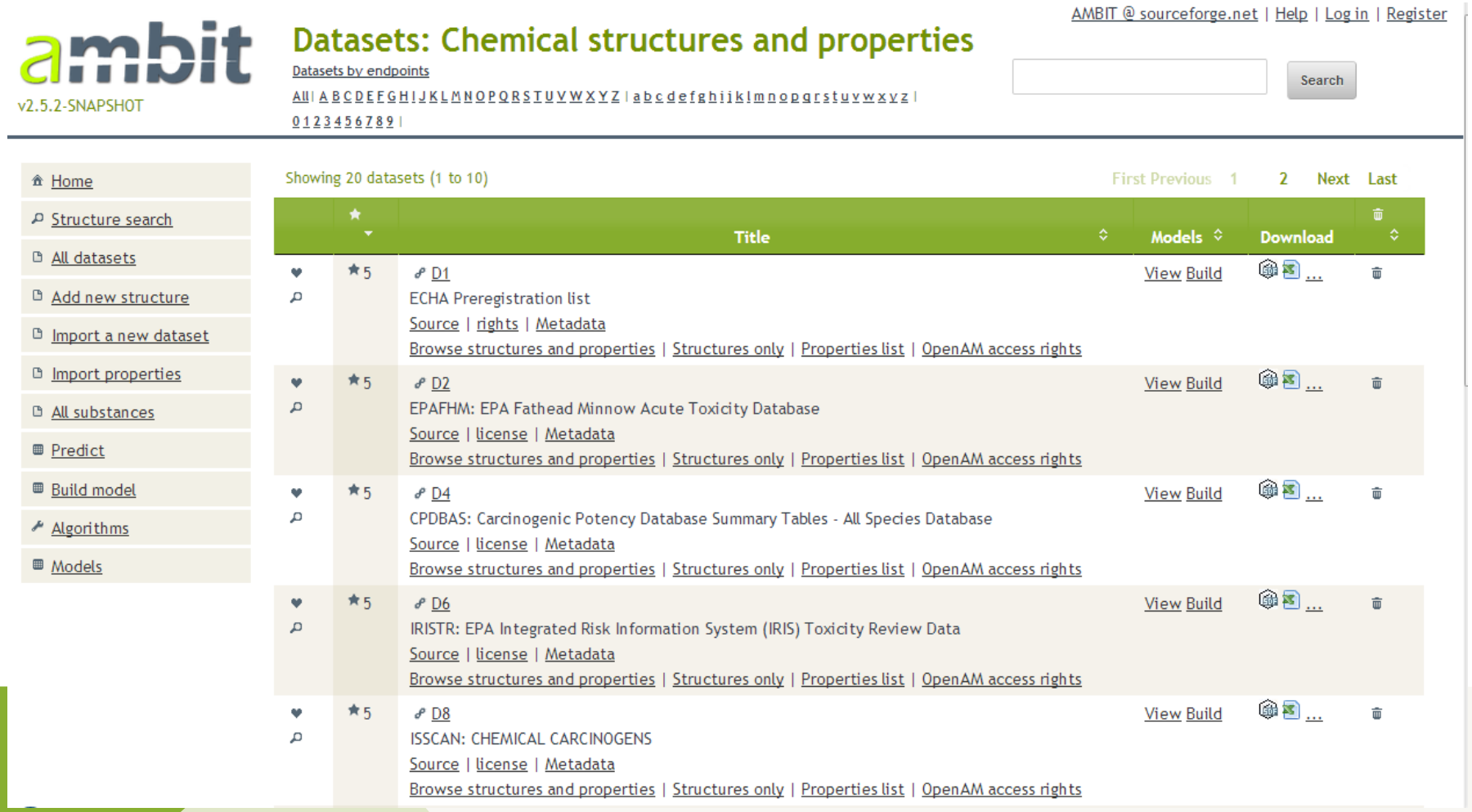
Select / Unselect	CAS	EC	Structure	Name	SMILES	InChI Key
	52-53-9	200-145-1		verapamil		
	68-19-9	200-680-0		cyanocobalamin		
	74-90-8	200-821-6		hydrocyanic acid hydrogen cyanide	C#N	
	75-05-8	200-835-2		acetonitrile AN; Cyanomethane; Ethane nitrile; ethanonitrile; Ethyl nitrile; methanecarbonitrile; Methyl cyanide. ACETONTRL	CC#N	

Identifiers | Data | Predictions | Composition

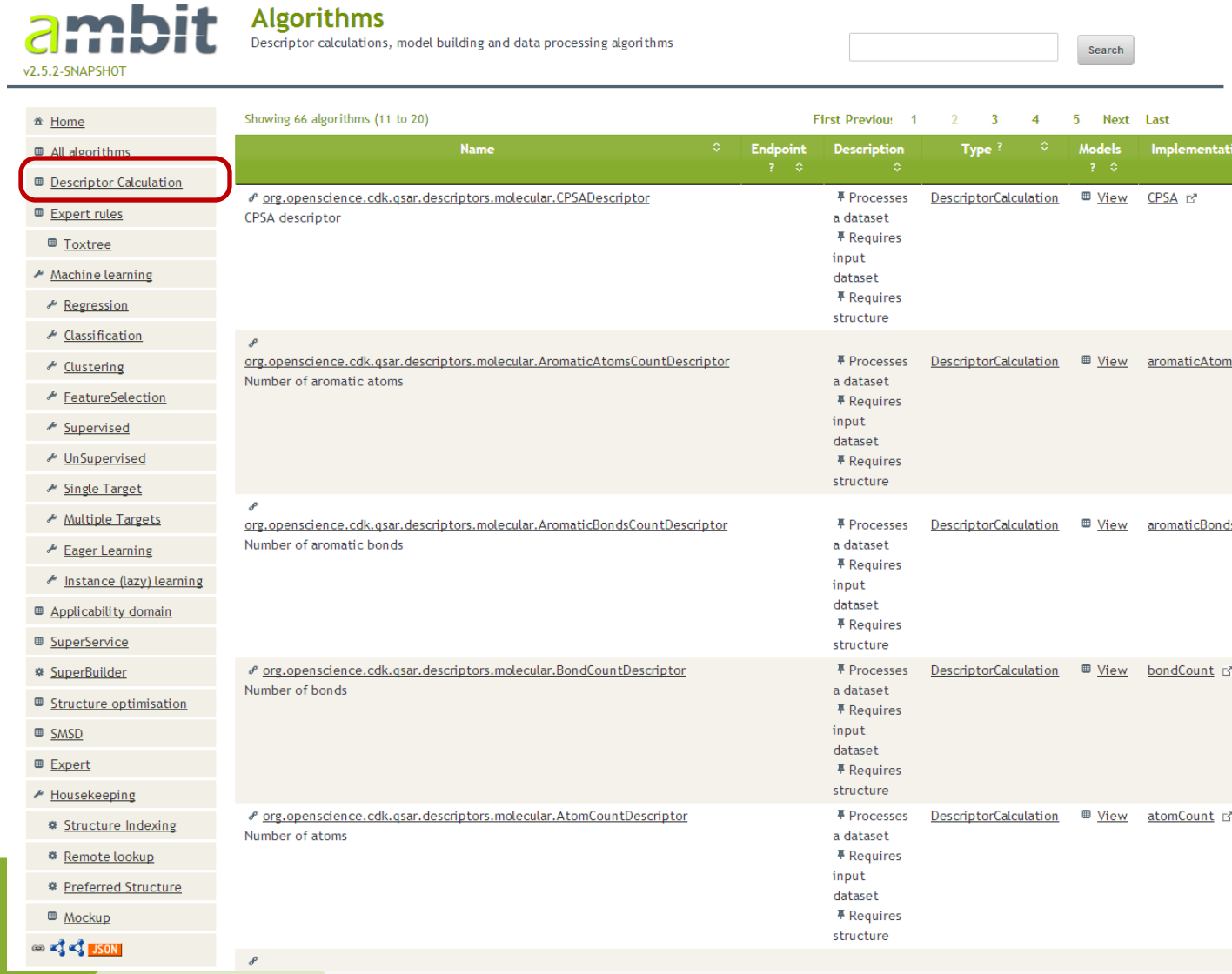
Data source	Property	Value	Endpoint
CPDBAS_v5d_1547_20Nov2008	ActivityOutcome_CPDBAS_MultiCellCall	inactive	Carcinogenicity
CPDBAS_v5d_1547_20Nov2008	STRUCTURE_Shown	tested chemical	STRUCTURE_Shown
CPDBAS_v5d_1547_20Nov2008	ActivityScore_CPDBAS_Rat	0	ActivityScore_CPDBAS_Rat
CPDBAS_v5d_1547_20Nov2008	ActivityOutcome_CPDBAS_SingleCellCall	inactive	Carcinogenicity
CPDBAS_v5d_1547_20Nov2008	TD50_Rat_Note	no positive results	Carcinogenicity
CPDBAS_v5d_1547_20Nov2008	STRUCTURE_MolecularWeight	41.052	STRUCTURE_MolecularWeight
CPDBAS_v5d_1547_20Nov2008	DSSTox_CID	9	DSSTox_CID
CPDBAS_v5d_1547_20Nov2008	NTP_TechnicalReport	TR 447	NTP TechnicalReport

HTTP GET

SEARCH ALL DATASETS

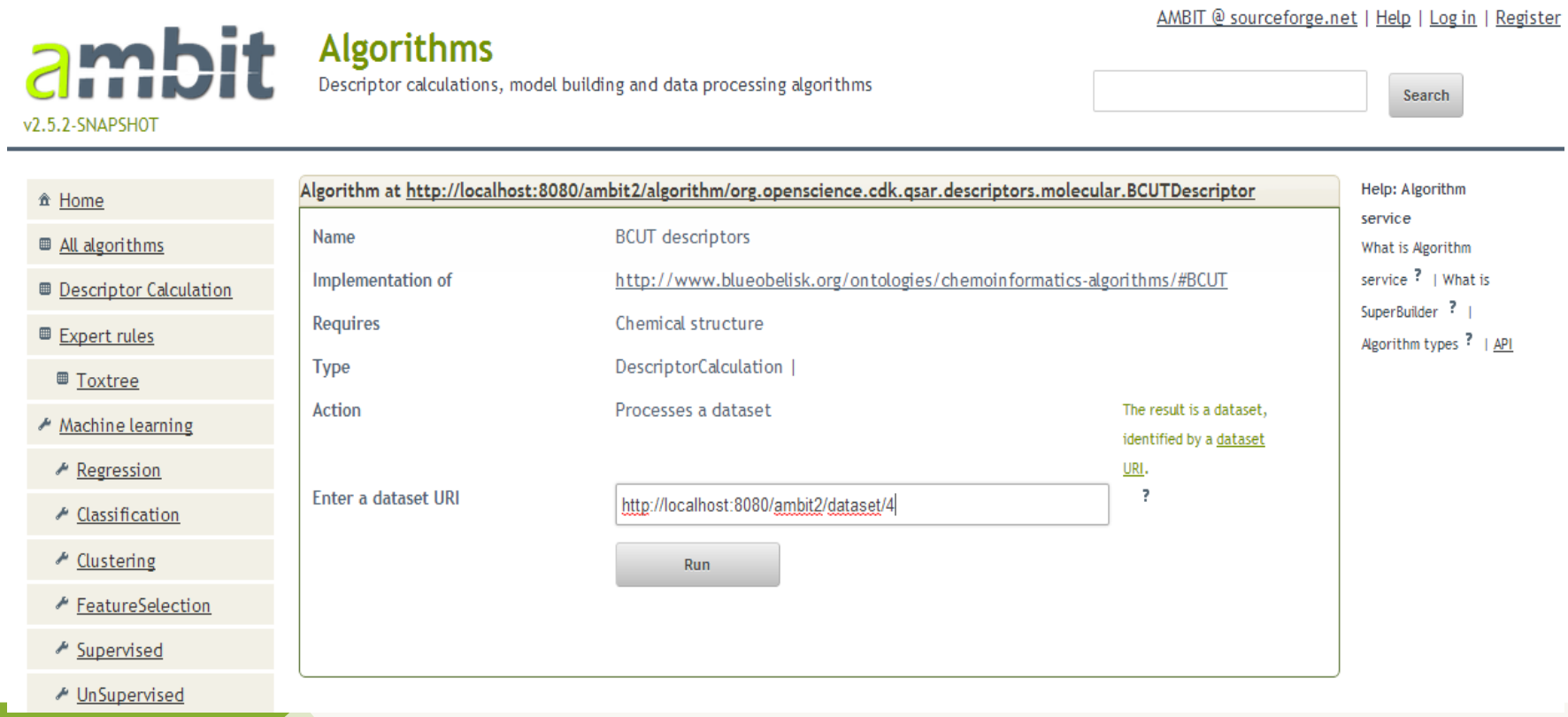


HTTP GET
http://host:port/ambit2/dataset



HTTP GET
http://host:port/ambit2/algorithm

CALCULATE DESCRIPTORS EXAMPLE:BCUT (CDK)

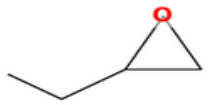
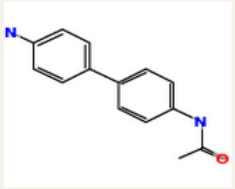
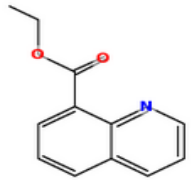
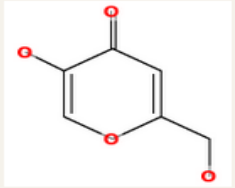


HTTP POST to
http://host:port/ambit2/algorithm
Parameter:
dataset_uri=http://host:port/ambit2/dataset/id

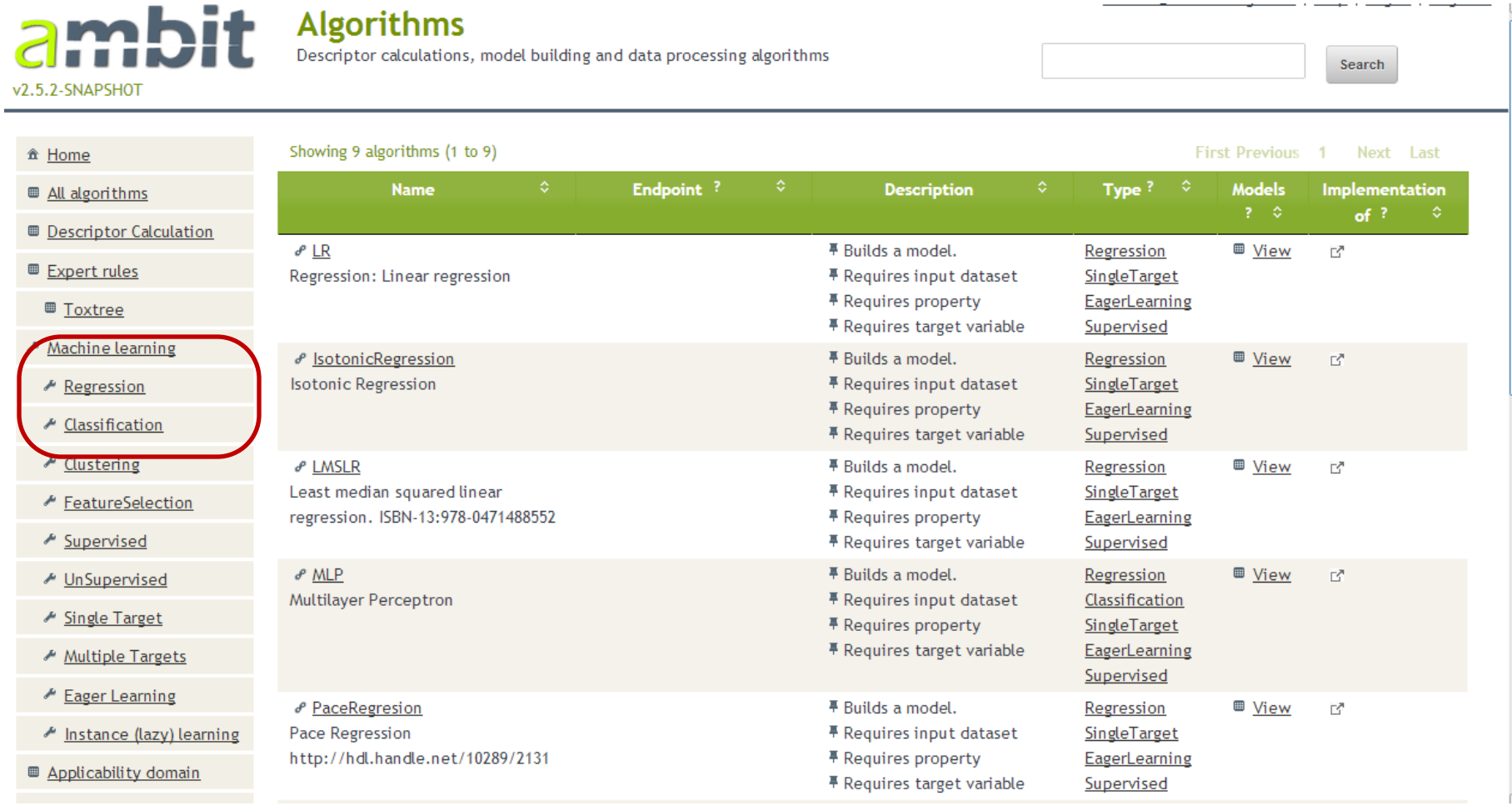
THE TRAINING DATASET (ACTIVITY & DESCRIPTORS)

0 Page size 100

v 10 entries Search:

		Activity	BCUTw-1l	BCUTw-1h	BCUTc-1l	BCUTc-1h	BCUTp-1l
1.		1	11.997	15.997	-0.394	0.102	4.502
2.		1	11.85	15.995	-0.303	0.201	4.242
3.		0	11.85	15.996	-0.312	0.295	4.8
4.		1	11.996	15.999	-0.389	0.213	4.26

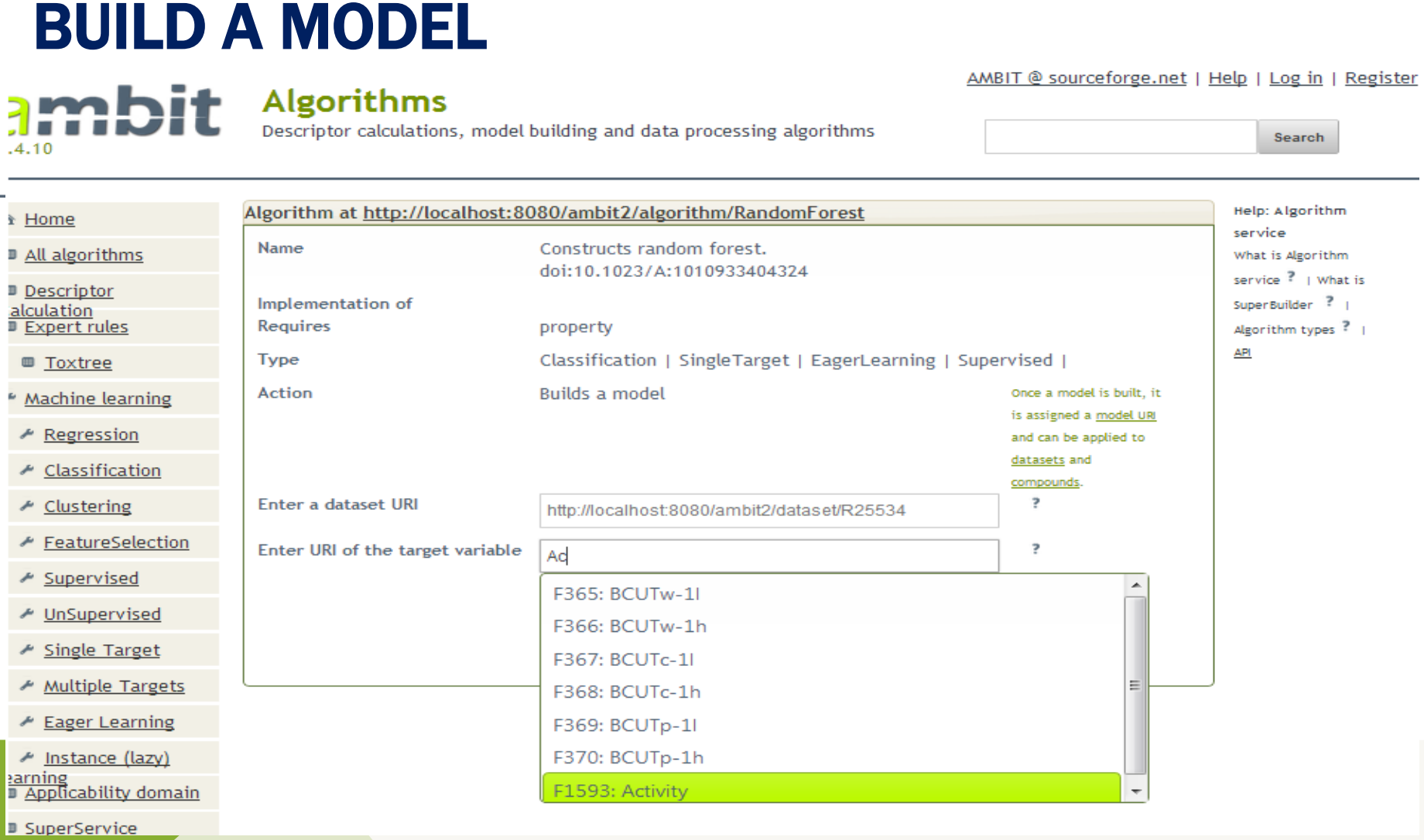
HTTP GET
http://host:port/ambit2/dataset/{id}



HTTP GET

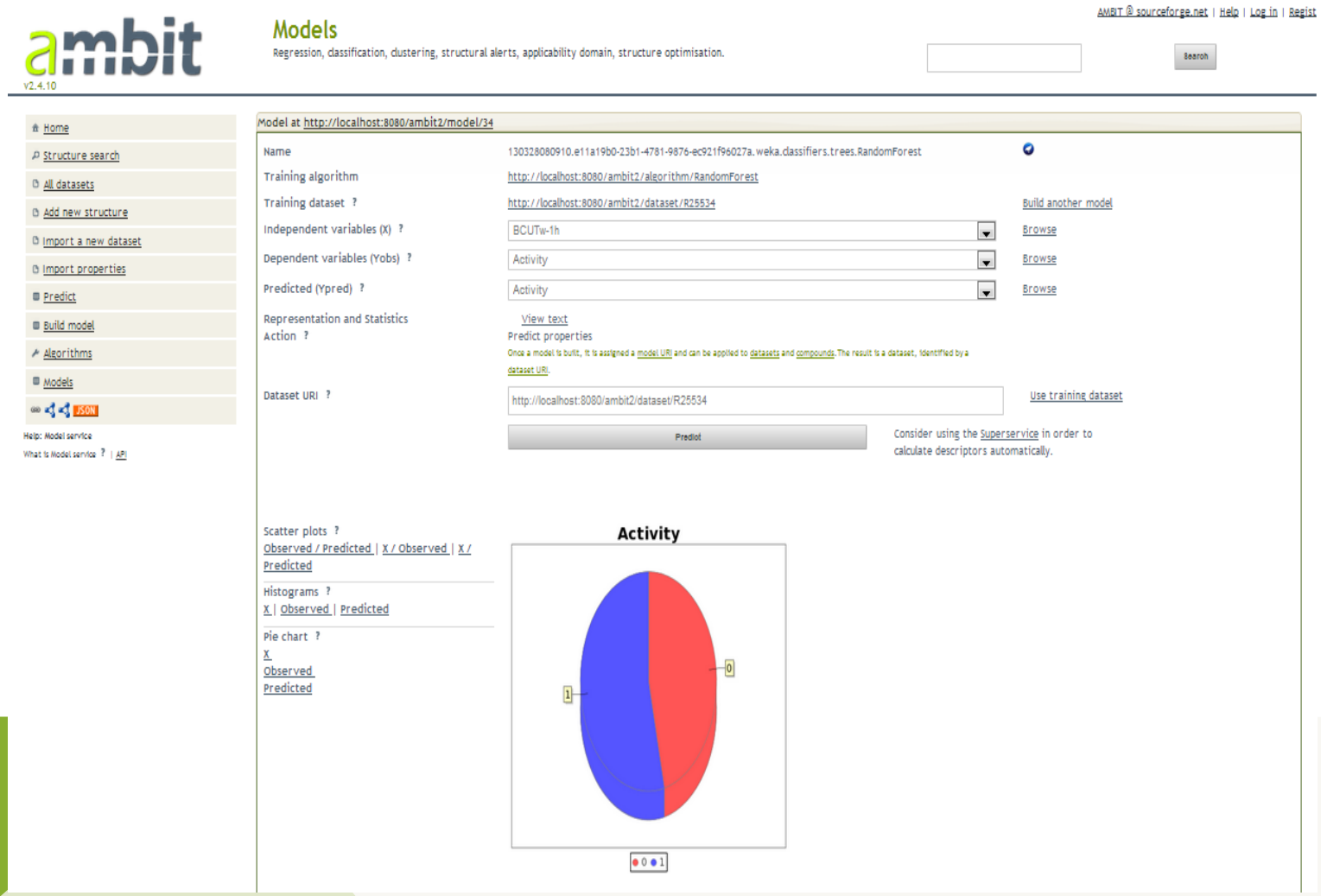
http://host:port/ambit2/algorithm

BUILD A MODEL



HTTP POST to
http://host:port/ambit2/algorithm/{id}
Creates a model
http://host:port/ambit2/model/{id}

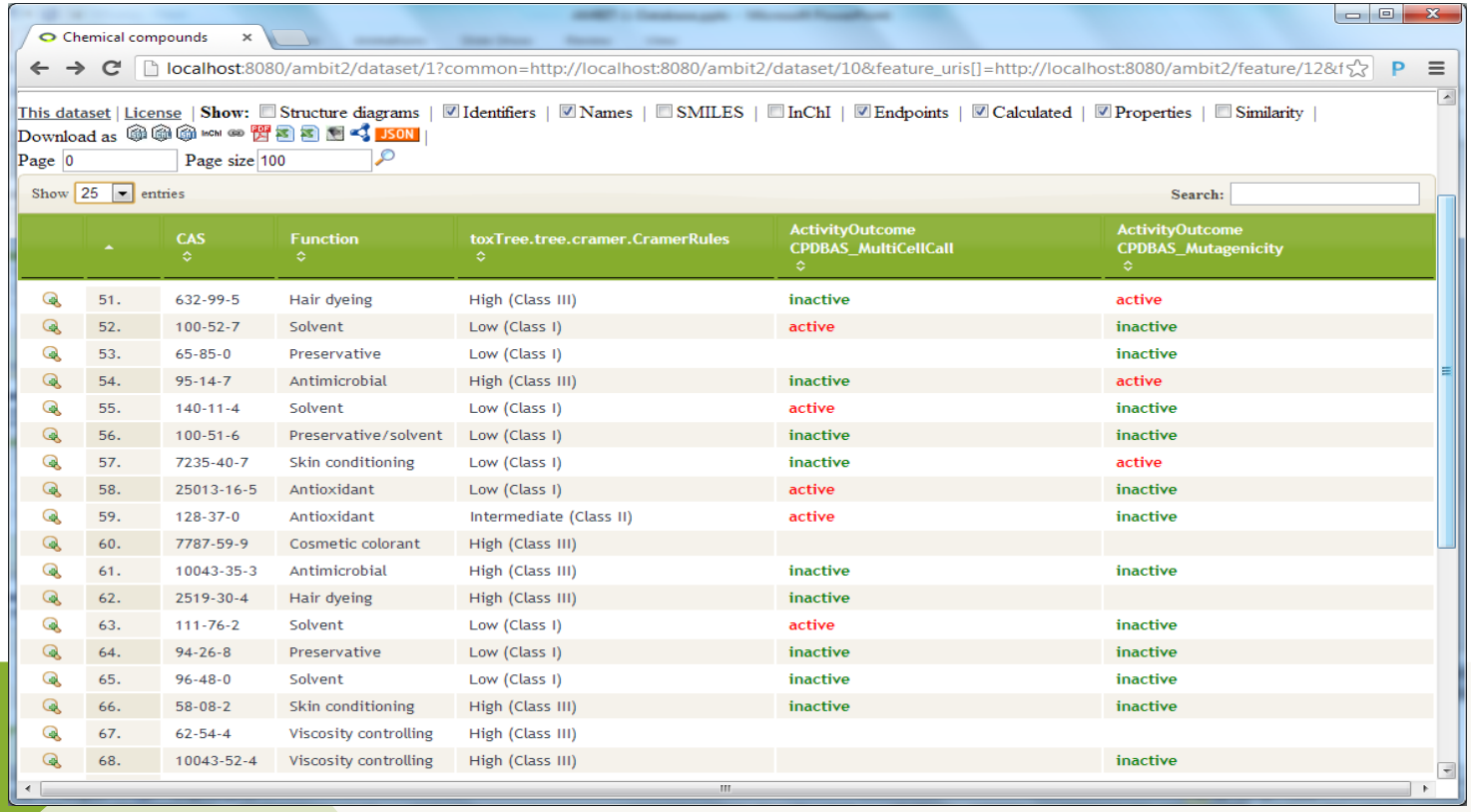
THE NEW MODEL



HTTP GET
http://host:port/ambit2/model/{id}

PREDICTION RESULTS

- Stored in the database. Also assigned web addresses (URIs)
- Collation with experimental data as previously shown (URI parameters)
- Downloadable in multiple formats

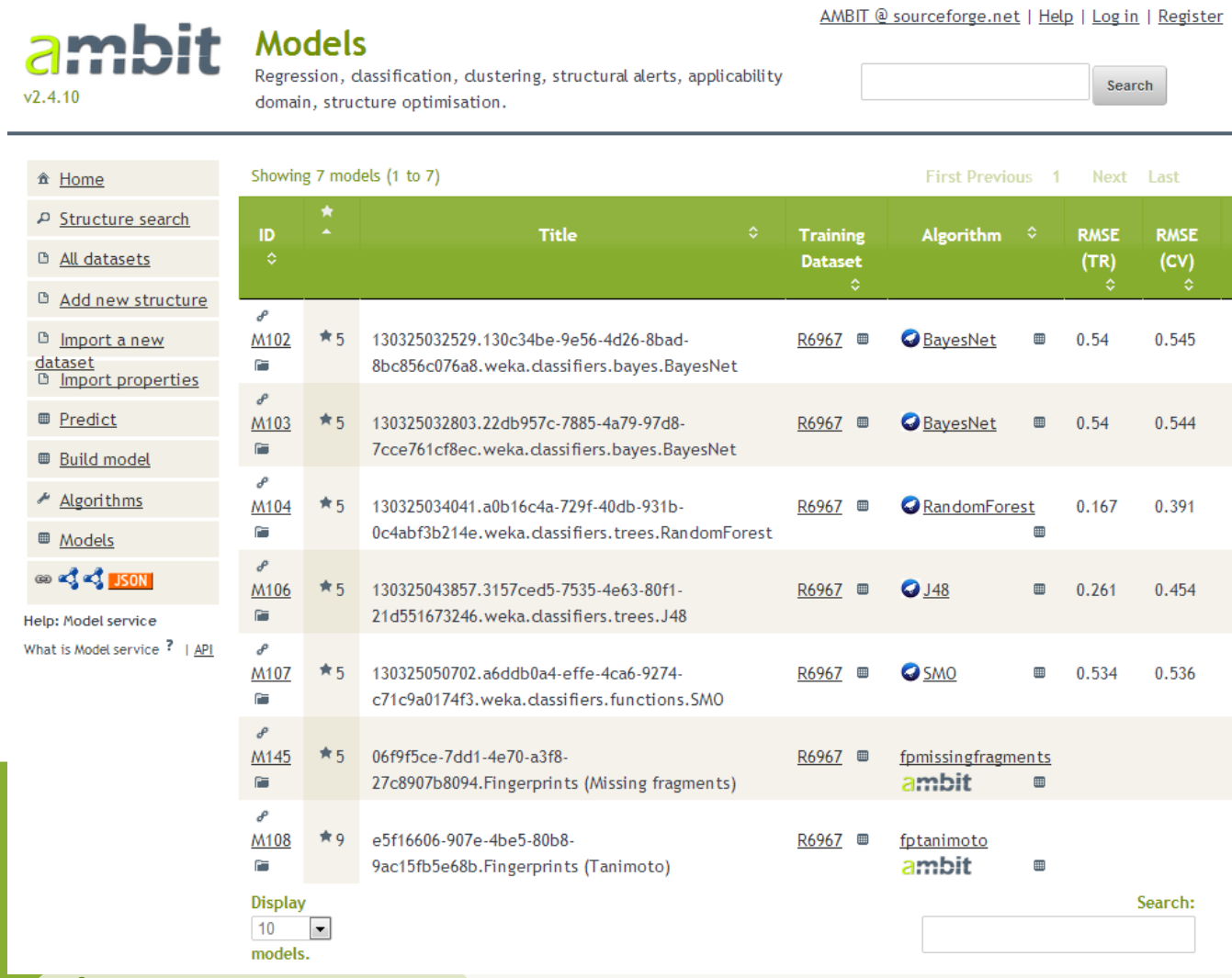


	CAS	Function	toxTree.tree.cramer.CramerRules	ActivityOutcome CPDBAS_MultiCellCall	ActivityOutcome CPDBAS_Mutagenicity
51.	632-99-5	Hair dyeing	High (Class III)	inactive	active
52.	100-52-7	Solvent	Low (Class I)	active	inactive
53.	65-85-0	Preservative	Low (Class I)		inactive
54.	95-14-7	Antimicrobial	High (Class III)	inactive	active
55.	140-11-4	Solvent	Low (Class I)	active	inactive
56.	100-51-6	Preservative/solvent	Low (Class I)	inactive	inactive
57.	7235-40-7	Skin conditioning	Low (Class I)	inactive	active
58.	25013-16-5	Antioxidant	Low (Class I)	active	inactive
59.	128-37-0	Antioxidant	Intermediate (Class II)	active	inactive
60.	7787-59-9	Cosmetic colorant	High (Class III)		
61.	10043-35-3	Antimicrobial	High (Class III)	inactive	inactive
62.	2519-30-4	Hair dyeing	High (Class III)	inactive	
63.	111-76-2	Solvent	Low (Class I)	active	inactive
64.	94-26-8	Preservative	Low (Class I)	inactive	inactive
65.	96-48-0	Solvent	Low (Class I)	inactive	inactive
66.	58-08-2	Skin conditioning	High (Class III)	inactive	inactive
67.	62-54-4	Viscosity controlling	High (Class III)		
68.	10043-52-4	Viscosity controlling	High (Class III)		inactive

Use a model:
HTTP POST to http://host:port/model/id
The result is a dataset URI

MODELS COMPARISON

Webpage Screenshot



HTTP GET
http://host:port/ambit2/model

HOW TO CREATE/PUBLISH A MODEL

Upload the training set and rebuild the model;

- Example: Upload the training dataset and run a classification algorithm

Develop an OpenTox API compatible solutions, allowing to train and run predictive models;

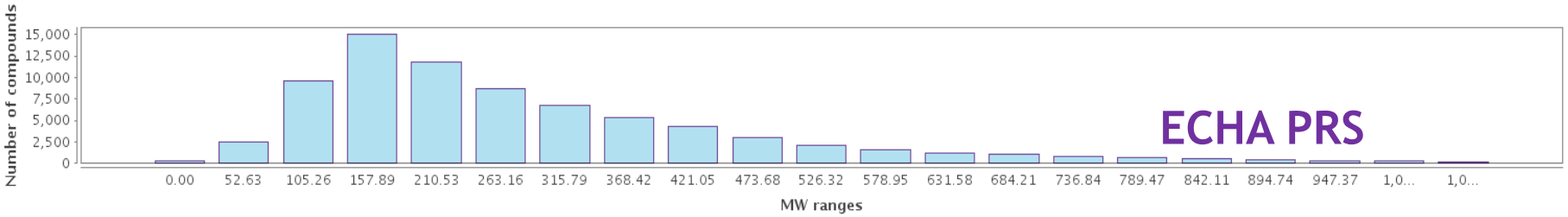
- Example: Lazar models, OpenTox partner models (BG, DE, CH, GR, RU)

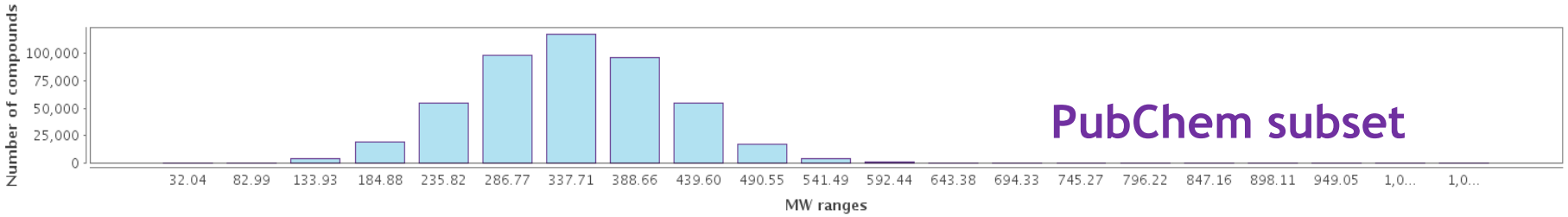
Use thin wrappers for third-party models, and exposing them through the compatible web service API.

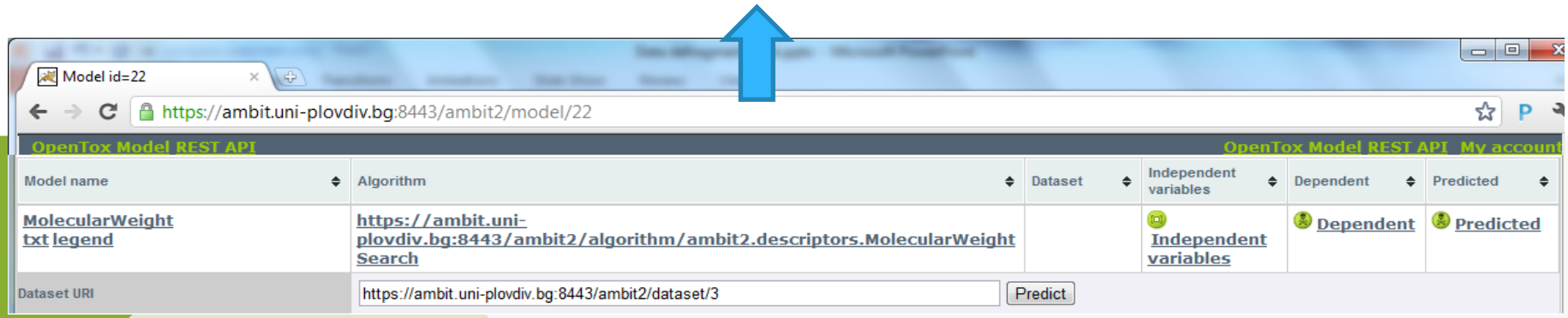
- Models available via SOAP services only had been exposed as OpenTox API compatible models using this approach

All models become potentially visible to client applications (ToxPredict, Bioclipse), subject to access rights.

DATASET COMPARISON
(USING AMBIT ALGORITHMS AND MODELS)







CHEMBL 16 EXAMPLE

Structures(1.2 mln) uploaded as a dataset

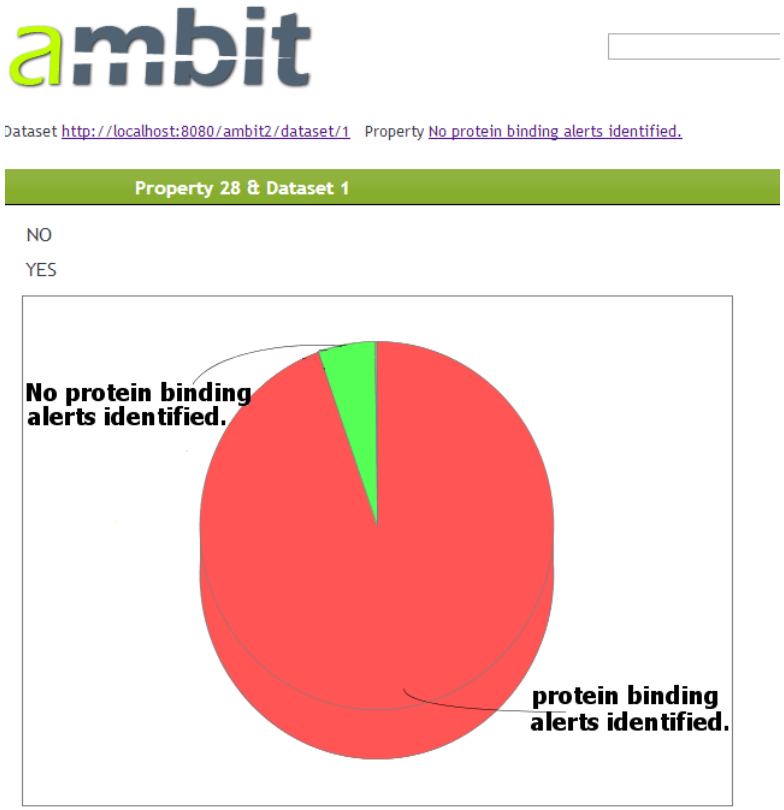
Toxtree protein binding prediction

available in AMBIT via OpenTox Algorithm API

results on the right

Compared with ECHA Preregistration Lis1

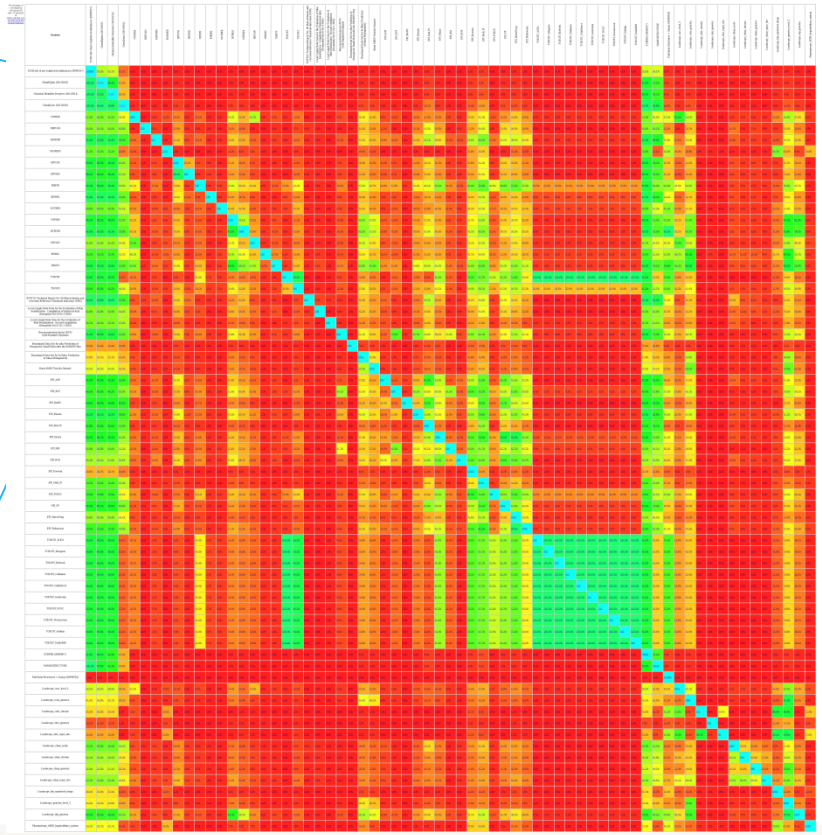
Only 10% (~13000) of ECHA PRS entries (~130000) are found in ChEMBL (~1.2 mln)



DATASET COMPARISON (MULTIPLE DATASETS)

Dataset	ECHA li	ChemIDp	Chemica	ChemDra	CPDBAS	DBPCAN	EPAFHM	FDAMDD
ECHA li	143835	55.9%	50.7%	15.7%	0.7%	0.1%	0.4%	0.6%
ChemIDp	100.0%	80468	88.4%	27.9%	1.2%	0.2%	0.7%	1.1%
Chemica	100.0%	97.5%	72985	30.5%	1.4%	0.2%	0.8%	1.2%
ChemDra	100.0%	99.6%	98.8%	22519	7.9%	0.6%	2.1%	1.0%
CPDBAS	67.3%	66.9%	67.2%	43.9%	1485	0.3%	6.5%	10.1%
DBPCAN	65.6%	65.1%	65.6%	60.8%	3.4%	209	6.2%	0.6%
EPAFHM	92.5%	92.5%	92.4%	78.3%	15.6%	2.1%	617	2.4%
FDAMDD	71.7%	71.7%	71.7%	19.8%	12.3%	0.6%	4.2%	1215
HPVCSI	89.6						1%	1.5%
HPVISD	88.1						8%	1.0%
IRISTR	90.1%	89.4%	90.1%	54.5%	40.1%	1.7%	17.4%	1.1%
KIERBL	95.3%	95.3%	95.0%	70.9%	11.9%	2.2%	6.1%	2.2%
NCTRER	69.8%	69.4%	69.4%	52.2%	22.4%	1.3%	10.8%	8.2%
NTPBSI	88.7%	88.1%	88.5%	67.6%	32.2%	2.0%	9.6%	6.4%
NTPHTS	92.3%	92.1%	92.1%	73.8%	32.2%	2.3%	12.6%	7.3%
ISSCAN	61.6%	61.4%	61.6%	45.9%	72.1%	0.3%	7.2%	9.0%
ISSMIC	76.7%	76.7%	76.7%	42.7%	48.7%	0.7%	10.7%	13.3%
ISSSTY	90.6%	90.1%	90.6%	72.6%	64.6%	3.1%	16.6%	7.2%
TOXCST	90.0%	89.4%	89.0%	13.2%	18.7%	0.0%	7.7%	0.3%
TXCST2	80.9%	80.8%	80.2%	46.6%	26.4%	1.7%	10.4%	6.3%
ECETOC	96.5%	95.8%	96.5%	77.5%	9.9%	6.3%	13.4%	1.4%
Local L	65.9%	65.9%	65.9%	56.3%	18.3%	1.9%	8.7%	2.4%
Local L	66.7%	66.7%	65.7%	57.4%	10.2%	0.0%	3.7%	1.9%
Bioconc	91.3%	89.9%	90.0%	70.5%	23.8%	0.6%	13.8%	1.3%
Benchma	25.4%	25.4%	25.4%	21.6%	1.2%	1.1%	2.2%	0.5%
Benchma	37.8%	37.7%	37.7%	29.1%	9.0%	1.0%	3.5%	2.8%
Bursi A	46.6%	46.6%	46.6%	36.5%	10.8%	1.1%	4.2%	2.9%
EPI_AOP	85.3%	85.0%	85.1%	79.3%	16.0%	3.0%	15.7%	0.7%
EPI_BCF	80.4%	80.4%	80.1%	65.3%	23.7%	0.4%	15.3%	1.2%
EPI_Bio	84.9%	84.9%	84.9%	82.6%	7.0%	0.6%	7.6%	0.0%

Compounds from dataset EPAFHM, found in the dataset ECHA list of pre-registered substances (20090327)
Number = 571 Percent = 92.54%

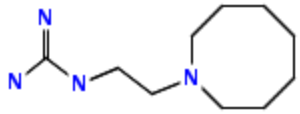


Legend:

No overlap — 100% overlap

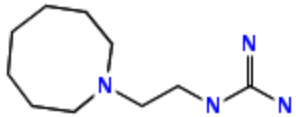
BUILT-IN HEURISTICS TO DISCOVER 2D CHEMICAL STRUCTURE INCONSISTENCIES

55-65-2
guanethidine



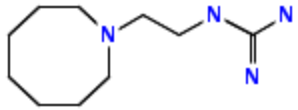
Consensus Quality
Label **Majority[1:3]**
Structure Quality
Label **ProbablyOK**

55-65-2
guanethidine



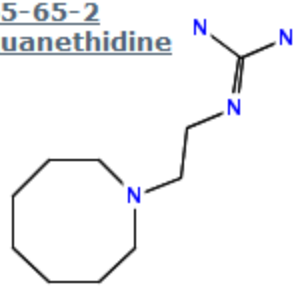
Consensus Quality
Label **Majority[1:3]**
Structure Quality
Label **ProbablyOK**

55-65-2
guanethidine



Consensus Quality
Label **Majority[1:3]**
Structure Quality
Label **ProbablyOK**

55-65-2
guanethidine



Consensus Quality
Label **Majority[1:3]**
Structure Quality
Label **ProbablyERROR**

Automatic Classification	Initial Quality Label Assigned
Consensus	OK
Majority	Probably OK for the structure that belongs to the majority Probably ERROR for the structure(s) that belong(s) to the minority
Ambiguous	Unknown (multiple sources)
Unconfirmed	Unknown (single source)

NEXT STEPS

- Substance composition support
- IUCLID5 import (substances and endpoint study records)
- Read across
- More descriptors and models
- Tautomer - enriched models (AMBIT-TAUTOMER) see poster #6)
- Metabolite generation (via SMARTCYP and AMBIT-SMIRKS package) (see poster #7)
- Improved user interface for model building
- JSON serialization of all resources
- JavaScript embeddable widgets
- Improved client libraries **https://github.com/ideaconsult/opentox-cli**
- Improved documentation and examples

  https://github.com/ideaconsult/examples-ambit
- Customized distributions (also VM)

THANK YOU

QUESTIONS?

Downloads:

http://ambit.sf.net

Acknowledgements:

Colleagues at Ideaconsult

Collaborators at Plovdiv University

OpenTox partners

The Open Source community

- CDK and Bioclipse in particular

Funding:

- CEFIC-LRI, FP7 OpenTox, FP7 Cadaster