

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>

7/18/2013

AMBIT

- **AMBIT, AMBIT XT**

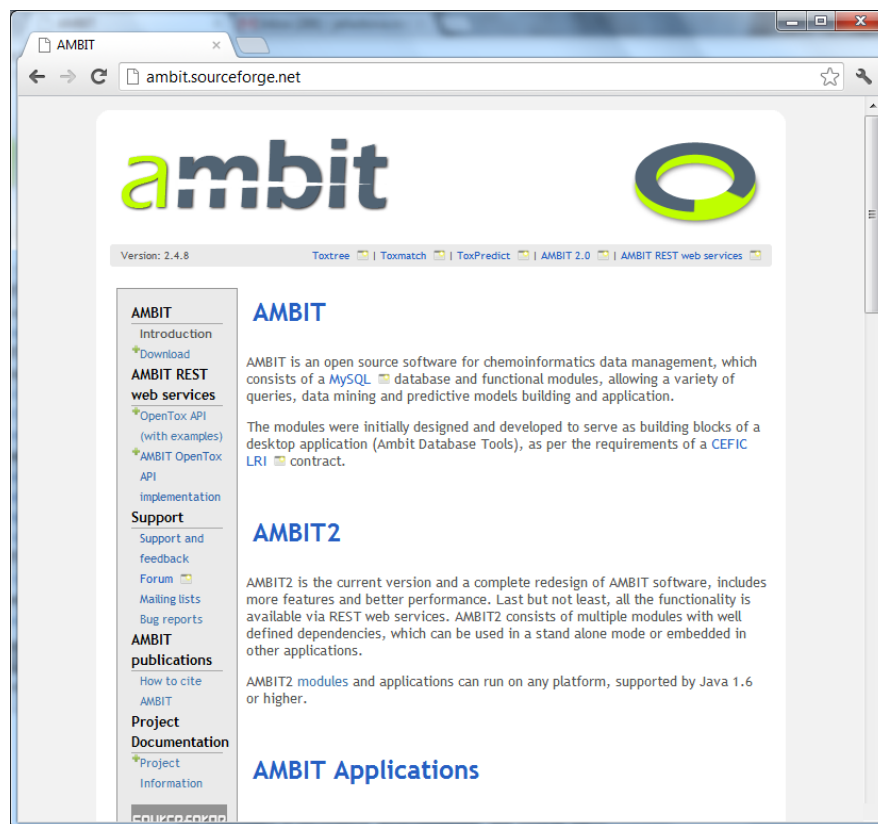
CEFIC Long Range Research Initiative project *“Building blocks for a future (Q)SAR decision support system: databases, applicability domain, similarity assessment and structure conversions”*; Ambit XT (Standalone GUI, workflow)

- **AMBIT REST web services**

- OpenTox Application Programming Interface (API)

- **Web Applications** using AMBIT REST web services

<http://ambit.sf.net>



Software

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



Journal of
Cheminformatics

Highly accessed

Open Access

11/4/2013

IDEACONSULT LTD.

3

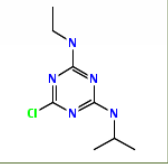
AMBIT WEB SERVICES

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;

The screenshot displays the AMBIT web service interface. The browser address bar shows the URL: <https://ambit.uni-plovdiv.bg:8443/ambit2/ui/query?option=auto&type=smiles&threshold=0.8&funcn>. The page title is "Structure search". Below the title, there are radio buttons for "Exact structure" (selected), "Similarity", and "Substructure". A text input field is labeled "Enter CAS, EINECS, Chemical name, SMILES or InChI". The version "v2.4.13-SNAPSHOT" is displayed below the logo.

The main content area shows "Showing 1 structures (1 to 1)". A table lists the search results:

Select	CAS	EC	Structure
<input checked="" type="checkbox"/>	1912-24-9	217-617-8	
<input type="checkbox"/>	1912-24-9		

Below the table, there is a detailed view of the selected structure. It includes tabs for "Identifiers", "Data", "Predictions", and "Composition". The "Identifiers" tab is active, showing a table of identifiers and their values:

Name	Value
CAS	1912-24-9 1912-24-9
EC	217-617-8
Name	6-chloro-N-ethyl-N'-isopropyl-1,3,5-triazine-2,4-diamine
Name	atrazine[Atrazine][1,3,5-Triazine-2,4-diamine, 6-chloro-N-ethyl-N'-(1-methylethyl)-]; chromozin; Crisamina; crisatrina; Crisazina; Cyazine; Extrazine II; fenamine; Fena-ATRAZINE[ATRAZINE][1,3,5-Triazine-2,4-diamine, 6-chloro-N-ethyl-N-(SPECTRUM33) isopropylamino-6-chlor-1,3,5-triazin (German)]2-Ethylamino-4-isopropylamino-6-diamine[Aktion][Atrazine [ANSE:BSE:ISO]]BRN 0612020]Crisazine[1-Chloro-3-ethyla
SMILES	ClC1=NC(=NC(=N1)NC(C)C)CNC1n(C)nc(C)nc1

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.

ambit Models
v2.4.10

Regression, classification, clustering, structural alerts, applicability domain, structure optimisation.

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

Search

Model at <http://localhost:8080/ambit2/model/1>

Name: ToxTree: Cramer rules

Training algorithm: <http://localhost:8080/ambit2/algorithm/toxtreecramer>

Training dataset: N/A [Build another model](#)

Independent variables (X): [Browse](#)

Dependent variables (Yobs): [Browse](#)

Predicted (Ypred): ToxTree.tree.cramer.CramerRules [Browse](#)

Action: Predict properties

Once a model is built, it is assigned a [model URI](#) and can be applied to [datasets](#) and [compounds](#). The result is a dataset, identified by a [dataset URI](#).

Dataset URI: <http://localhost:8080/ambit2/dataset/20> [Use training dataset](#)

Consider using the [Superservice](#) in order to calculate descriptors automatically.

Scatter plots: [Observed / Predicted](#) | [X /](#)
[Observed](#) | [X / Predicted](#)

Histograms: [X](#) | [Observed](#) | [Predicted](#)

Pie chart: [X](#)
[Observed](#)
[Predicted](#)

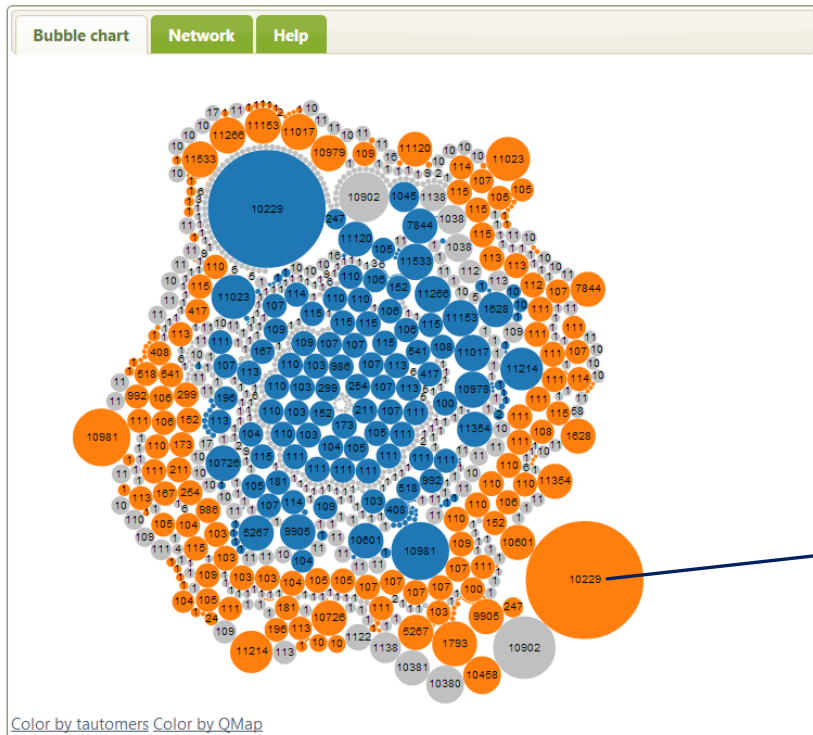
toxTree.tree.cramer.CramerRules

Low (Class I)
Intermediate (Class II)
High (Class III)

High (Class III) Intermediate (Class II) Low (Class I)

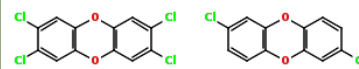
Calculations launched in in an uniform way (REST API)

CHEMICAL LANDSCAPES ANALYSIS



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

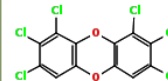
Similar compounds (Activity cliffs)



Tanimoto=1
Activity=100
ID =10229
2,3,7,8-
tetrachloroox...

Tanimoto=0.95
Activity=0
ID =10705
2,7-
dichlorooxant...

Similar compounds (Smooth landscape)



Tanimoto=0.82
Activity=85
ID =10902
1,2,3,3,7,8,9-
hexachloroox...

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

11/4/2013

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

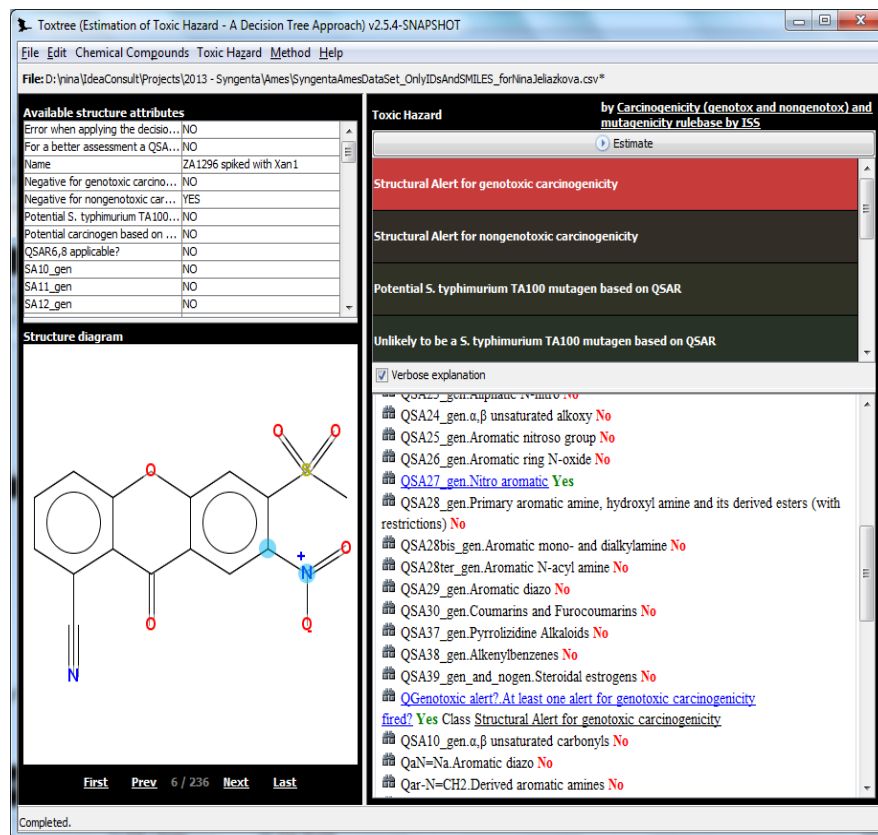
- Web applications
- <http://ToxPredict.org> – aggregates remote predictions
- OpenTox API Wrapper for SOAP web services
- QMRF database (server & client)
- Applicability domain used by CADASTER web site (client)
- REPDOSE structure & similarity search www.repdose.de
- Xenobiotics Metabolism Database 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 plugins in Java code



The screenshot displays the Toxtree software interface. The window title is "Toxtree (Estimation of Toxic Hazard - A Decision Tree Approach) v2.5.4-SNAPSHOT". The file path is "D:\yina\IdeaConsult\Projects\2013 - Syngenta\Ames\SyngentaAmesDataSet_Only\IDsAndSMILES_forNinaJelazkova.csv".

Available structure attributes:

Error when applying the decision tree...	NO
For a better assessment a QSAR...	NO
Name	ZA1296 spiked with Xan1
Negative for genotoxic carcinogenicity...	NO
Negative for nongenotoxic carcinogenicity...	YES
Potential S. typhimurium TA100 mutagen based on QSAR	NO
Potential carcinogen based on QSAR	NO
QSAR6,8 applicable?	NO
SA10_gen	NO
SA11_gen	NO
SA12_gen	NO

Structure diagram:

CC(=O)N(C)C1=CC=C2C(=C1)C(=O)OC3=CC=C(C#N)C=C23

Toxic Hazard: by Carcinogenicity (genotoxic and nongenotoxic) and mutagenicity rulebase by ISS

Estimate

- Structural Alert for genotoxic carcinogenicity
- Structural Alert for nongenotoxic carcinogenicity
- Potential S. typhimurium TA100 mutagen based on QSAR
- Unlikely to be a S. typhimurium TA100 mutagen based on QSAR

Verbose explanation

- QSA24_gen.α,β unsaturated alkoxy No
- QSA25_gen.Aromatic nitroso group No
- QSA26_gen.Aromatic ring N-oxide No
- QSA27_gen.Nitro aromatic Yes
- QSA28_gen.Primary aromatic amine, hydroxyl amine and its derived esters (with restrictions) No
- QSA28bis_gen.Aromatic mono- and dialkylamine No
- QSA28ter_gen.Aromatic N-acyl amine No
- QSA29_gen.Aromatic diazo No
- QSA30_gen.Coumarins and Furocoumarins No
- QSA37_gen.Pyrrolizidine Alkaloids No
- QSA38_gen.Alkenylbenzenes No
- QSA39_gen.and_nogen.Steroidal estrogens No
- QGenotoxic alert?.At least one alert for genotoxic carcinogenicity fired? Yes Class Structural Alert for genotoxic carcinogenicity
- QSA10_gen.α,β unsaturated carbonyls No
- QaN=Na.Aromatic diazo No
- Qar-N=CH2.Derived aromatic amines No

Completed.

Download at
<http://toxtree.sourceforge.net>

11/4/2013

HTTP://TOXTREE.SF.NET/PREDICT

The screenshot shows the ToxTree for Web interface. The search term is "Piroctone olamine". The interface displays various prediction modules and their results:

- Available structure attributes:** CasRN: 68890-66-4, IUPAC name: 1-hydroxy-4-methyl-6-pyridin-2-yl-1,2,3,4-tetrahydropyridine, Names: Piroctone olamine, SMILES: CC=1C=C(C)C(C)C(C)C1=O
- Toxicity prediction modules:**
 - Cramer rules:** High (Class III) (checked)
 - Extended Cramer rules:** (Auto)
 - Verhaar scheme for predicting toxicity mode of action:** (Auto)
 - Verhaar scheme (modified) for predicting toxicity mode of action:** (Auto)
 - Eye irritation:** (Auto)
 - Skin irritation:** (Auto)
 - Structure Alerts for the in vivo micronucleus assay in rodents:** (Auto)
 - Skin sensitisation alerts (M. Cronin):**
 - Alert for SNAr identified: NO
 - Alert for Schiff base formation identified: YES
 - Alert for Michael Acceptor identified: NO
 - Alert for Acyl Transfer agent identified: NO
 - Alert for SN2 identified: NO
 - No skin sensitisation reactivity domains alerts identified: NO
- Structure diagram:** A chemical structure diagram of Piroctone olamine is shown.

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

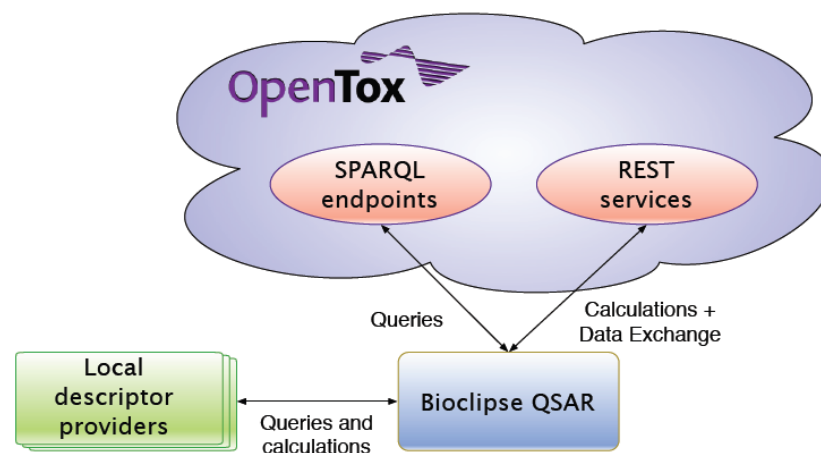
Ambit web services serve as a backend database and calculation

11/4/2013

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., Jeliaskova 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

Research article

Highly accessed

Open Access

Collaborative development of predictive toxicology applications



Journal of
Cheminformatics

11/4/2013

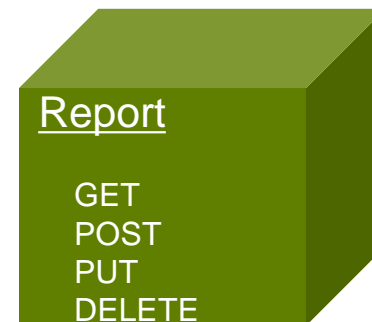
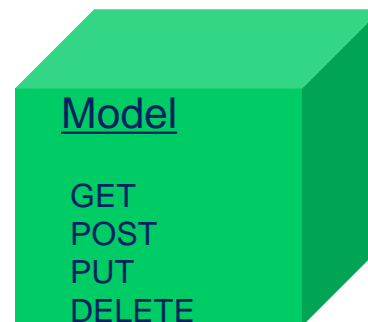
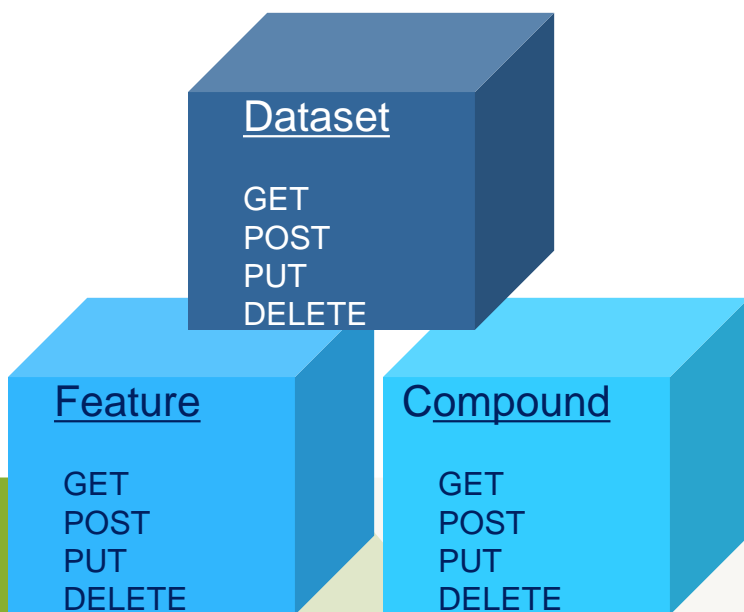
IDEACONSULT LTD.

13

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



11/4/2013

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

ambit
v2.4.10

Structure and properties import

Upload a dataset of chemical structures and properties. Supported formats are SDF, MOL, SMI, CSV, TXT, XLS, ToxML (.xml)

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Register](#)

<ul style="list-style-type: none">HomeStructure searchAll datasetsAdd new structureImport a new datasetImport propertiesPredictBuild modelAlgorithmsModels	<h3>Import new dataset</h3> <p>File ? * <input type="button" value="Choose File"/> tox_benchmark_N6512.sdf</p> <p>Dataset name ? <input type="text" value="Benchmark Data Set for In Silico Prediction of Ame"/></p> <p>URL ? <input type="text" value="http://doc.ml.tu-berlin.de/toxbenchmark/"/></p> <p>Match ? <input type="text" value="Match by InChI"/></p> <p>License ? <input type="text" value="http://www.opendatacommons.org/licenses/pdd"/></p> <p><input type="button" value="Submit"/></p>	<p>Help: Import structures</p> <p>This page allows to upload a file with chemical structures and properties. In order to add a single chemical structure instead, use Add a new structure</p>
---	---	---

HTTP POST to

<http://host:port/ambit2/dataset>

<http://host:port/ambit2/compound>

Uniform interface: (OpenTox web services API)

11/4/2013


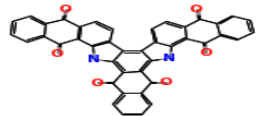

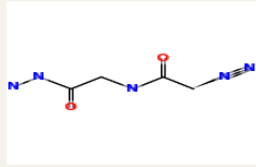

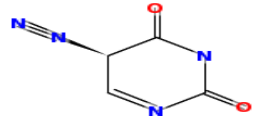

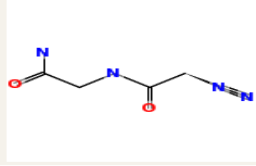

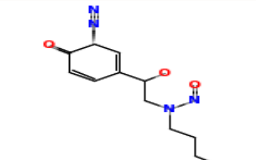
THE DATASET UPLOADED

Summary |

Download as            |

Page 0 Page size 100

Show 10 entries Search:

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

11/4/2013

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

IDEACONSULT LTD.

VIEW DATASET COLUMNS

ambit

Features

Features (identifiers, measured and calculated properties)

v2.5.2-SNAPSHOT

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Register](#)

Showing 60 properties (1 to 10)

First Previous 1 2 3 4 5 Next Last

ID	Name ?	Units	Same As ?	Origin ?	Values Type ?	Nominal values ?	More ?
F113	STRUCTURE_Parent_SMILES		OpenTox: SMILES owl:sameAs	Dataset	<input type="checkbox"/> String	<input type="checkbox"/> No	More
F114	STRUCTURE_InChI		OpenTox: InChI owl:sameAs	Dataset	<input type="checkbox"/> String	<input type="checkbox"/> No	More
F115	ActivityOutcome_CPDBAS_MultiCellCall		Endpoint: Carcinogenicity owl:sameAs	Dataset	<input type="checkbox"/> String	<input type="checkbox"/> No	More
F116	STRUCTURE_Shown		STRUCTURE_Shown owl:sameAs	Dataset	<input type="checkbox"/> String	<input type="checkbox"/> No	More
F117	TestSubstance_ChemicalName		OpenTox: ChemicalName owl:sameAs	Dataset	<input type="checkbox"/> String	<input type="checkbox"/> No	More

Help: Feature service
What is a Feature service ? | API

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

11/4/2013

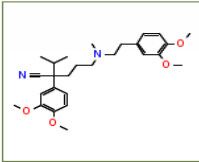
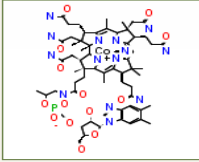
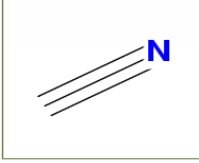
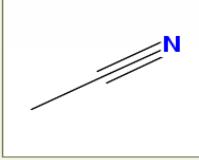
[Home](#)
[Query](#)
 Substructure search
 N#C Max number of hits

 Datasets

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

 D1 ECHA Preregistration list
 Predictions
 M61 130326084909.e0...
 M62 130326085806.d7...
 M63 130326105220.3d...
 Download

Showing 10 structures (1 to 10) First Previous 1 Next Last

Select	CAS	EC	Structure	Name	SMILES	InChI Key
<input checked="" type="checkbox"/>	52-53-9	200-145-1		verapamil		
<input checked="" type="checkbox"/>	68-19-9	200-680-0		cyanocobalamin		
<input checked="" type="checkbox"/>	74-90-8	200-821-6		hydrocyanic acid hydrogen cyanide	C#N	
<input checked="" type="checkbox"/>	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

SEARCH ALL DATASETS

ambit

v2.5.2-SNAPSHOT

Datasets: Chemical structures and properties

Datasets by endpoints

All | [A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [F](#) | [G](#) | [H](#) | [I](#) | [J](#) | [K](#) | [L](#) | [M](#) | [N](#) | [O](#) | [P](#) | [Q](#) | [R](#) | [S](#) | [T](#) | [U](#) | [V](#) | [W](#) | [X](#) | [Y](#) | [Z](#) | [a](#) | [b](#) | [c](#) | [d](#) | [e](#) | [f](#) | [g](#) | [h](#) | [i](#) | [j](#) | [k](#) | [l](#) | [m](#) | [n](#) | [o](#) | [p](#) | [q](#) | [r](#) | [s](#) | [t](#) | [u](#) | [v](#) | [w](#) | [x](#) | [y](#) | [z](#) | [0](#) | [1](#) | [2](#) | [3](#) | [4](#) | [5](#) | [6](#) | [7](#) | [8](#) | [9](#) |

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Register](#)

- Home
- Structure search
- All datasets
- Add new structure
- Import a new dataset
- Import properties
- All substances
- Predict
- Build model
- Algorithms
- Models

Showing 20 datasets (1 to 10)

First Previous 1 2 Next Last

	★	Title	Models	Download	
♥	★ 5	D1 ECHA Preregistration list Source rights Metadata Browse structures and properties Structures only Properties list OpenAM access rights	View Build	...	
♥	★ 5	D2 EPAFHM: EPA Fathead Minnow Acute Toxicity Database Source license Metadata Browse structures and properties Structures only Properties list OpenAM access rights	View Build	...	
♥	★ 5	D4 CPDBAS: Carcinogenic Potency Database Summary Tables - All Species Database Source license Metadata Browse structures and properties Structures only Properties list OpenAM access rights	View Build	...	
♥	★ 5	D6 IRISTR: EPA Integrated Risk Information System (IRIS) Toxicity Review Data Source license Metadata Browse structures and properties Structures only Properties list OpenAM access rights	View Build	...	
♥	★ 5	D8 ISSCAN: CHEMICAL CARCINOGENS Source license Metadata Browse structures and properties Structures only Properties list OpenAM access rights	View Build	...	

HTTP GET

<http://host:port/ambit2/dataset>

11/4/2013

- Home
- All algorithms
- Descriptor Calculation**
- Expert rules
- Toxtree
- Machine learning
 - Regression
 - Classification
 - Clustering
 - FeatureSelection
 - Supervised
 - UnSupervised
 - Single Target
 - Multiple Targets
 - Eager Learning
 - Instance (lazy) Learning
- Applicability domain
- SuperService
- SuperBuilder
- Structure optimisation
- SMSD
- Expert
- Housekeeping
 - Structure Indexing
 - Remote lookup
 - Preferred Structure
- Mockup
- JSON

Showing 66 algorithms (11 to 20)

First Prev: 1 2 3 4 5 Next Last

Name	Endpoint	Description	Type ?	Models ?	Implementation of ?
org.openscience.cdk.qsar.descriptors.molecular.CPSADescriptor CPSA descriptor		Processes a dataset Requires input dataset Requires structure	DescriptorCalculation	View	CPSA
org.openscience.cdk.qsar.descriptors.molecular.AromaticAtomsCountDescriptor Number of aromatic atoms		Processes a dataset Requires input dataset Requires structure	DescriptorCalculation	View	aromaticAtomsCount
org.openscience.cdk.qsar.descriptors.molecular.AromaticBondsCountDescriptor Number of aromatic bonds		Processes a dataset Requires input dataset Requires structure	DescriptorCalculation	View	aromaticBondsCount
org.openscience.cdk.qsar.descriptors.molecular.BondCountDescriptor Number of bonds		Processes a dataset Requires input dataset Requires structure	DescriptorCalculation	View	bondCount
org.openscience.cdk.qsar.descriptors.molecular.AtomCountDescriptor Number of atoms		Processes a dataset Requires input dataset Requires structure	DescriptorCalculation	View	atomCount

11/4/2013

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

CALCULATE DESCRIPTORS

EXAMPLE:BCUT (CDK)

ambit

v2.5.2-SNAPSHOT

Algorithms

Descriptor calculations, model building and data processing algorithms

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Register](#)

Search

Home

All algorithms

Descriptor Calculation

Expert rules

Toxtree

Machine learning

Regression

Classification

Clustering

FeatureSelection

Supervised

UnSupervised

Algorithm at <http://localhost:8080/ambit2/algorithm/org.openscience.cdk.qsar.descriptors.molecular.BCUTDescriptor>

Name	BCUT descriptors
Implementation of	http://www.blueobelisk.org/ontologies/chemoinformatics-algorithms/#BCUT
Requires	Chemical structure
Type	DescriptorCalculation
Action	Processes a dataset

Enter a dataset URI

Run

The result is a dataset,
identified by a [dataset](#)
[URI](#).

?

Help: Algorithm
service
What is Algorithm
service ? | What is
SuperBuilder ? |
Algorithm types ? | [API](#)

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

11/4/2013

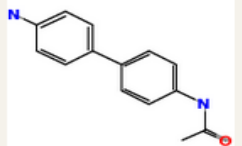
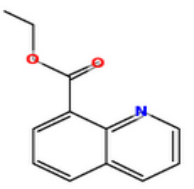
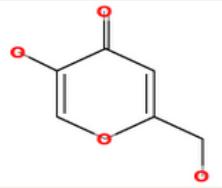
IDEACONSULT LTD.

22

THE TRAINING DATASET (ACTIVITY & DESCRIPTORS)

0 Page size 100

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}>

11/4/2013

IDEACONSULT LTD.

23

- Home
- All algorithms
- Descriptor Calculation
- Expert rules
- Toxtree
- Machine Learning**
- Regression
- Classification
- Clustering
- FeatureSelection
- Supervised
- UnSupervised
- Single Target
- Multiple Targets
- Eager Learning
- Instance (lazy) learning
- Applicability domain

Showing 9 algorithms (1 to 9)

First Previous 1 Next Last

Name	Endpoint ?	Description	Type ?	Models ?	Implementation of ?
LR Regression: Linear regression		<ul style="list-style-type: none"> Builds a model. Requires input dataset Requires property Requires target variable 	Regression SingleTarget EagerLearning Supervised	View	View
IsotonicRegression Isotonic Regression		<ul style="list-style-type: none"> Builds a model. Requires input dataset Requires property Requires target variable 	Regression SingleTarget EagerLearning Supervised	View	View
LMSLR Least median squared linear regression. ISBN-13:978-0471488552		<ul style="list-style-type: none"> Builds a model. Requires input dataset Requires property Requires target variable 	Regression SingleTarget EagerLearning Supervised	View	View
MLP Multilayer Perceptron		<ul style="list-style-type: none"> Builds a model. Requires input dataset Requires property Requires target variable 	Regression Classification SingleTarget EagerLearning Supervised	View	View
PaceRegression Pace Regression http://hdl.handle.net/10289/2131		<ul style="list-style-type: none"> Builds a model. Requires input dataset Requires property Requires target variable 	Regression SingleTarget EagerLearning Supervised	View	View

HTTP GET

<http://host:port/ambit2/algorithm>

BUILD A MODEL



Algorithms

Descriptor calculations, model building and data processing algorithms

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Register](#)

- Home
- All algorithms
- Descriptor calculation
- Expert rules
- Toxtree
- Machine learning
 - Regression
 - Classification
 - Clustering
 - FeatureSelection
 - Supervised
 - UnSupervised
 - Single Target
 - Multiple Targets
 - Eager Learning
 - Instance (lazy)
- Learning
 - Applicability domain
- SuperService

Algorithm at <http://localhost:8080/ambit2/algorithm/RandomForest>

Name	Constructs random forest. doi:10.1023/A:1010933404324
Implementation of Requires	property
Type	Classification SingleTarget EagerLearning Supervised
Action	Builds a model

Once a model is built, it is assigned a [model URI](#) and can be applied to [datasets](#) and [compounds](#).

Enter a dataset URI ?

Enter URI of the target variable ?

- F365: BCUTw-1l
- F366: BCUTw-1h
- F367: BCUTc-1l
- F368: BCUTc-1h
- F369: BCUTp-1l
- F370: BCUTp-1h
- F1593: Activity**

Help: Algorithm service
What is Algorithm service ? | What is SuperBuilder ? | Algorithm types ? | API

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

11/4/2013

THE NEW MODEL



v2.4.10

Models

Regression, classification, clustering, structural alerts, applicability domain, structure optimisation.

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Regist](#)

Search

- Home
- Structure search
- All datasets
- Add new structure
- Import a new dataset
- Import properties
- Predict
- Build model
- Algorithms
- Models
- JSON

Help: Model service
What is Model service ? | [API](#)

Model at <http://localhost:8080/ambit2/model/34>

Name: 130328080910.e11a19b0-23b1-4781-9876-ec921f96027a.weka.classifiers.trees.RandomForest

Training algorithm: <http://localhost:8080/ambit2/algorithm/RandomForest>

Training dataset: <http://localhost:8080/ambit2/dataset/R25534> [Build another model](#)

Independent variables (X): [Browse](#)

Dependent variables (Yobs): [Browse](#)

Predicted (Ypred): [Browse](#)

Representation and Statistics
Action: [View text](#)
Predict properties
Once a model is built, it is assigned a [model URI](#) and can be applied to [datasets](#) and [compounds](#). The result is a dataset, identified by a [dataset URI](#).

Dataset URI: [Use training dataset](#)

Consider using the [Superservice](#) in order to calculate descriptors automatically.

Scatter plots: [Observed / Predicted](#) | [X / Observed](#) | [X / Predicted](#)

Histograms: [X](#) | [Observed](#) | [Predicted](#)

Pie chart: [X](#) | [Observed](#) | [Predicted](#)

The pie chart is titled "Activity" and is divided into two equal halves. The left half is blue and labeled "1", while the right half is red and labeled "0". Below the chart is a legend with a blue circle next to "0" and a red circle next to "1".

11/4/2013

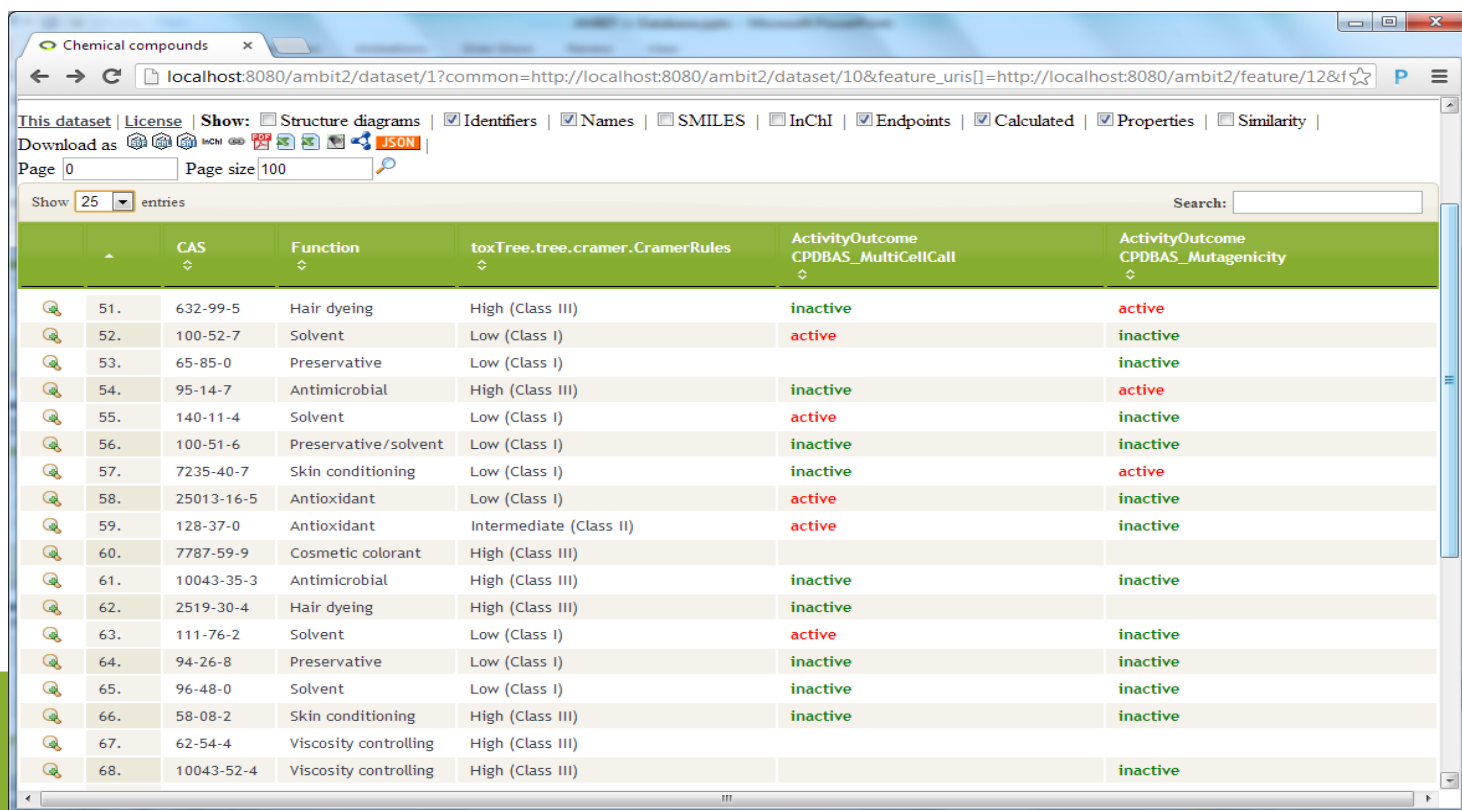
HTTP GET
<http://host:port/ambit2/model/{id}>

IDEACONSULT LTD.

26

PREDICTION RESULTS

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



The screenshot shows a web browser window displaying a table of chemical compounds. The browser address bar shows the URL: `localhost:8080/ambit2/dataset/1?common=http://localhost:8080/ambit2/dataset/10&feature_uris[]=http://localhost:8080/ambit2/feature/12&f`. The page title is "Chemical compounds". The table has the following columns: CAS, Function, toxTree.tree.cramer.CramerRules, ActivityOutcome CPDBAS_MultiCellCall, and ActivityOutcome CPDBAS_Mutagenicity. The table contains 18 rows of data, numbered 51 to 68. The activity outcomes are either "inactive" or "active".

	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	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

11/4/2013

MODELS COMPARISON

Webpage Screenshot



v2.4.10

Models

Regression, classification, clustering, structural alerts, applicability domain, structure optimisation.

[AMBIT @ sourceforge.net](#) | [Help](#) | [Log in](#) | [Register](#)

- Home
- Structure search
- All datasets
- Add new structure
- Import a new dataset
- Import properties
- Predict
- Build model
- Algorithms
- Models
- JSON

Help: Model service
What is Model service ? | API

Showing 7 models (1 to 7)

First Previous 1 Next Last

ID	★	Title	Training Dataset	Algorithm	RMSE (TR)	RMSE (CV)	Correct % (TR)	Correct % (CV)	Stats
M102	★ 5	130325032529.130c34be-9e56-4d26-8bad-8bc856c076a8.weka.classifiers.bayes.BayesNet	R6967	BayesNet	0.54	0.545	66.759	66.375	CV Training
M103	★ 5	130325032803.22db957c-7885-4a79-97d8-7cce761cf8ec.weka.classifiers.bayes.BayesNet	R6967	BayesNet	0.54	0.544	66.759	66.406	CV Training
M104	★ 5	130325034041.a0b16c4a-729f-40db-931b-0c4abf3b214e.weka.classifiers.trees.RandomForest	R6967	RandomForest	0.167	0.391	98.74	78.203	CV Training
M106	★ 5	130325043857.3157ced5-7535-4e63-80f1-21d551673246.weka.classifiers.trees.J48	R6967	J48	0.261	0.454	92.012	76.221	CV Training
M107	★ 5	130325050702.a6ddb0a4-effe-4ca6-9274-c71c9a0174f3.weka.classifiers.functions.SMO	R6967	SMO	0.534	0.536	71.521	71.321	CV Training
M145	★ 5	06f9f5ce-7dd1-4e70-a3f8-27c8907b8094.Fingerprints (Missing fragments)	R6967	fpmissingfragments ambit					
M108	★ 9	e5f16606-907e-4be5-80b8-9ac15fb5e68b.Fingerprints (Tanimoto)	R6967	fptanimoto ambit					

Display
10
models.

Search:

11/4/2013

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)

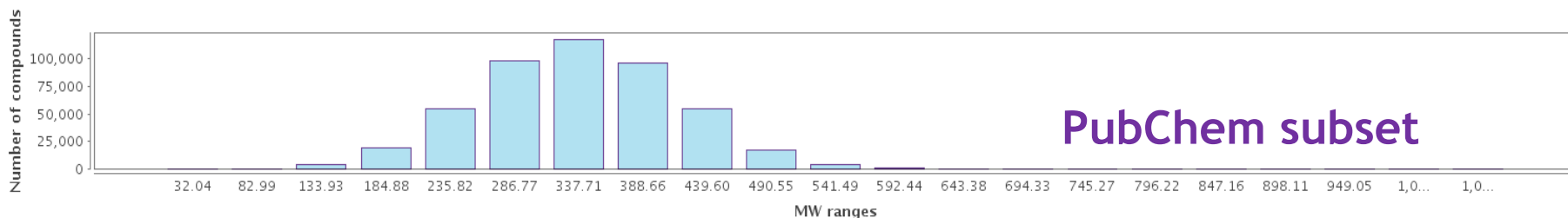
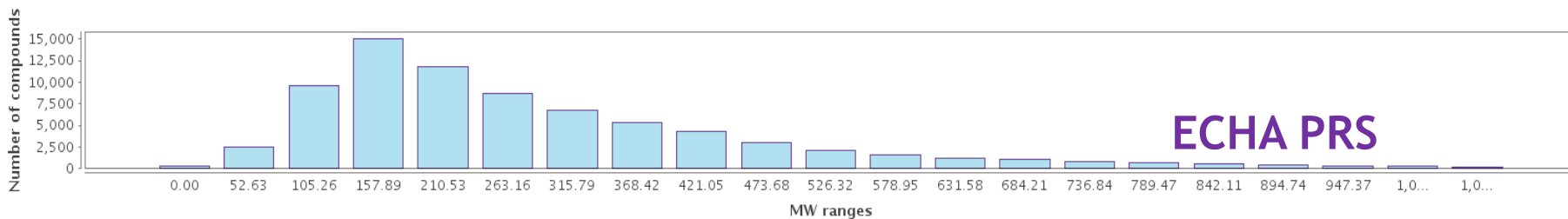
Use thin wrappers for third-party models, and exposing them through the compatible web service API.

- 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.

11/4/2013

DATASET COMPARISON (USING AMBIT ALGORITHMS AND MODELS)



Model id=22

<https://ambit.uni-plovdiv.bg:8443/ambit2/model/22>

OpenTox Model REST API

Model name	Algorithm	Dataset	Independent variables	Dependent	Predicted
MolecularWeight.txt legend	https://ambit.uni-plovdiv.bg:8443/ambit2/algorithm/ambit2.descriptors.MolecularWeightSearch		Independent variables	Dependent	Predicted

Dataset URI:

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)

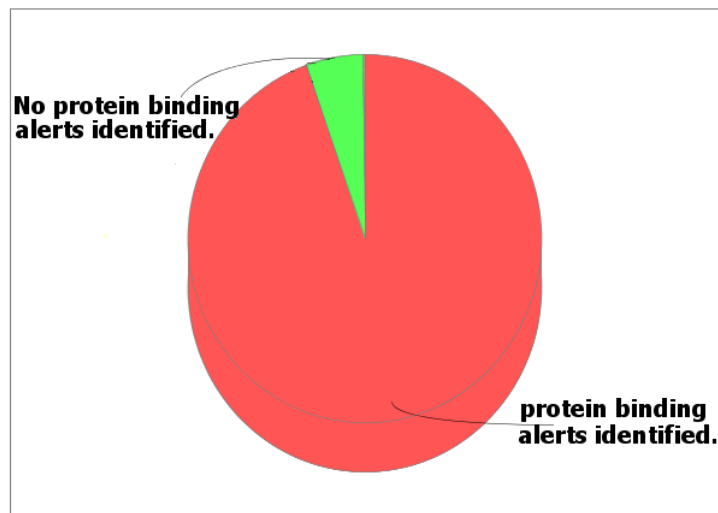
ambit

Dataset <http://localhost:8080/ambit2/dataset/1> Property [No protein binding alerts identified.](#)

Property 28 & Dataset 1

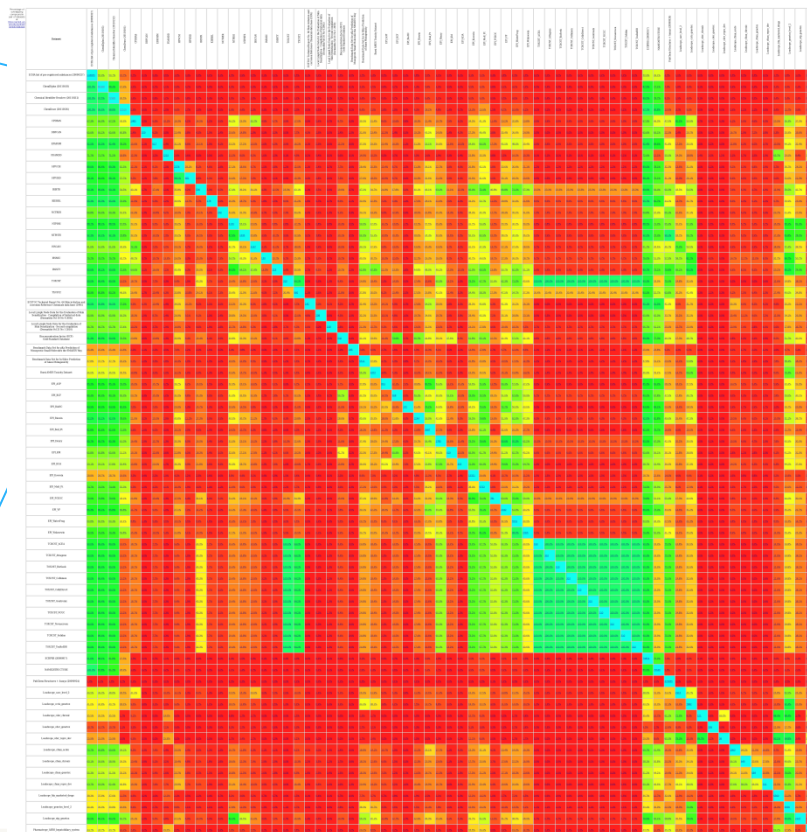
NO

YES



DATASET COMPARISON (MULTIPLE DATASETS)

Dataset	ECHA li	ChemIDp	Chemica	ChemDra	CPDBAS	DBPCAN	EPAFHM	FDAMDD	
ECHA li	143833	55.9%	50.7%	15.2%	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%	72983	30.5%	1.4%	0.2%	0.8%	1.2%	
ChemDra	100.0%	99.6%	98.8%	72519	2.9%	0.6%	2.1%	1.0%	
CPDBAS	67.3%	66.9%	67.2%	43.9%	1485	0.8%	6.5%	10.1%	
DBPCAN	65.6%	65.1%	65.6%	60.8%	2.4%	202	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%	19.9%	19.3%	0.6%	1.5%	1215	
HPVCSI	82.6	Compounds from dataset EPAFHM, found in the dataset ECHA list of pre-registered substances (20090327)						1%	1.2%
HPVISED	88.1	Number = 571 Percent = 92.54%						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%	6.5%	2.2%	
NCTREER	69.8%	69.4%	69.4%	52.2%	22.4%	1.3%	10.8%	8.2%	
NTPBSI	83.7%	88.1%	88.5%	67.6%	32.2%	2.0%	8.6%	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%	9.6%	
ISSMIC	76.7%	76.7%	76.7%	42.7%	48.7%	0.7%	10.7%	13.3%	
ISSSY	90.6%	90.1%	90.6%	72.6%	64.6%	2.1%	16.6%	7.2%	
TOXCST1	90.0%	89.4%	89.0%	13.2%	18.7%	0.0%	7.7%	0.3%	
TOXCST2	80.9%	80.8%	80.7%	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%	
LocalL	65.9%	65.9%	65.9%	56.3%	18.3%	1.9%	8.7%	2.4%	
LocalL	66.7%	66.7%	65.7%	57.4%	10.2%	0.0%	3.7%	1.0%	
Bioconc	91.3%	89.9%	90.0%	70.3%	23.8%	0.6%	13.8%	1.3%	
Benchma	25.4%	25.4%	25.4%	21.6%	2.2%	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%	4.7%	2.9%	
EPI_AOP	84.3%	85.0%	85.1%	79.3%	16.6%	1.0%	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%	



Legend:

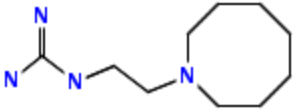
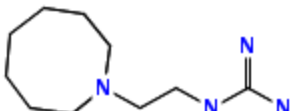
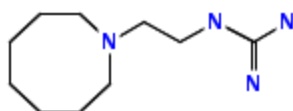
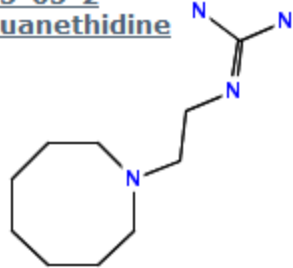
No overlap



100% overlap

11/4/2013

BUILT-IN HEURISTICS TO DISCOVER 2D CHEMICAL STRUCTURE INCONSISTENCIES

<p><u>55-65-2</u> <u>guanethidine</u></p>  <p><u>Consensus Quality</u> <u>Label</u> Majority[1:3] <u>Structure Quality</u> <u>Label</u> ProbablyOK</p>	<p><u>55-65-2</u> <u>guanethidine</u></p>  <p><u>Consensus Quality</u> <u>Label</u> Majority[1:3] <u>Structure Quality</u> <u>Label</u> ProbablyOK</p>	<p><u>55-65-2</u> <u>guanethidine</u></p>  <p><u>Consensus Quality</u> <u>Label</u> Majority[1:3] <u>Structure Quality</u> <u>Label</u> ProbablyOK</p>	<p><u>55-65-2</u> <u>guanethidine</u></p>  <p><u>Consensus Quality</u> <u>Label</u> Majority[1:3] <u>Structure Quality</u> <u>Label</u> ProbablyERROR</p>
---	---	---	--

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 Ideaconult

Collaborators at Plovdiv University

OpenTox partners

The Open Source community

- CDK and Bioclipse in particular

Funding:

- CEFIC-LRI, FP7 OpenTox, FP7
Cadaster

11/4/2013