HelmholtzZentrum münchen
German Research Center for Environmental Health

eADMET
THE REFERENCE IN CHEMOINFORMATICS

Web Tools for Predictive Toxicology Model Building

Igor V. Tetko
Helmholtz Zentrum Muenchen & eADMET GmbH

OpenTox, Munich, August 11, 2011

Web tools – why?

The widest possible dissemination of information

Instant sharing of data and models

Data integration from different partners

Private/shared/public access to data and models

Access to original sources of data

Collaboration to develop common models

Sharing of developed models

Reuse of knowledge and best modeling practices

Highlighting web tools & resources

CDD LtD (Spin off Eli Lilly)	2004 collaborativedrug.com
VCCLAB (INTAS project)	2005 vcclab.org
AMBIT (IDEA Ltd)	2005 ambit.sourceforge.net
OCHEM (GO-Bio project, eADMET GmbH)	2011 ochem.eu
OpenTox (FP7 project)	2010 opentox.org
CADASTER QSPR-THESAURUS (FP7 project)	2009 cadaster.eu
JRC QMRF Database (JRC, EU)	2008 qsardb.jrc.it
ChemBench (UNC, NIH & EPA)	2010 chembench.mml.unc.edu

Grouping by functionality

Database sharing

CDD LtD (Spin off Eli Lilly)

Descriptor calculation & modelling

VCCLAB (INTAS project)

Collection of descriptions and models

JRC QMRF Database (JRC, EU)

CADASTER QSPR-THESAURUS (FP7 project)

Workflow and API for model development and publishing

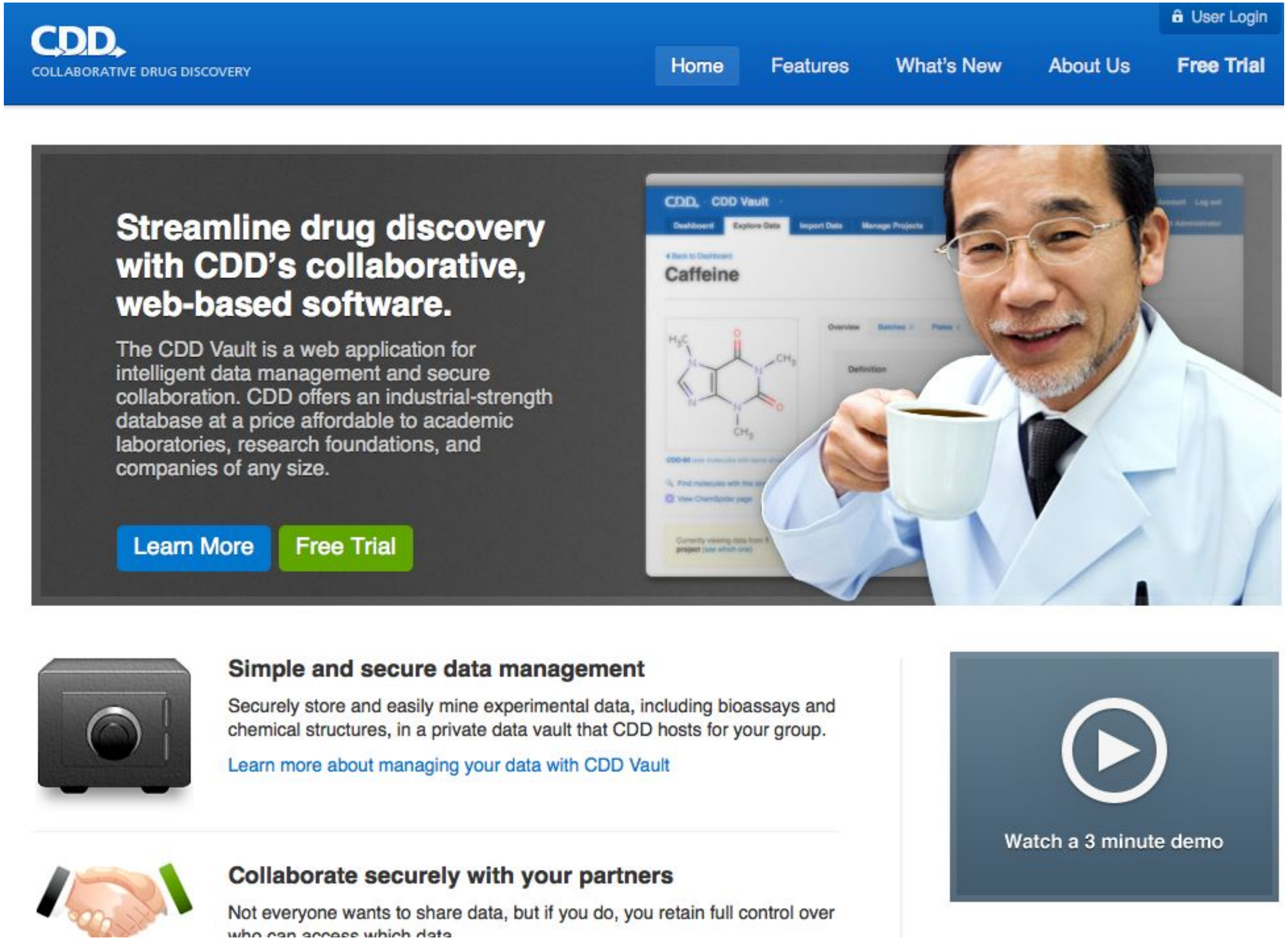
ChemBench (UNC, NIH & EPA)

OpenTox (FP7 project)

Database and workflow for model development and publishing

OCHEM (GO-Bio project, eADMET GmbH)

DATA sharing: Collaborative Drug Discovery (CDD)



Streamline drug discovery with CDD's collaborative, web-based software.

The CDD Vault is a web application for intelligent data management and secure collaboration. CDD offers an industrial-strength database at a price affordable to academic laboratories, research foundations, and companies of any size.

Simple and secure data management

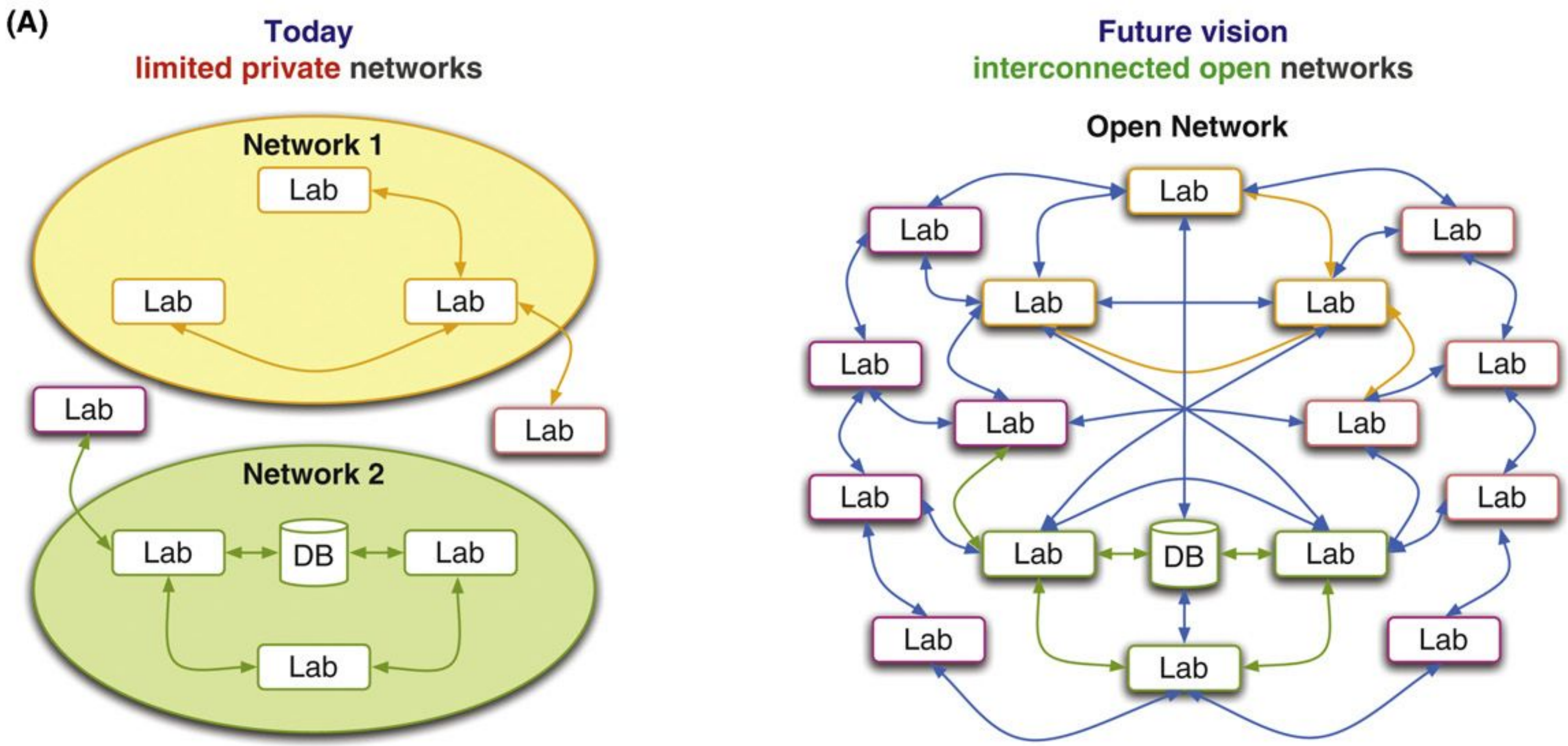
Securely store and easily mine experimental data, including bioassays and chemical structures, in a private data vault that CDD hosts for your group.

Learn more about managing your data with CDD Vault

Collaborate securely with your partners

Not everyone wants to share data, but if you do, you retain full control over who can access which data

Concept: social network for drug discovery



- Fragmented/duplicative efforts
- Difficulty spanning disciplines
- IP issues inhibit open sharing
- Lack of project management

- Interconnect the whole community
- Cross-pollinate groups
- Preserve IP/streamline agreements
- Advance drug candidates faster

Collaborative Drug Discovery (CDD) -- an importance of data sharing

2011: CDD wins Bio-IT World Editors' Choice Best Practices Award

2011: MM4TB (More Medicines for Tuberculosis) 5 year EU funded project with AstraZeneca, Sanofi-Aventis

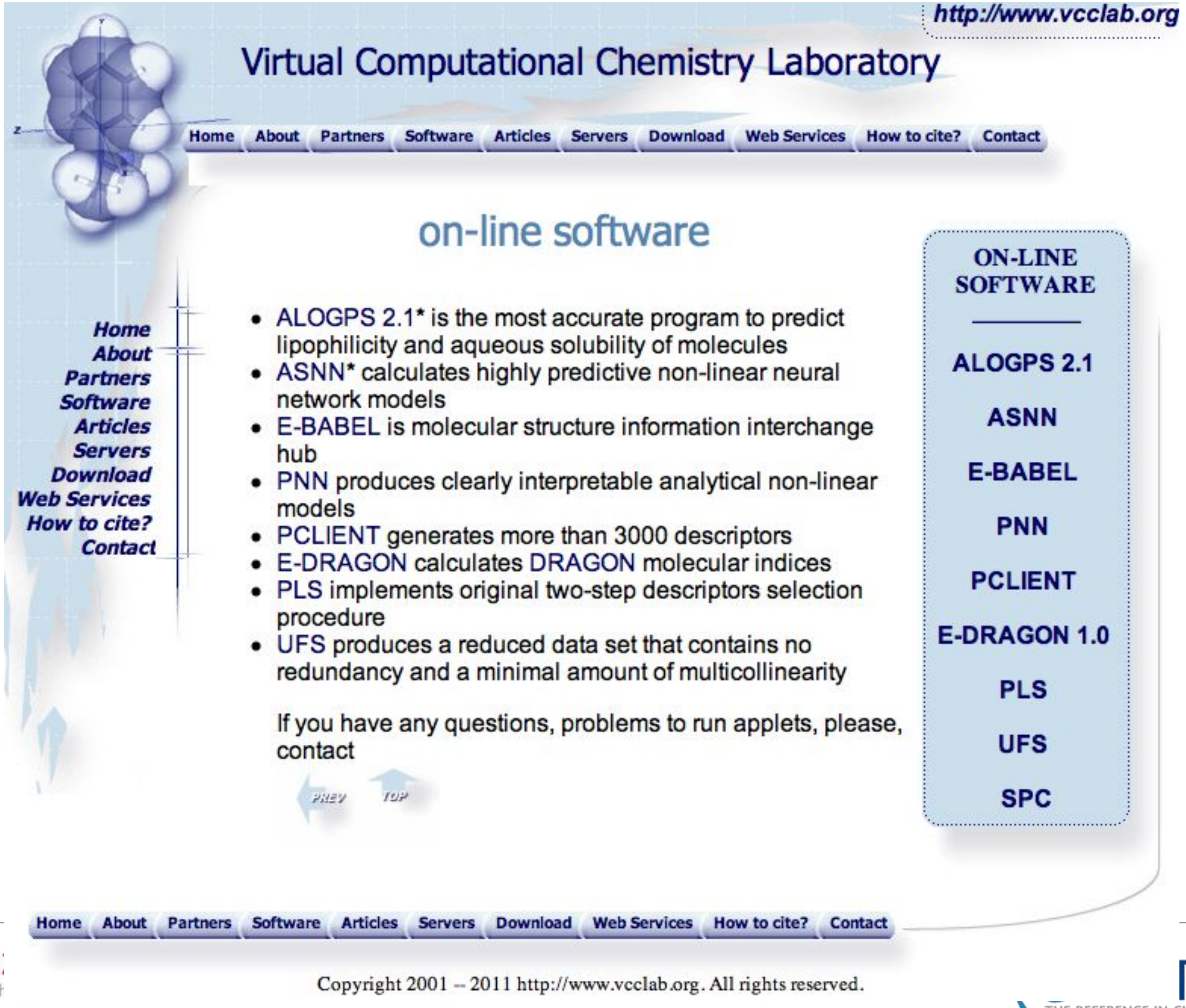
2010: GSK, Novartis, Pfizer, and NIH Collaborations Announced

2008: Gates Foundation 2 year grant for TB database (extended to 5 years)

2005: Eli Lilly co-invested in a syndicate with Omidyar Network and Founders Fund

2004: CDD spun out of Lilly, UCSF signs up as first customer

Descriptor calculations & models development

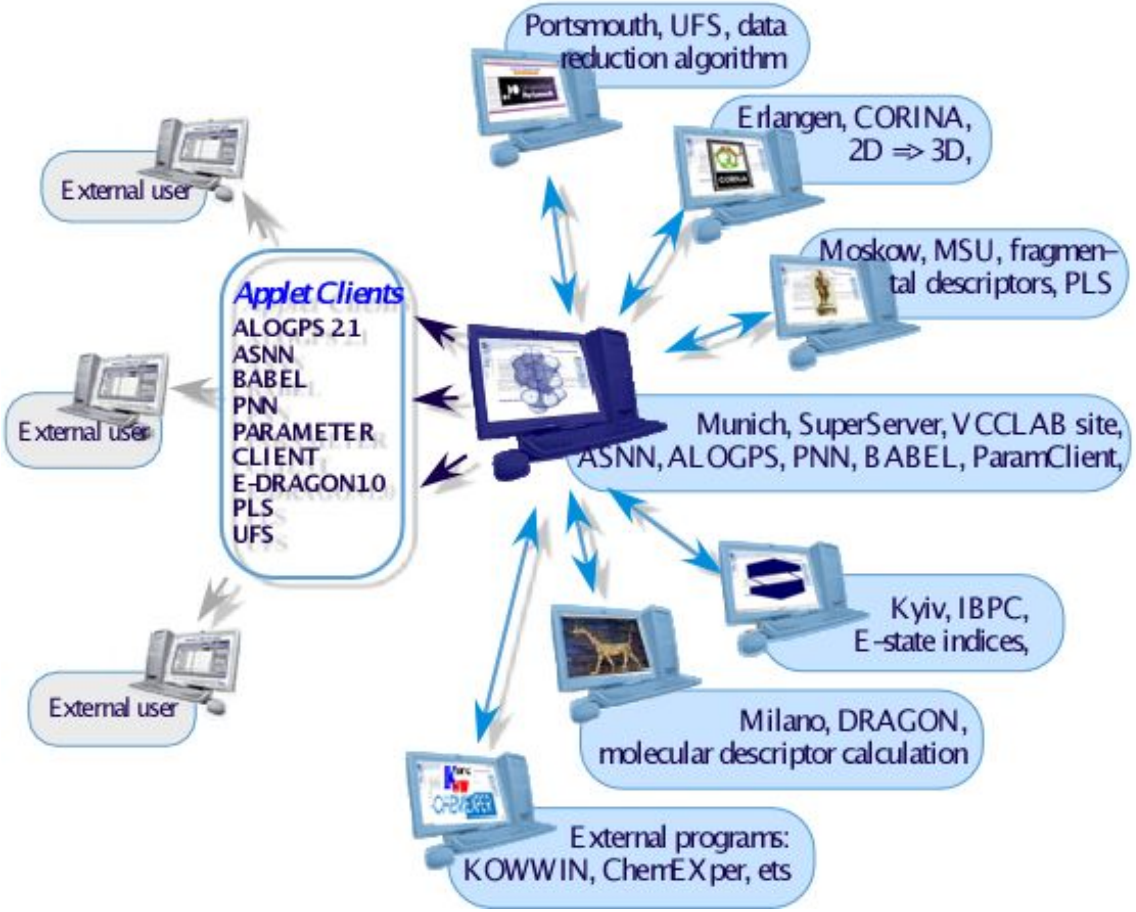

http://www.vcclab.org
Virtual Computational Chemistry Laboratory
Home About Partners Software Articles Servers Download Web Services How to cite? Contact
Home
About
Partners
Software
Articles
Servers
Download
Web Services
How to cite?
Contact
ON-LINE SOFTWARE
ALOGPS 2.1
ASNN
E-BABEL
PNN
PCLIENT
E-DRAGON 1.0
PLS
UFS
SPC
PREV
TOP
Home About Partners Software Articles Servers Download Web Services How to cite? Contact

on-line software

- ALOGPS 2.1* is the most accurate program to predict lipophilicity and aqueous solubility of molecules
- ASNN* calculates highly predictive non-linear neural network models
- E-BABEL is molecular structure information interchange hub
- PNN produces clearly interpretable analytical non-linear models
- PCLIENT generates more than 3000 descriptors
- E-DRAGON calculates DRAGON molecular indices
- PLS implements original two-step descriptors selection procedure
- UFS produces a reduced data set that contains no redundancy and a minimal amount of multicollinearity

If you have any questions, problems to run applets, please, contact

Collaboration in VCCLAB


Portsmouth, UFS, data reduction algorithm
Erlangen, CORINA, 2D => 3D,
Moskow, MSU, fragmen-tal descriptors, PLS
Munich, SuperServer, VCCLAB site, ASNN, ALOGPS, PNN, BABEL, ParamClient,
Kyiv, IBPC, E-state indices,
Milano, DRAGON, molecular descriptor calculation
External programs: KOWWIN, ChemEXper, ets
Applet Clients
ALOGPS 2.1
ASNN
BABEL
PNN
PARAMETER CLIENT
E-DRAGON1.0
PLS
UFS
External user
External user
External user

VCCLAB offers:

Descriptor/property calculations

PCLIENT: Descriptor calculation using several algorithms

eDRAGON: Web interface to Dragon (Talete Srl) descriptor calculation tool using user provided structures or CORINA (Molecular Networks GmbH)

ALOGPS: calculation of logP and logS using ALOGPS as well as >10 other methods

On-line model development

ASNN – Associative Neural Networks

PNN – Polynomial Neural Network

PLS – Partial Least Squares

Other tools

SPC – Supermagnetic Clustering

E-babel – on-line conversion of molecules using OpenBabel

UFS – Unsupervised Forward Selection of descriptors

Virtual Computational Chemistry Laboratory statistics

VCCLAB: 2001-2004 INTAS project “Virtual Computational Chemistry Laboratory”

- More than 5000 unique users per month
- About 4500 registered users
- > 230,000 tasks are calculated per year
- ~ 300 citations of the primary article (~ 600 citations in Google Scholar)

But …

- Across labs EU collaboration – was great to develop but difficult to support

Model storage: JRC, CADASTER

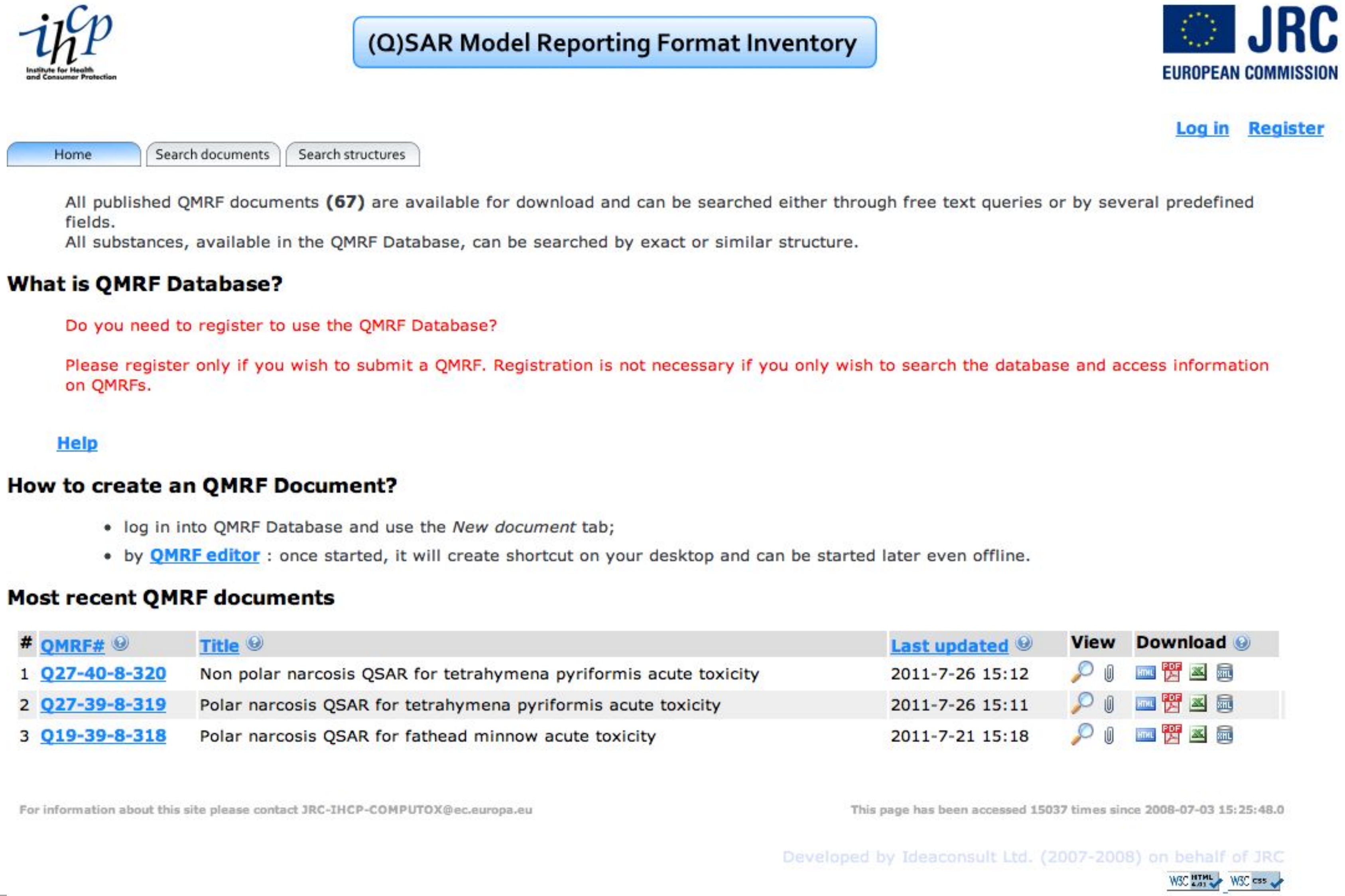
REACH guidelines (OECD principles):

1) a defined endpoint
2) an unambiguous algorithm
3) defined domain of applicability
4) appropriate measures of goodness-of-fit, robustness and predictivity
5) a mechanistic interpretation, if possible

→ Published models may not be sufficient to fulfill all principles

→ Am authority decision is required to decide whether models fulfills “OECD principles”; once validated such “OECD” models should be stored and be available for use in REACH registration

(Q)SAR Model Reporting Format (QMRF) Inventory



Institute for Health and Consumer Protection

(Q)SAR Model Reporting Format Inventory

JRC
EUROPEAN COMMISSION

Log in Register

Home | Search documents | Search structures

All published QMRF documents **(67)** are available for download and can be searched either through free text queries or by several predefined fields.
All substances, available in the QMRF Database, can be searched by exact or similar structure.

What is QMRF Database?

Do you need to register to use the QMRF Database?

Please register only if you wish to submit a QMRF. Registration is not necessary if you only wish to search the database and access information on QMRFs.

Help

How to create an QMRF Document?

- log in into QMRF Database and use the *New document* tab;
- by QMRF editor : once started, it will create shortcut on your desktop and can be started later even offline.

Most recent QMRF documents

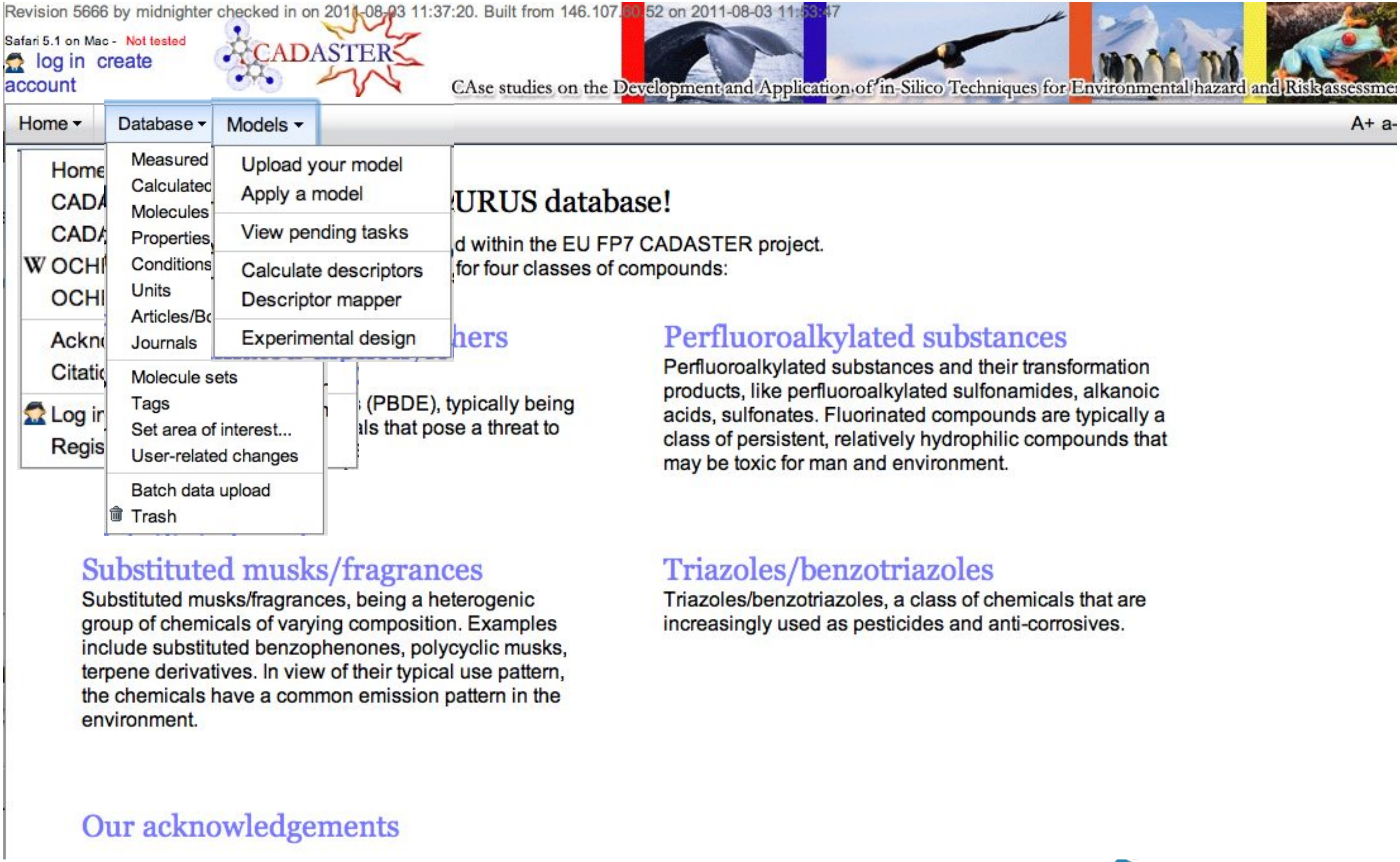
#	QMRF#	Title	Last updated	View	Download
1	Q27-40-8-320	Non polar narcosis QSAR for tetrahymena pyriformis acute toxicity	2011-7-26 15:12	🔍 📎	HTML PDF X XML
2	Q27-39-8-319	Polar narcosis QSAR for tetrahymena pyriformis acute toxicity	2011-7-26 15:11	🔍 📎	HTML PDF X XML
3	Q19-39-8-318	Polar narcosis QSAR for fathead minnow acute toxicity	2011-7-21 15:18	🔍 📎	HTML PDF X XML

For information about this site please contact JRC-IHCP-COMPUTOX@ec.europa.eu

This page has been accessed 15037 times since 2008-07-03 15:25:48.0

Developed by Ideaconsult Ltd. (2007-2008) on behalf of JRC

CADASTER QSPR-THESAURUS DB



Revision 5666 by midnighter checked in on 2011-08-03 11:37:20. Built from 146.107.60.52 on 2011-08-03 11:53:47

Safari 5.1 on Mac - Not tested

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessme

URUS database!

d within the EU FP7 CADASTER project.

,for four classes of compounds:

hers

(PBDE), typically being

als that pose a threat to

Perfluoroalkylated substances

Perfluoroalkylated substances and their transformation products, like perfluoroalkylated sulfonamides, alkanoic acids, sulfonates. Fluorinated compounds are typically a class of persistent, relatively hydrophilic compounds that may be toxic for man and environment.

Substituted musks/fragrances

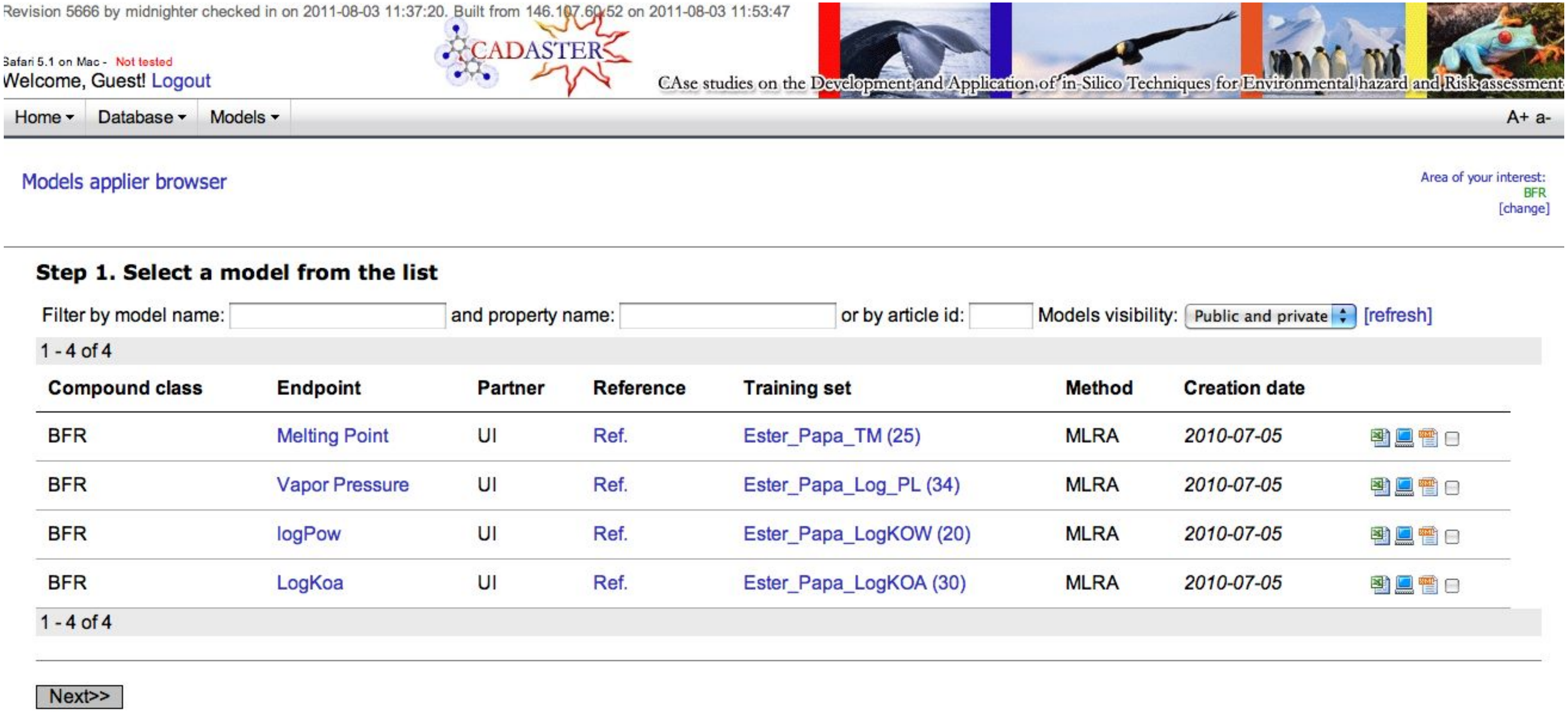
Substituted musks/fragrances, being a heterogenic group of chemicals of varying composition. Examples include substituted benzophenones, polycyclic musks, terpene derivatives. In view of their typical use pattern, the chemicals have a common emission pattern in the environment.

Triazoles/benzotriazoles

Triazoles/benzotriazoles, a class of chemicals that are increasingly used as pesticides and anti-corrosives.

Our acknowledgements

QSPR THESAURUS: available models



Revision 5666 by midnighter checked in on 2011-08-03 11:37:20. Built from 146.107.60.52 on 2011-08-03 11:53:47

Safari 5.1 on Mac - Not tested

Welcome, Guest! Logout

CADASTER
CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment
Home ▾
Database ▾
Models ▾
A+ a-

Models applier browser

Area of your interest:
BFR
[change]

Step 1. Select a model from the list

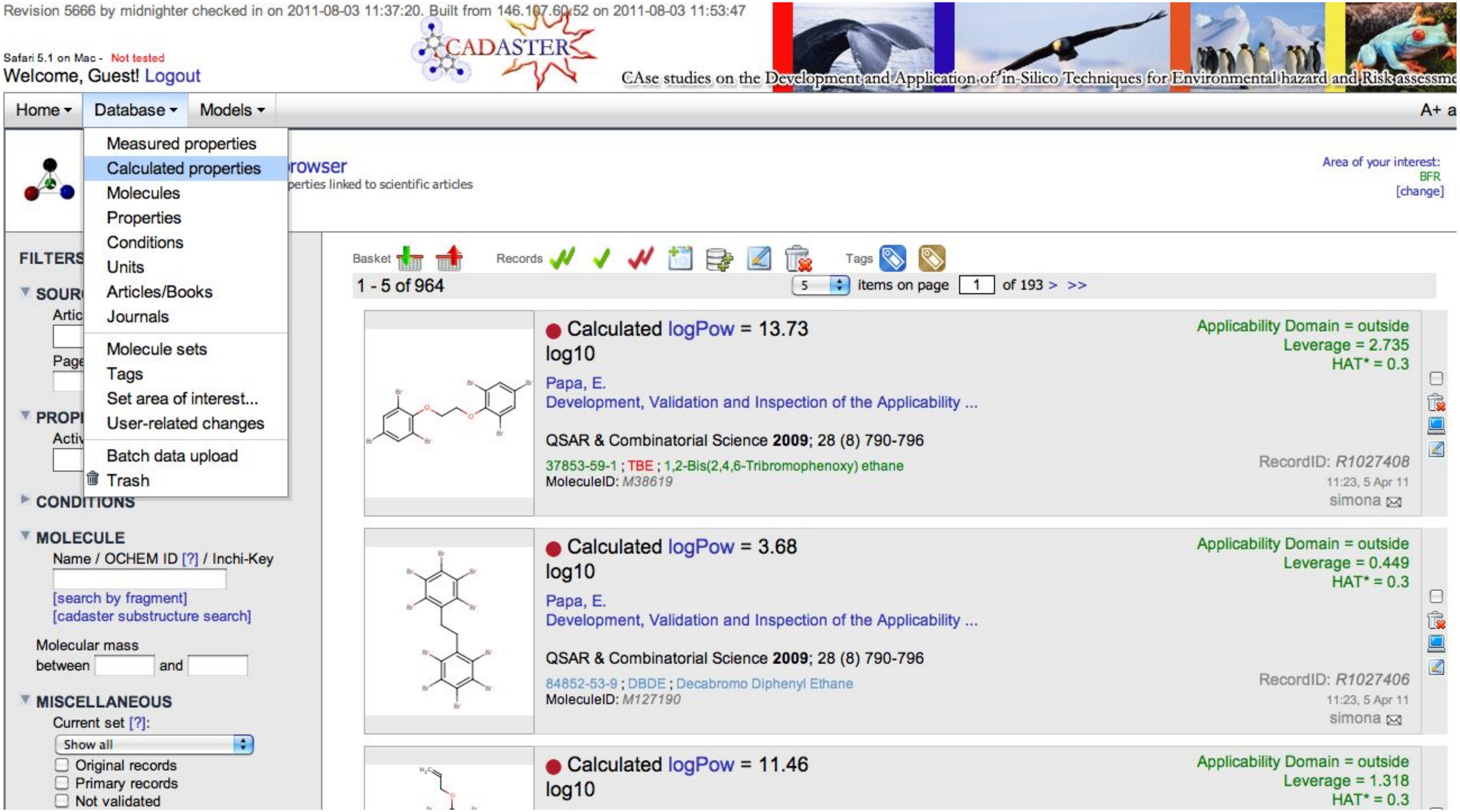
Filter by model name: and property name: or by article id: Models visibility: Public and private [refresh]

1 - 4 of 4

Compound class	Endpoint	Partner	Reference	Training set	Method	Creation date	
BFR	Melting Point	UI	Ref.	Ester_Papa_TM (25)	MLRA	*2010-07-05*	□
BFR	Vapor Pressure	UI	Ref.	Ester_Papa_Log_PL (34)	MLRA	*2010-07-05*	□
BFR	logPow	UI	Ref.	Ester_Papa_LogKOW (20)	MLRA	*2010-07-05*	□
BFR	LogKoa	UI	Ref.	Ester_Papa_LogKOA (30)	MLRA	*2010-07-05*	□

1 - 4 of 4

QSPR-THESAURUS: Browser of calculated values



Revision 5666 by midnighter checked in on 2011-08-03 11:37:20. Built from 146.107.60/52 on 2011-08-03 11:53:47

Safari 5.1 on Mac - Not tested

Welcome, Guest! Logout

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessme

Home ▾ Database ▾ Models ▾ A+ a

- Measured properties
- Calculated properties
- Molecules
- Properties
- Conditions
- Units
- Articles/Books
- Journals
- Molecule sets
- Tags
- Set area of interest...
- User-related changes
- Batch data upload
- Trash

rowser

perties linked to scientific articles

Area of your interest:
BFR
[change]

FILTERS

SOUR
Artic
Page

PROPI
Activ

CONDITIONS

MOLECULE
Name / OCHEM ID [?] / Inchi-Key
[search by fragment]
[cadaster substructure search]

Molecular mass
between and

MISCELLANEOUS
Current set [?]:
Show all
□ Original records
□ Primary records
□ Not validated

Basket Records Tags

1 - 5 of 964 5 items on page 1 of 193 > >>

● Calculated logPow = 13.73
log10
Papa, E.
Development, Validation and Inspection of the Applicability ...
QSAR & Combinatorial Science **2009**; 28 (8) 790-796
37853-59-1 ; TBE ; 1,2-Bis(2,4,6-Tribromophenoxy) ethane
MoleculeID: *M38619*

Applicability Domain = outside
Leverage = 2.735
HAT* = 0.3
RecordID: *R1027408*
11:23, 5 Apr 11
simona

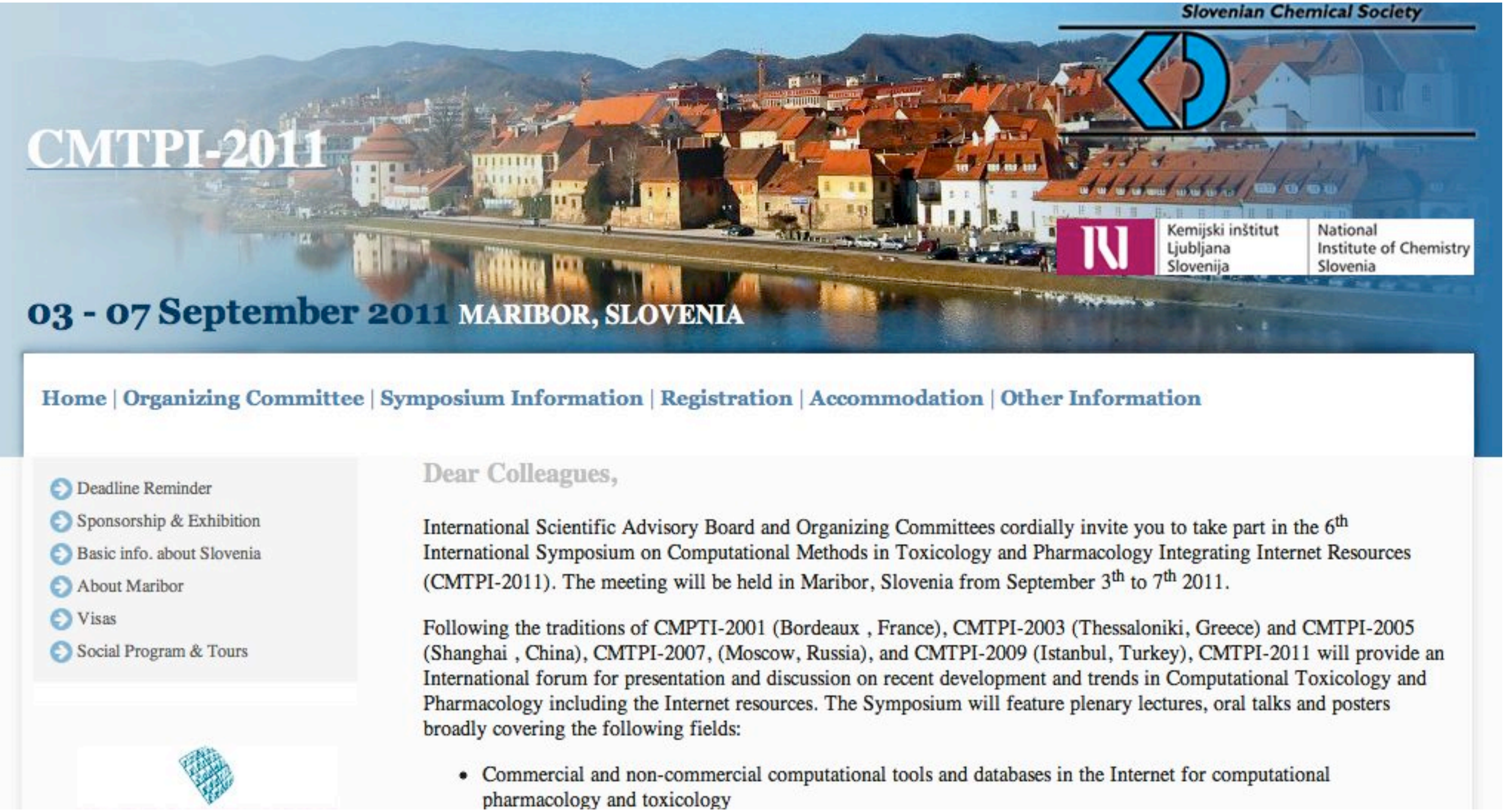
● Calculated logPow = 3.68
log10
Papa, E.
Development, Validation and Inspection of the Applicability ...
QSAR & Combinatorial Science **2009**; 28 (8) 790-796
84852-53-9 ; DBDE ; Decabromo Diphenyl Ethane
MoleculeID: *M127190*

Applicability Domain = outside
Leverage = 0.449
HAT* = 0.3
RecordID: *R1027406*
11:23, 5 Apr 11
simona

● Calculated logPow = 11.46
log10

Applicability Domain = outside
Leverage = 1.318
HAT* = 0.3

CADASTER workshop, Maribor, September 1-2, 2011


CMTPI-2011
Slovenian Chemical Society
Kemijski inštitut Ljubljana Slovenija
National Institute of Chemistry Slovenia
03 - 07 September 2011 MARIBOR, SLOVENIA
Home | Organizing Committee | Symposium Information | Registration | Accommodation | Other Information
Deadline Reminder
Sponsorship & Exhibition
Basic info. about Slovenia
About Maribor
Visas
Social Program & Tours

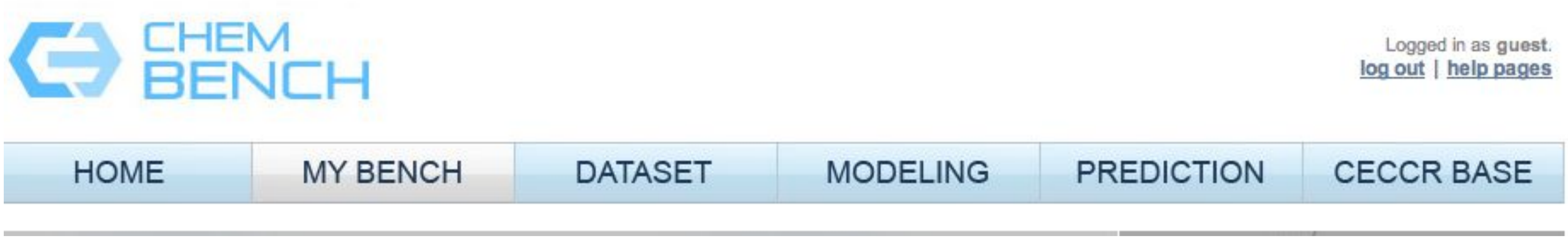
Dear Colleagues,

International Scientific Advisory Board and Organizing Committees cordially invite you to take part in the 6th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011). The meeting will be held in Maribor, Slovenia from September 3th to 7th 2011.

Following the traditions of CMPTI-2001 (Bordeaux , France), CMTPI-2003 (Thessaloniki, Greece) and CMTPI-2005 (Shanghai , China), CMTPI-2007, (Moscow, Russia), and CMTPI-2009 (Istanbul, Turkey), CMTPI-2011 will provide an International forum for presentation and discussion on recent development and trends in Computational Toxicology and Pharmacology including the Internet resources. The Symposium will feature plenary lectures, oral talks and posters broadly covering the following fields:

- Commercial and non-commercial computational tools and databases in the Internet for computational pharmacology and toxicology

Model Development & Publishing: ChemBench



Logged in as guest.

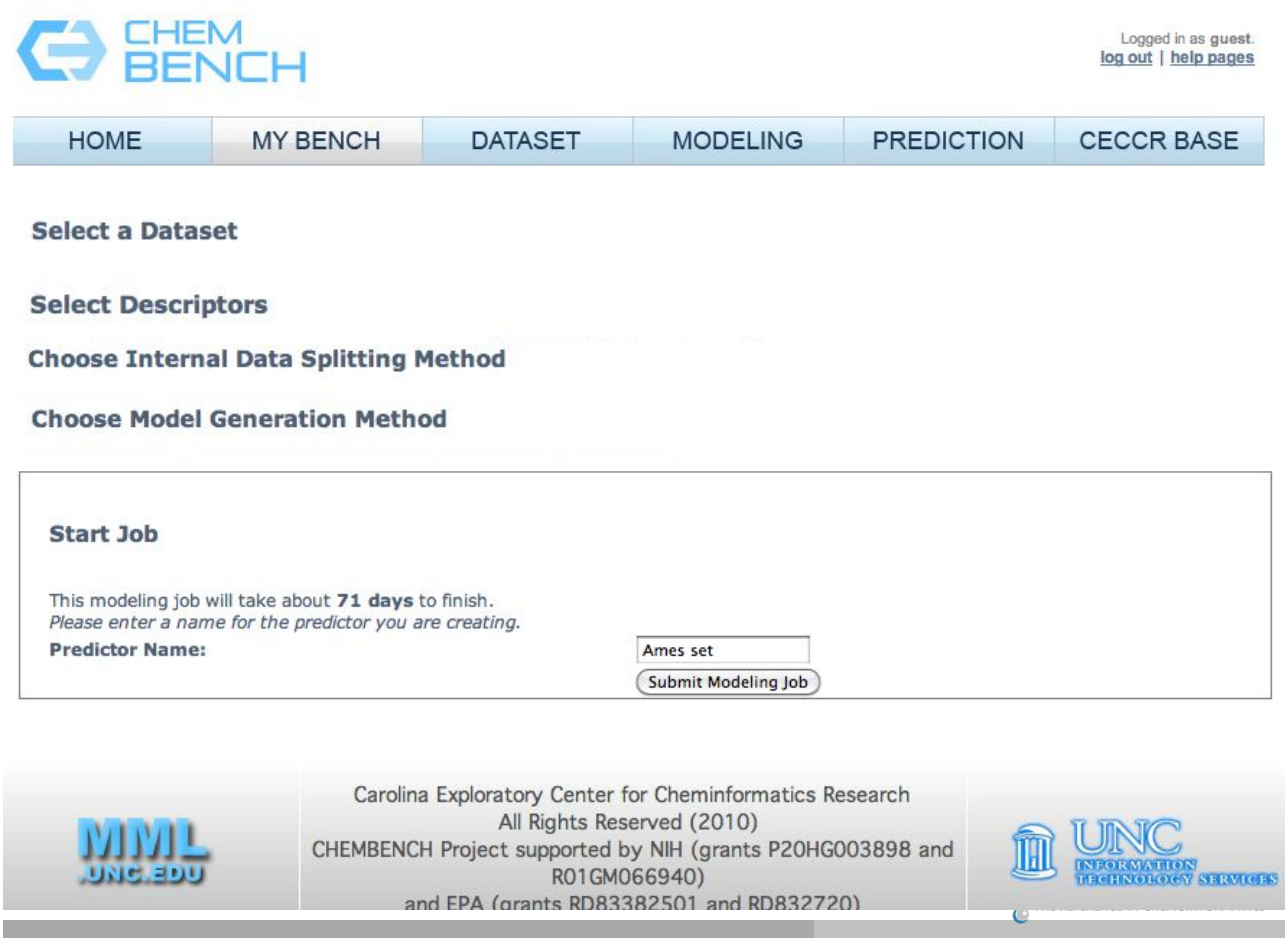
Developed with help of NIH and EPA projects

Main motivation: enable processing of large HTS datasets, like those produced by PubChem (2600 bioassays with nearly 300 000 active molecules)

Allows: Dataset Creation, Visualization, Modeling, Model Validation, Virtual Screening (predictions)

Includes:

- Random Forest
- SVM
- GA-kNN
- SA-kNN



Select a Dataset

Select Descriptors

Choose Internal Data Splitting Method

Choose Model Generation Method

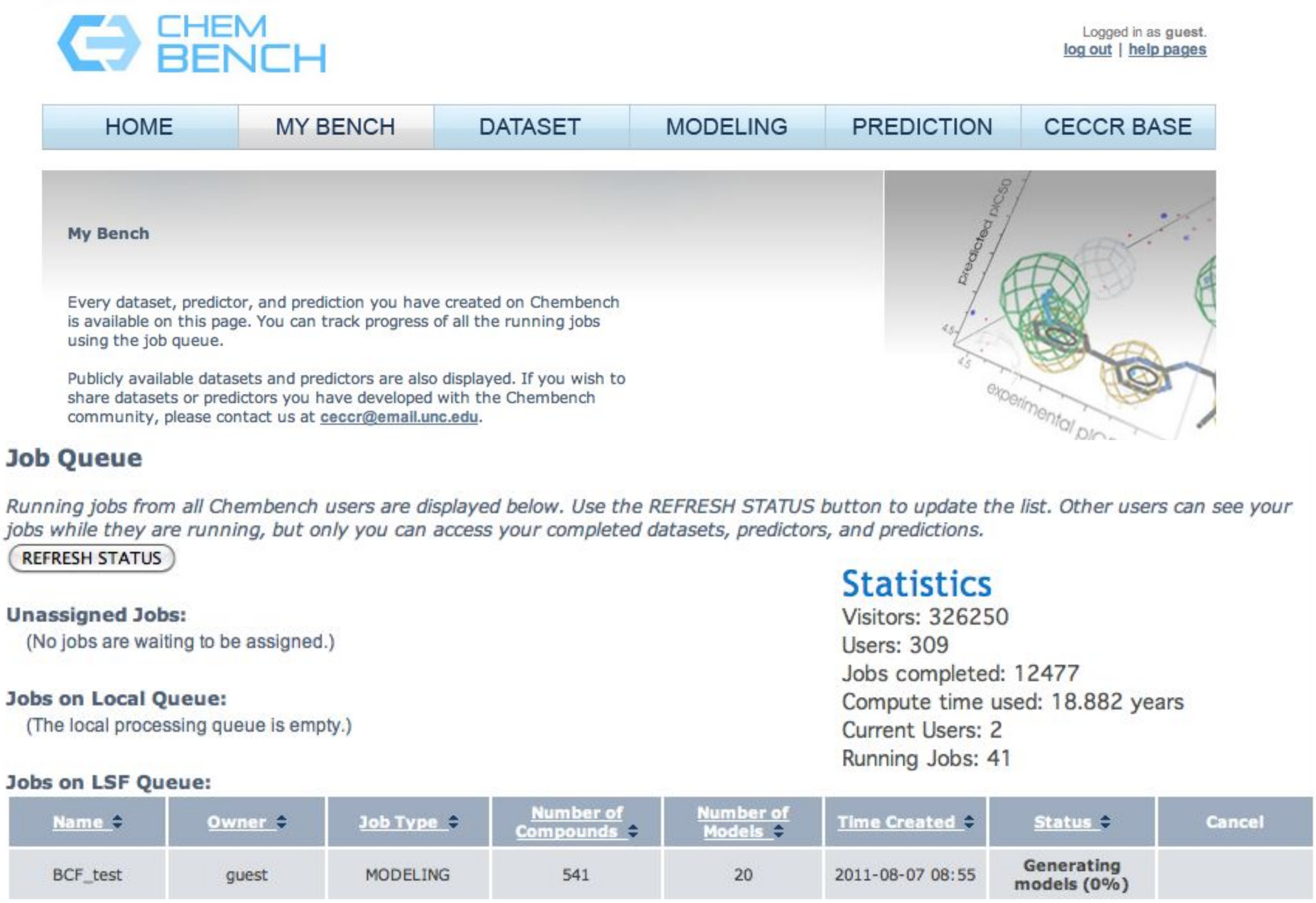
Start Job

This modeling job will take about **71 days** to finish.
Please enter a name for the predictor you are creating.

Predictor Name: Ames set

Submit Modeling Job

Carolina Exploratory Center for Cheminformatics Research
All Rights Reserved (2010)
CHEMBENCH Project supported by NIH (grants P20HG003898 and R01GM066940)
and EPA (grants RD83382501 and RD832720)



My Bench

Every dataset, predictor, and prediction you have created on Chembench is available on this page. You can track progress of all the running jobs using the job queue.

Publicly available datasets and predictors are also displayed. If you wish to share datasets or predictors you have developed with the Chembench community, please contact us at ceccr@email.unc.edu.

Job Queue

Running jobs from all Chembench users are displayed below. Use the REFRESH STATUS button to update the list. Other users can see your jobs while they are running, but only you can access your completed datasets, predictors, and predictions.

Unassigned Jobs:

(No jobs are waiting to be assigned.)

Jobs on Local Queue:

(The local processing queue is empty.)

Jobs on LSF Queue:

Name	Owner	Job Type	Number of Compounds	Number of Models	Time Created	Status	Cancel
BCF_test	guest	MODELING	541	20	2011-08-07 08:55	**Generating models (0%)**	

Statistics

Visitors: 326250

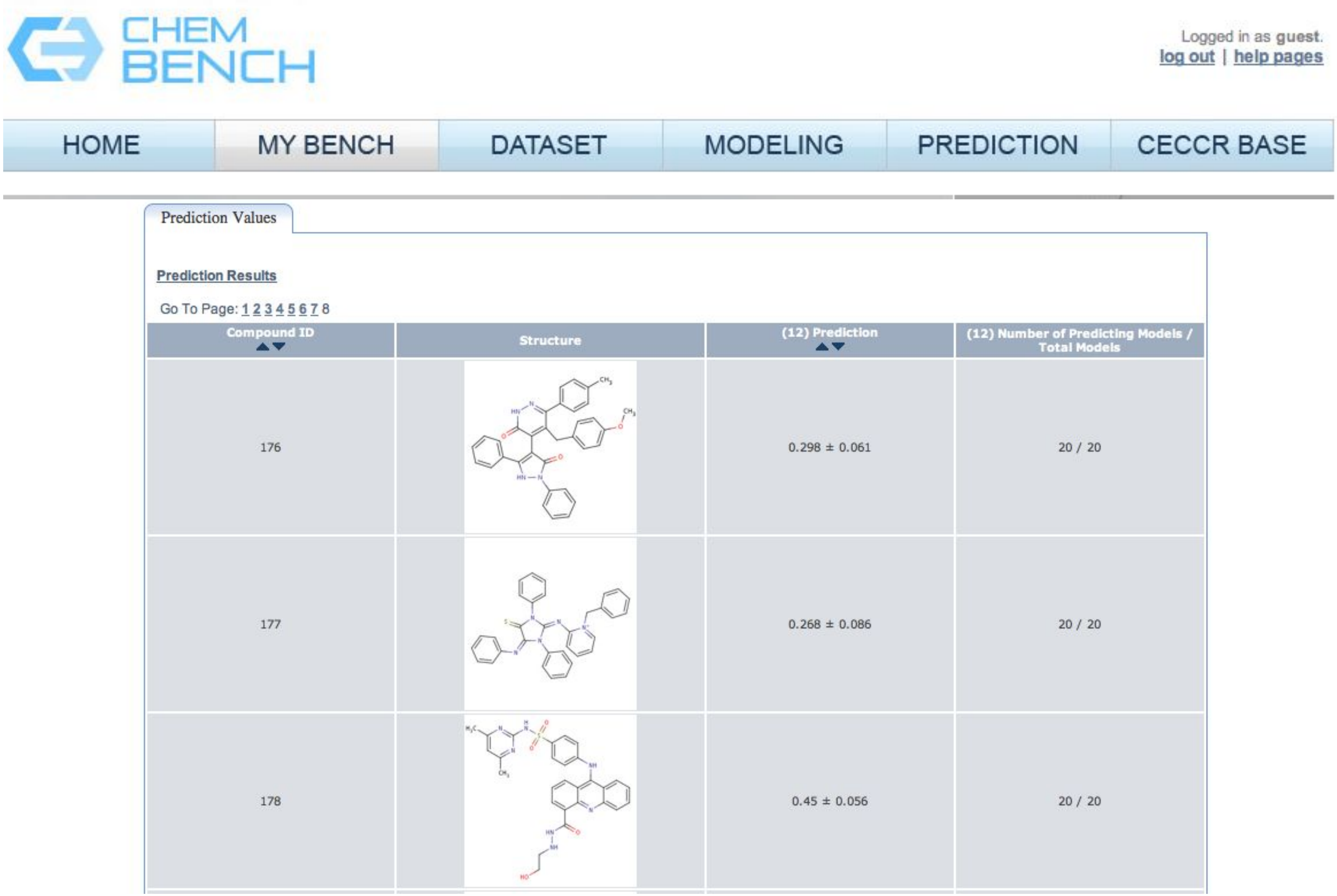
Users: 309

Jobs completed: 12477

Compute time used: 18.882 years

Current Users: 2

Running Jobs: 41



Prediction Results

Go To Page: 1 2 3 4 5 6 7 8

Compound ID	Structure	(12) Prediction	(12) Number of Predicting Models / Total Models
176		0.298 ± 0.061	20 / 20
177		0.268 ± 0.086	20 / 20
178		0.45 ± 0.056	20 / 20

Web Tools for QSAR/QSPR modeling: OpenTox

FP7 project OpenTox 2008-2011

Based on AMBIT, RESTful services

Provides webservices for model development, publishing and prediction:

Extended API:

Enables 3rd parties to develop their workflow

ToxCreate:

Method: Lazar regression/classification

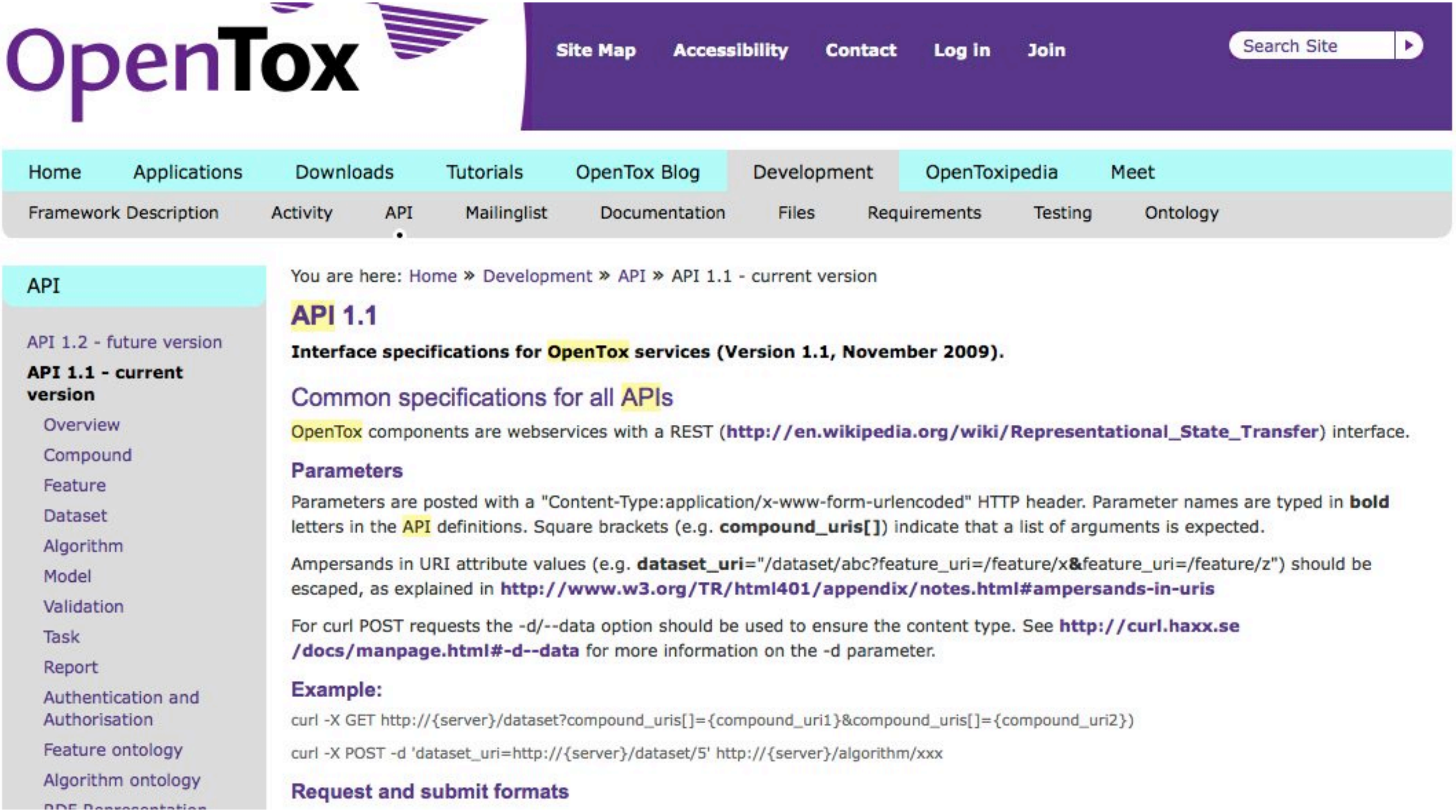
Descriptors: Fminer backbone refinement classes

ToxPredict:

Stores/calculates predictions for chemical compounds from ECHA list

Uses about 20 models collected from literature and developed by the authors

Web Tools for QSAR/QSPR modeling: OpenTox



You are here: Home » Development » API » API 1.1 - current version

API 1.1

Interface specifications for OpenTox services (Version 1.1, November 2009).

Common specifications for all APIs

OpenTox components are webservices with a REST (**http://en.wikipedia.org/wiki/Representational_State_Transfer**) interface.

Parameters

Parameters are posted with a "Content-Type:application/x-www-form-urlencoded" HTTP header. Parameter names are typed in **bold** letters in the API definitions. Square brackets (e.g. **compound_uris[]**) indicate that a list of arguments is expected.

Ampersands in URI attribute values (e.g. **dataset_uri**="/dataset/abc?feature_uri=/feature/x&feature_uri=/feature/z") should be escaped, as explained in **http://www.w3.org/TR/html401/appendix/notes.html#ampersands-in-uris**

For curl POST requests the -d/--data option should be used to ensure the content type. See **http://curl.haxx.se/docs/manpage.html#-d--data** for more information on the -d parameter.

Example:

```
curl -X GET http://{server}/dataset?compound_uris[]={compound_uri1}&compound_uris[]={compound_uri2})

curl -X POST -d 'dataset_uri=http://{server}/dataset/5' http://{server}/algorithm/xxx
```

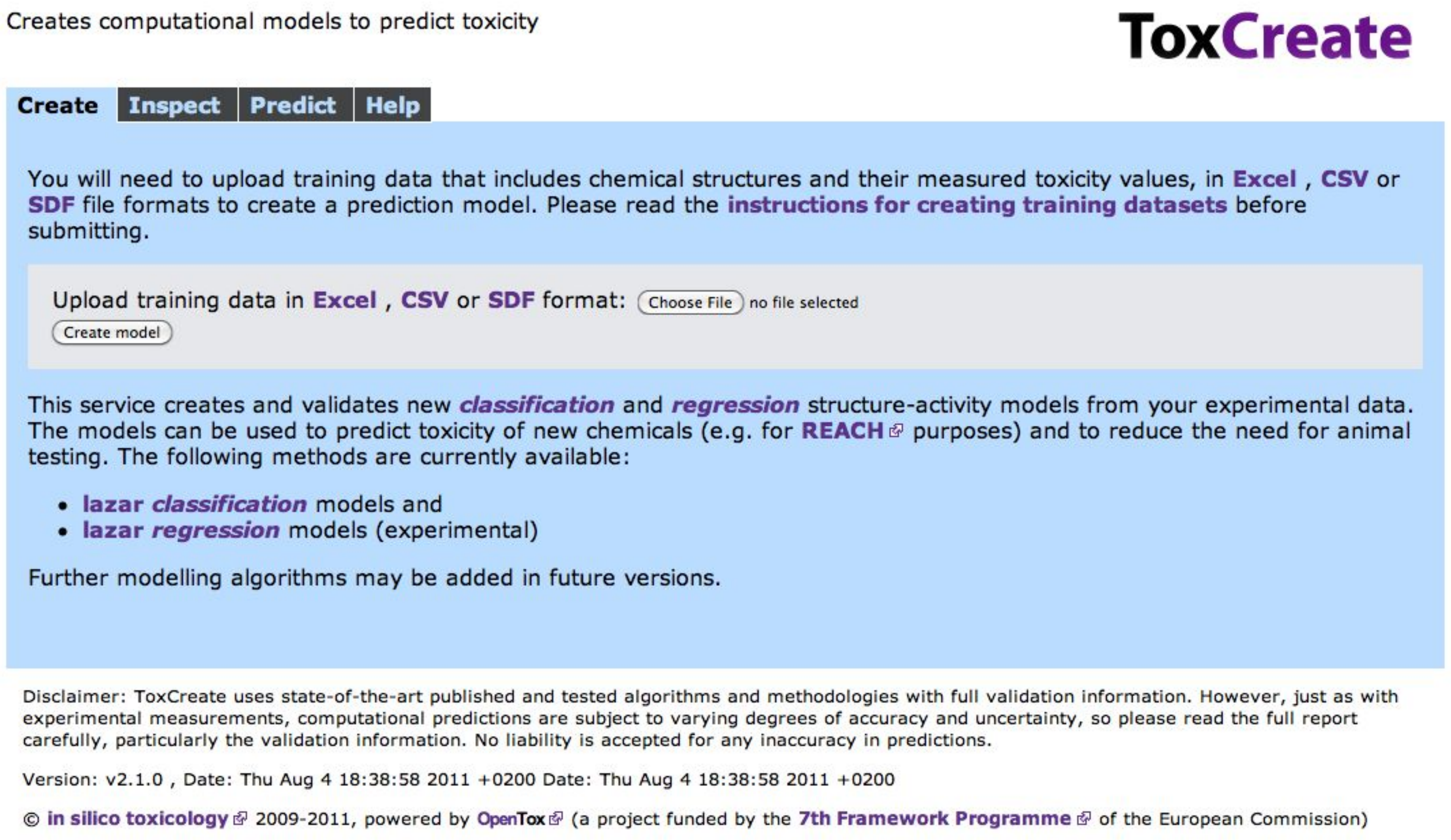
Request and submit formats

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

OpenTox meeting

eADMET THE REFERENCE IN CHEMOINFORMATICS

Model development using OpenTox



Creates computational models to predict toxicity

You will need to upload training data that includes chemical structures and their measured toxicity values, in **Excel** , **CSV** or **SDF** file formats to create a prediction model. Please read the **instructions for creating training datasets** before submitting.

Upload training data in **Excel** , **CSV** or **SDF** format:

This service creates and validates new ***classification*** and ***regression*** structure-activity models from your experimental data. The models can be used to predict toxicity of new chemicals (e.g. for **REACH** purposes) and to reduce the need for animal testing. The following methods are currently available:

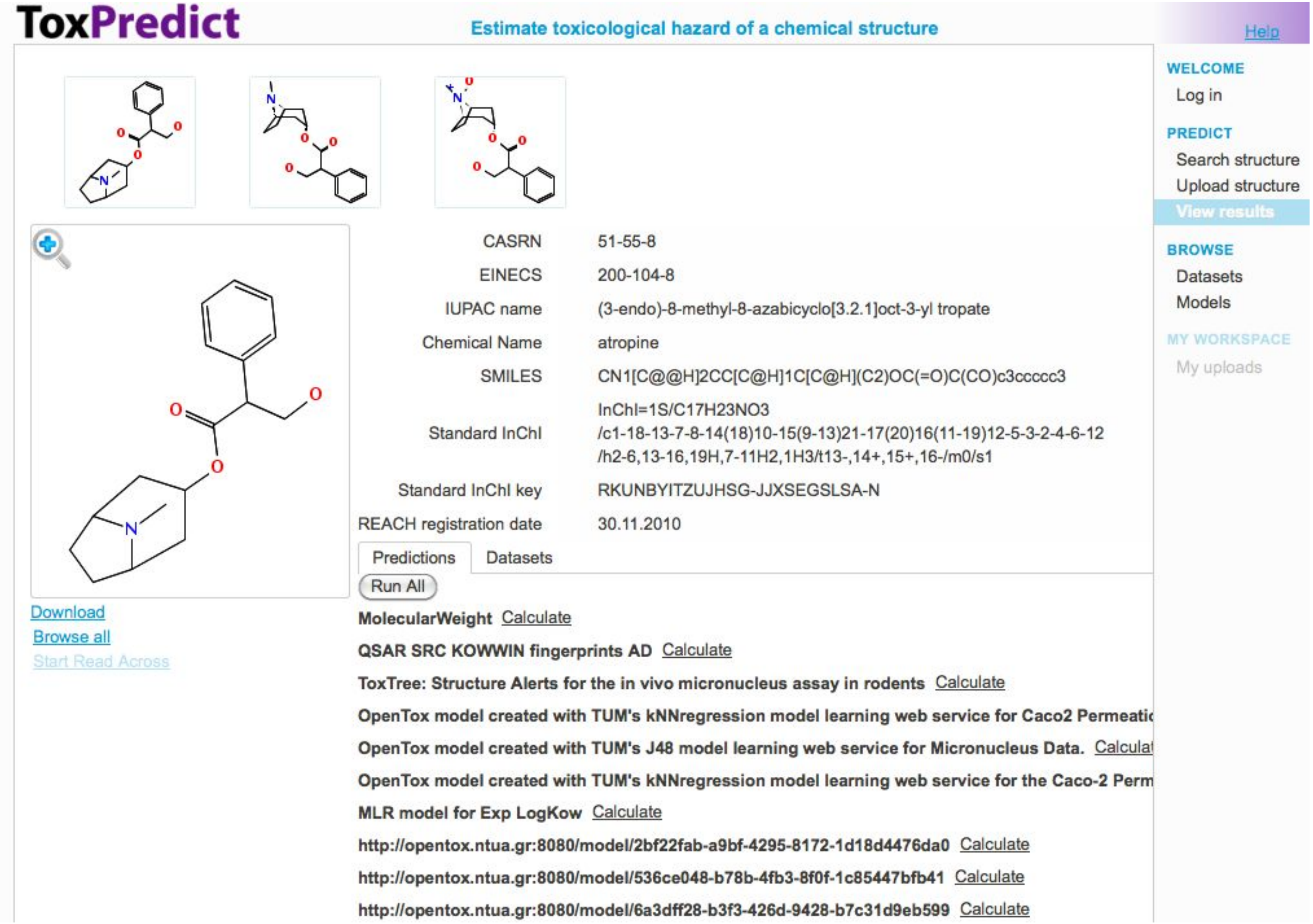
- **lazar *classification*** models and
- **lazar *regression*** models (experimental)

Further modelling algorithms may be added in future versions.

Disclaimer: ToxCreate uses state-of-the-art published and tested algorithms and methodologies with full validation information. However, just as with experimental measurements, computational predictions are subject to varying degrees of accuracy and uncertainty, so please read the full report carefully, particularly the validation information. No liability is accepted for any inaccuracy in predictions.

Version: v2.1.0 , Date: Thu Aug 4 18:38:58 2011 +0200 Date: Thu Aug 4 18:38:58 2011 +0200

© **in silico toxicology** 2009-2011, powered by **OpenTox** (a project funded by the **7th Framework Programme** of the European Commission)



CASRN	51-55-8
EINECS	200-104-8
IUPAC name	(3-endo)-8-methyl-8-azabicyclo[3.2.1]oct-3-yl tropate
Chemical Name	atropine
SMILES	CN1[C@@H]2CC[C@H]1C[C@H](C2)OC(=O)C(CO)c3ccccc3
Standard InChI	InChI=1S/C17H23NO3 /c1-18-13-7-8-14(18)10-15(9-13)21-17(20)16(11-19)12-5-3-2-4-6-12 /h2-6,13-16,19H,7-11H2,1H3/t13-,14+,15+,16-/m0/s1
Standard InChI key	RKUNBYITZUJHSG-JJXSEGSLSA-N
REACH registration date	30.11.2010

MolecularWeight Calculate

QSAR SRC KOWWIN fingerprints AD Calculate

ToxTree: Structure Alerts for the in vivo micronucleus assay in rodents Calculate

OpenTox model created with TUM's kNNregression model learning web service for Caco2 Permeatic

OpenTox model created with TUM's J48 model learning web service for Micronucleus Data. Calculat

OpenTox model created with TUM's kNNregression model learning web service for the Caco-2 Perm

MLR model for Exp LogKow Calculate

http://opentox.ntua.gr:8080/model/2bf22fab-a9bf-4295-8172-1d18d4476da0 Calculate

http://opentox.ntua.gr:8080/model/536ce048-b78b-4fb3-8f0f-1c85447bfb41 Calculate

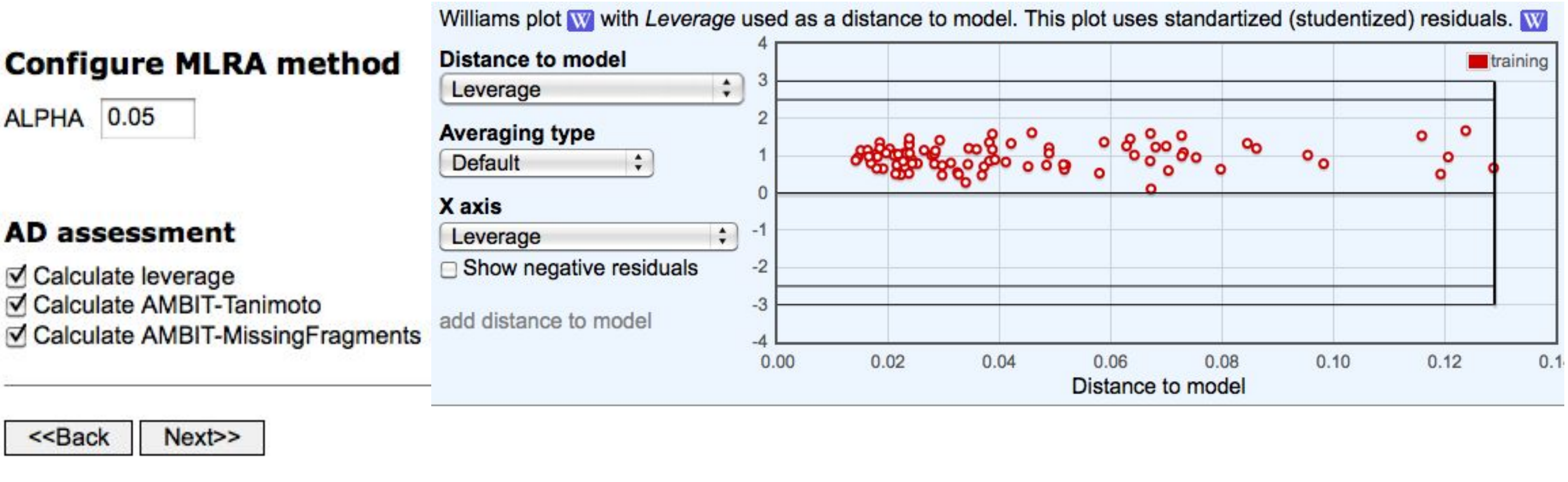
http://opentox.ntua.gr:8080/model/6a3dff28-b3f3-426d-9428-b7c31d9eb599 Calculate

OpenTox & CADASTER collaboration using API:

Web tools for similarity searching

Web tools for Applicability Domain calculations

Realized as calls to respective web services of OpenTox by means of LINUX "curl" command line tools



Configure MLRA method

ALPHA 0.05

AD assessment

☑ Calculate leverage
☑ Calculate AMBIT-Tanimoto
☑ Calculate AMBIT-MissingFragments

Williams plot with *Leverage* used as a distance to model. This plot uses standartized (studentized) residuals.

Distance to model Leverage

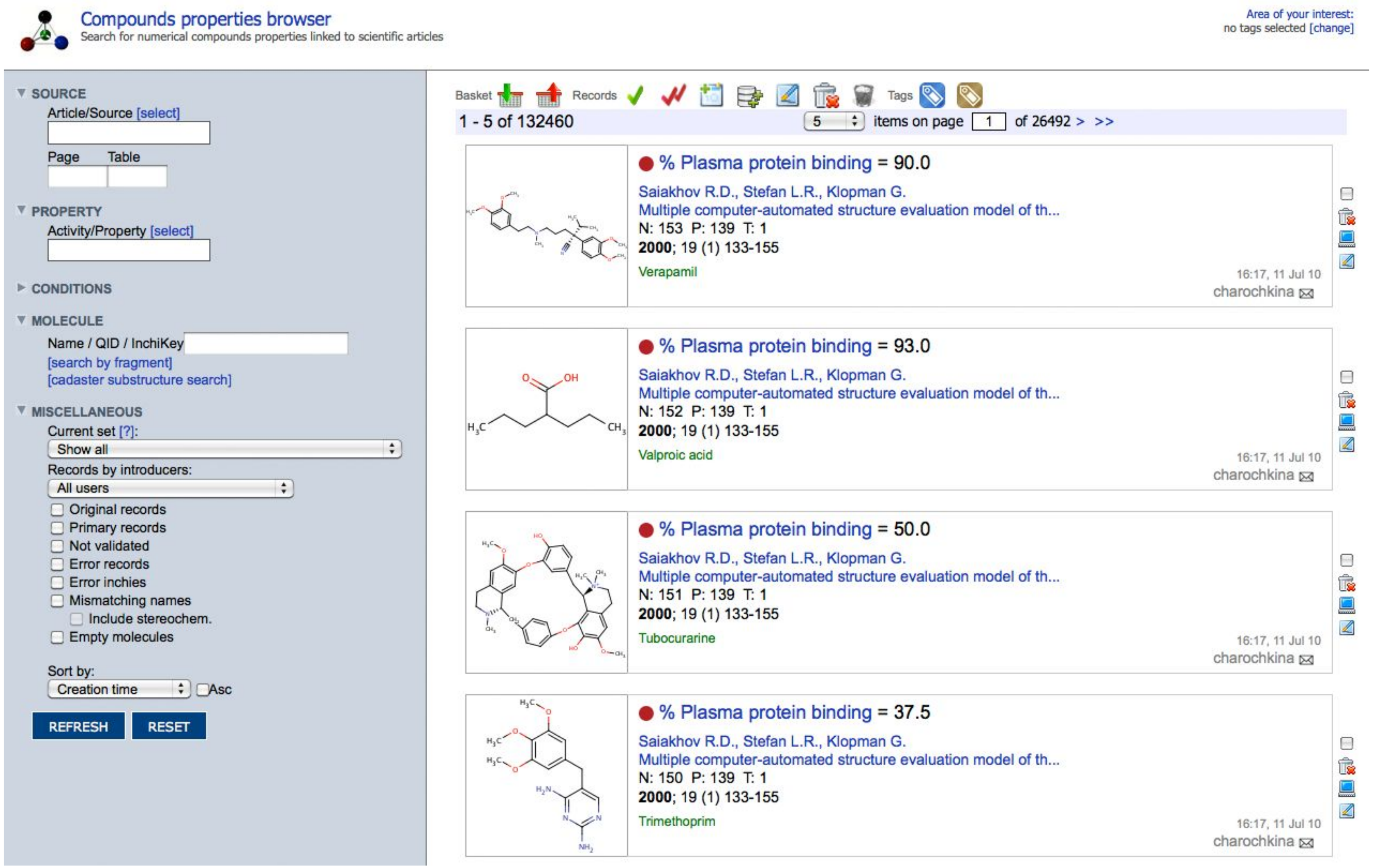
Averaging type Default

X axis Leverage

□ Show negative residuals

add distance to model

Database & model: OCHEM – On-line Chemical Modeling Environment http://ochem.eu



Compounds properties browser
Search for numerical compounds properties linked to scientific articles

Area of your interest:
no tags selected [change]

▼ SOURCE
Article/Source [select]
Page Table

▼ PROPERTY
Activity/Property [select]

▶ CONDITIONS

▼ MOLECULE
Name / QID / InchiKey
[search by fragment]
[cadaster substructure search]

▼ MISCELLANEOUS
Current set [?]: Show all
Records by introducers: All users

- □ Original records
- □ Primary records
- □ Not validated
- □ Error records
- □ Error inchies
- □ Mismatching names
  - □ Include stereochem.
- □ Empty molecules

Sort by: Creation time □Asc

REFRESH RESET

Basket Records Tags

1 - 5 of 132460 5 items on page 1 of 26492 > >>

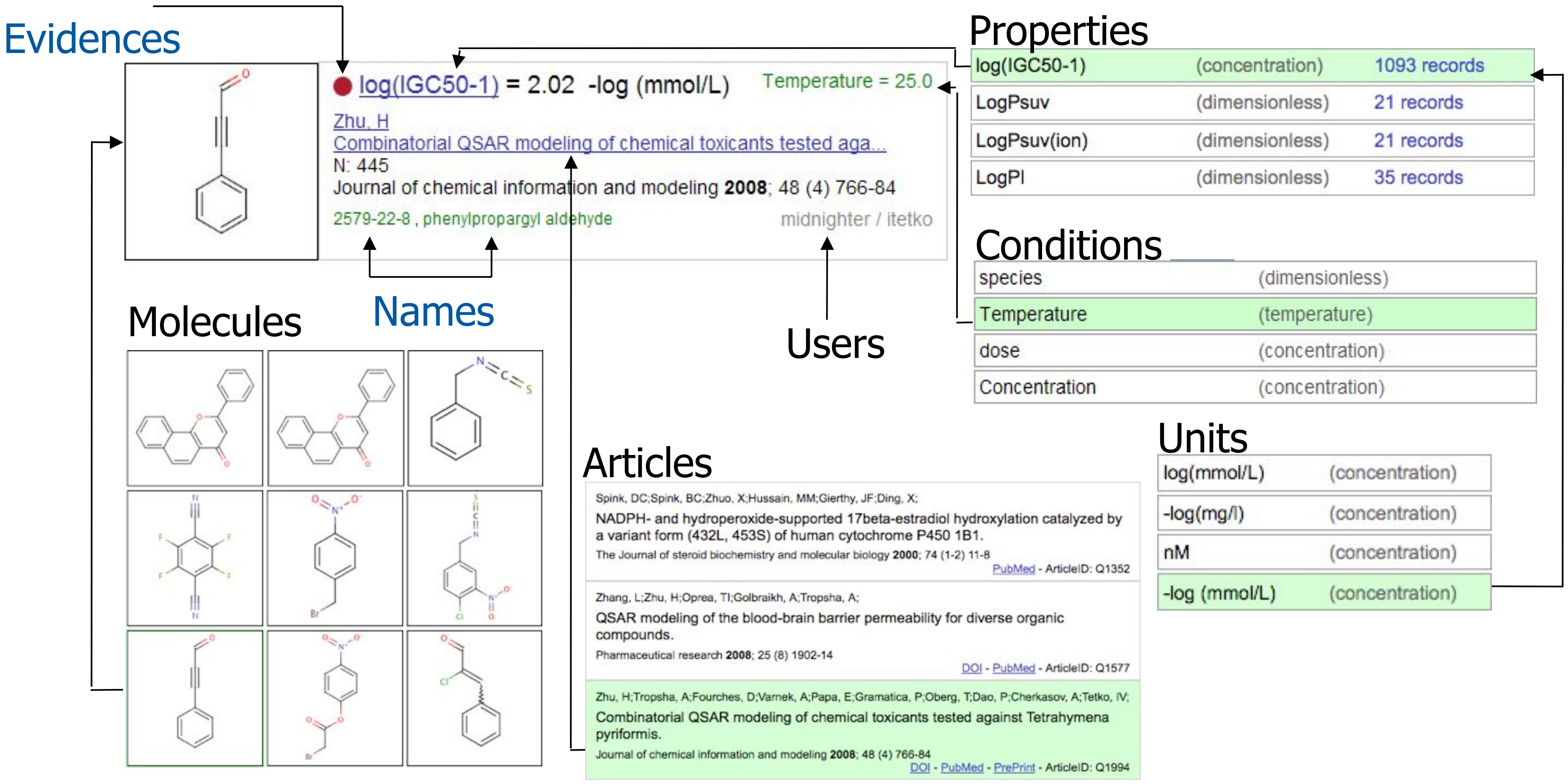
● % Plasma protein binding = 90.0
Saiakhov R.D., Stefan L.R., Klopman G.
Multiple computer-automated structure evaluation model of th...
N: 153 P: 139 T: 1
2000; 19 (1) 133-155
Verapamil
16:17, 11 Jul 10
charochkina

● % Plasma protein binding = 93.0
Saiakhov R.D., Stefan L.R., Klopman G.
Multiple computer-automated structure evaluation model of th...
N: 152 P: 139 T: 1
2000; 19 (1) 133-155
Valproic acid
16:17, 11 Jul 10
charochkina

● % Plasma protein binding = 50.0
Saiakhov R.D., Stefan L.R., Klopman G.
Multiple computer-automated structure evaluation model of th...
N: 151 P: 139 T: 1
2000; 19 (1) 133-155
Tubocurarine
16:17, 11 Jul 10
charochkina

● % Plasma protein binding = 37.5
Saiakhov R.D., Stefan L.R., Klopman G.
Multiple computer-automated structure evaluation model of th...
N: 150 P: 139 T: 1
2000; 19 (1) 133-155
Trimethoprim
16:17, 11 Jul 10
charochkina

Database schema
Simplified overview



Evidences

● log(IGC50-1) = 2.02 -log (mmol/L) Temperature = 25.0

Zhu, H
Combinatorial QSAR modeling of chemical toxicants tested aga...
N: 445
Journal of chemical information and modeling **2008**; 48 (4) 766-84
2579-22-8 , phenylpropargyl aldehyde midnighter / itetko

Molecules

Names

Users

Properties

log(IGC50-1)	(concentration)	1093 records
LogPsuv	(dimensionless)	21 records
LogPsuv(ion)	(dimensionless)	21 records
LogPl	(dimensionless)	35 records

Conditions

species	(dimensionless)
Temperature	(temperature)
dose	(concentration)
Concentration	(concentration)

Units

log(mmol/L)	(concentration)
-log(mg/l)	(concentration)
nM	(concentration)
-log (mmol/L)	(concentration)

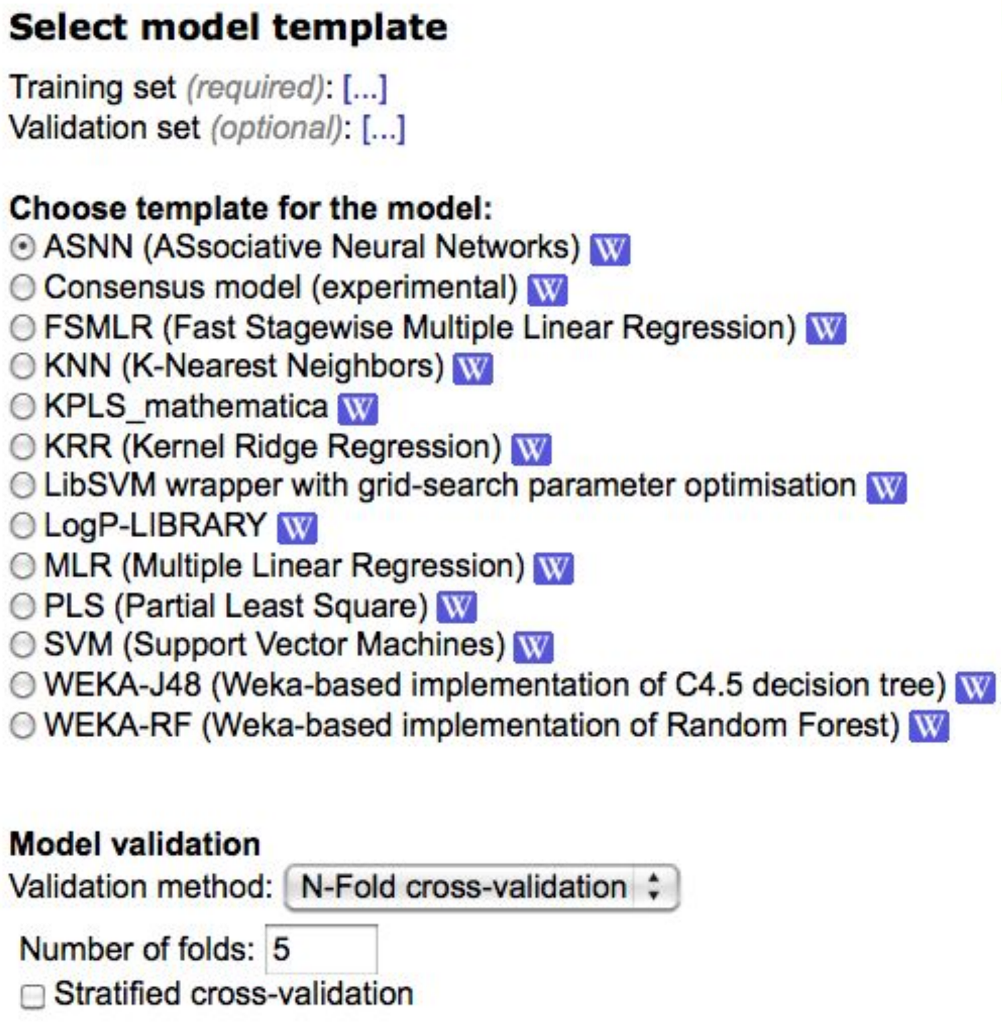
Articles

Spink, DC;Spink, BC;Zhuo, X;Hussain, MM;Gierthy, JF;Ding, X;
NADPH- and hydroperoxide-supported 17beta-estradiol hydroxylation catalyzed by a variant form (432L, 453S) of human cytochrome P450 1B1.
The Journal of steroid biochemistry and molecular biology **2000**; 74 (1-2) 11-8
PubMed - ArticleID: Q1352

Zhang, L;Zhu, H;Oprea, TI;Golbraikh, A;Tropsha, A;
QSAR modeling of the blood-brain barrier permeability for diverse organic compounds.
Pharmaceutical research **2008**; 25 (8) 1902-14
DOI - PubMed - ArticleID: Q1577

Zhu, H;Tropsha, A;Fourches, D;Varnek, A;Papa, E;Gramatica, P;Oberg, T;Dao, P;Cherkasov, A;Tetko, IV;
Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.
Journal of chemical information and modeling **2008**; 48 (4) 766-84
DOI - PubMed - PrePrint - ArticleID: Q1994

QSPR/QSAR modelling in OCHEM



Select model template

Training set *(required)*: [...]
Validation set *(optional)*: [...]

Choose template for the model:

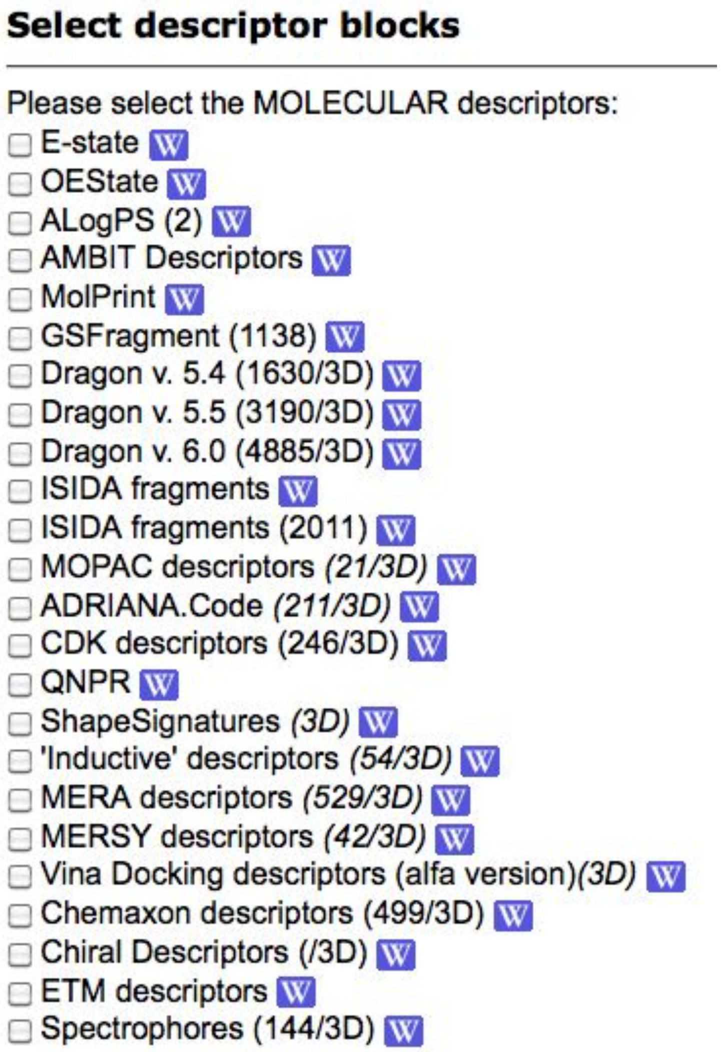
● ASNN (ASsociative Neural Networks) W
○ Consensus model (experimental) W
○ FSMLR (Fast Stagewise Multiple Linear Regression) W
○ KNN (K-Nearest Neighbors) W
○ KPLS_mathematica W
○ KRR (Kernel Ridge Regression) W
○ LibSVM wrapper with grid-search parameter optimisation W
○ LogP-LIBRARY W
○ MLR (Multiple Linear Regression) W
○ PLS (Partial Least Square) W
○ SVM (Support Vector Machines) W
○ WEKA-J48 (Weka-based implementation of C4.5 decision tree) W
○ WEKA-RF (Weka-based implementation of Random Forest) W

Model validation

Validation method: N-Fold cross-validation

Number of folds: 5

□ Stratified cross-validation

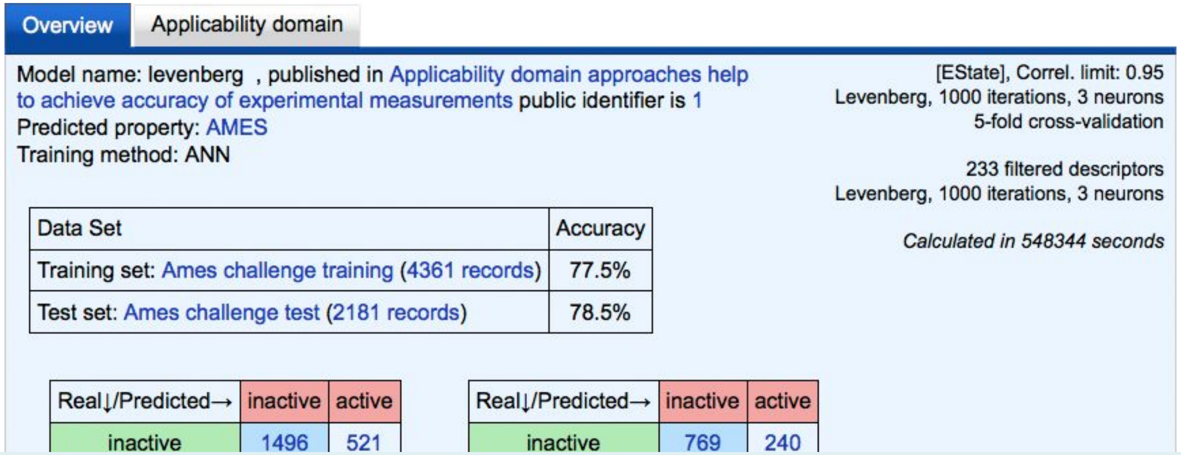


Select descriptor blocks

Please select the MOLECULAR descriptors:

□ E-state W
□ OEState W
□ ALogPS (2) W
□ AMBIT Descriptors W
□ MolPrint W
□ GSFragment (1138) W
□ Dragon v. 5.4 (1630/3D) W
□ Dragon v. 5.5 (3190/3D) W
□ Dragon v. 6.0 (4885/3D) W
□ ISIDA fragments W
□ ISIDA fragments (2011) W
□ MOPAC descriptors *(21/3D)* W
□ ADRIANA.Code *(211/3D)* W
□ CDK descriptors (246/3D) W
□ QNPR W
□ ShapeSignatures *(3D)* W
□ 'Inductive' descriptors *(54/3D)* W
□ MERA descriptors *(529/3D)* W
□ MERSY descriptors *(42/3D)* W
□ Vina Docking descriptors (alfa version)*(3D)* W
□ Chemaxon descriptors (499/3D) W
□ Chiral Descriptors (/3D) W
□ ETM descriptors W
□ Spectrophores (144/3D) W

Models: it’s everything about reliability



Model name: levenberg , published in Applicability domain approaches help to achieve accuracy of experimental measurements public identifier is 1
Predicted property: AMES
Training method: ANN

[EState], Correl. limit: 0.95
Levenberg, 1000 iterations, 3 neurons
5-fold cross-validation

233 filtered descriptors
Levenberg, 1000 iterations, 3 neurons

Calculated in 548344 seconds

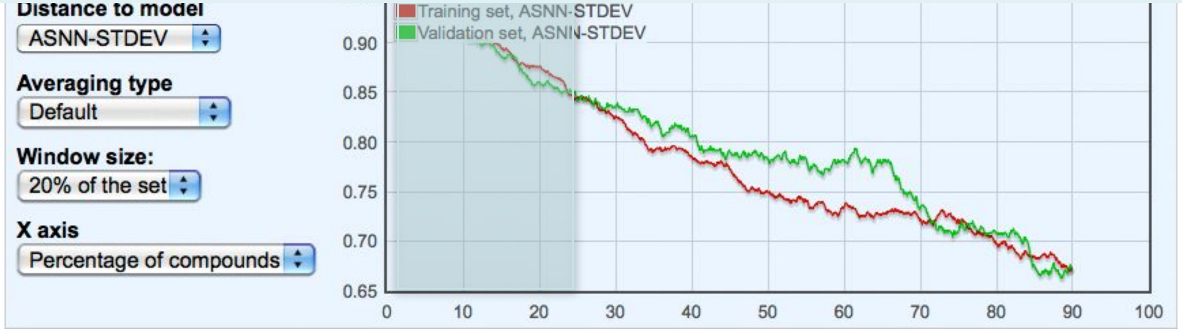
Data Set	Accuracy
Training set: Ames challenge training (4361 records)	77.5%
Test set: Ames challenge test (2181 records)	78.5%

Real↓/Predicted→	inactive	active
inactive	1496	521

Real↓/Predicted→	inactive	active
inactive	769	240

For more details: see poster

Applicability domain estimation for classification QSARs on example of Ames test and CYP450 inhibition



Distance to model: ASNN-STDEV
Averaging type: Default
Window size: 20% of the set
X axis: Percentage of compounds

OCHEM statistics

Available data	***Descriptors***	***Algorithms***
>200k records	20 providers	10 methods
>200 properties		
>7000 articles		

Model development facilities

120M entries, e.g. 120,000 molecules with 1,000 descriptors each

Limitation on the size of the model is 1G (mysql query)

Parallel processing on more than 400 CPUs

Available models

LogP, aqueous solubility, solubility in 5% and 100% DMSO, several pKa models, AMES test, CYP450, environmental toxicity + several other properties in development

“Environmental ChemOInformatics” (ECO) school at UFS Schneefernerhaus, Zugspitze, http://eco-itn.eu

Conclusions

Web tools are invaluable resources for collaboration on the web

- Important tools for REACH
- Development of web tools become more and more popular
- Several large projects, like eTOX, Open PHACTS, were launched
- Large companies demonstrate their interest in sharing pre-competitive data

Future trends:

Ontologies for better structuring of complex data and for integration of heterogeneous databases

Data sharing

Integration of *in vitro* and *in silico* data to model *in vivo* toxicities

Data curation and annotation

Text mining

Eva Schlosser
Iurii Sushko
Vlad Kholodovych
Wolfram Teetz
Robert Körner
Ahmed Abdelaziz
Sergii Novotarskyi
Stefan Brandmaier
Jacques Ehret

CADASTER FP7
ECO MC ITN