QSAR in Green Chemistry

- QSAR is the acronym for Quantitative Structure-Activity Relationship
- Chemistry is based on the premise that similar chemicals will behave similarly
- The behavior/activity of a chemical is derived from its structure
- QSAR research searches for relationships between chemical structure and activity

Why QSAR?

- Limited toxicity data on majority of chemicals
- Priority setting for testing
- In QSAR, estimating behavior of untested chemicals has been used for >60 years
QSAR basics

- QSARs associated with endpoint (i.e. LC50)
- QSAR attempt to group chemical behavior in terms of toxicity
- QSAR predicts biological activity (endpoints) directly from models of chemical structure
- Each QSAR predicts a specific endpoint

- Discrete Endpoints (e.g., Carcinogenicity, Mutagenicity)

```
Toxicity + Chemical + SAR
Endpoint Data Structure Data + Software -> Toxicity Response Predictions
```
What a QSAR needs

- Set of chemicals
- Molecular Descriptors for chemicals
- Data – Training and Test
- Model or Algorithm linking descriptors to endpoint
- Independent Validation (hopefully)
The use case can be divided into the following five steps:

1. Enter/select a chemical compound
2. Display selected/found structures
3. Select models
4. Perform the estimation
5. Display the results

→ Live Demo: www.toxpredict.org
Open Tox Examples

Input Structure

Out - Toxic or Not?
- LD50
- Liver Toxicity
- Secondary Metabolites
- Bioavailability
- Mutagenicity
- Carcogenicity
- Reproductive Toxicology
- Skin Irritation
- Aqua Toxicity
- Combined predictions for arrays of multiple end points
Molecular Descriptors

• Descriptors representing properties of complete molecules

• Descriptors calculated from 2D structure

• Descriptors requiring 3D representations
Descriptors representing properties of complete molecules

- **SMILES**-simplified molecular-input line-entry system - a machine-readable format - 2D representation of molecules as linear strings of alpha-numeric characters.

- **LogP**, Hydrophobicity
  - Log$P$ – the logarithm of the partition coefficient between $n$-octanol and water
  - ClogP (Leo and Hansch) – based on small set of values from a small set of simple molecules
## Molecular Descriptors

**SMILES**-simplified molecular-input line-entry system

<table>
<thead>
<tr>
<th>Structures</th>
<th>Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Structure 1" /></td>
<td>c1cccccc1</td>
</tr>
<tr>
<td><img src="image2" alt="Structure 2" /></td>
<td>Oe1cc(C)ccc1OC</td>
</tr>
<tr>
<td><img src="image3" alt="Structure 3" /></td>
<td>s1c2[nH0]ce[nH0]c2c(N)c1C(=0)OCC</td>
</tr>
<tr>
<td><img src="image4" alt="Structure 4" /></td>
<td>[S + 2][O−]<a href="CCC">O−</a>C1 = Cc2cccccc2OC1 = O</td>
</tr>
</tbody>
</table>
Descriptors calculated from 2D structure

- Single-valued descriptors calculated from the 2D graph of the molecule
  - Simple counts of features
  - Lipinski Rule of Five (H bonds, MW, etc.)
  - Number of ring systems
  - Number of rotatable bonds

- Characterize structures according to size, degree of branching, connectivity and overall shape
  - Wiener Index – counts the number of bonds between pairs of atoms and sums the distances between all pairs
  - Randić (et al.) branching index - Defines a “degree” of an atom as the number of adjacent non-hydrogen atoms
    - Bond connectivity value is the reciprocal of the square root of the product of the degree of the two atoms in the bond.
    - Branching index is the sum of the bond connectivities over all bonds in the molecule.
  - Chi indexes – introduces valence values to encode sigma, pi, and lone pair electrons
  - Kappa Shape Indexes - characterize aspects of molecular shape
    - Compare the molecule with the “extreme shapes” possible for that number of atoms
    - Range from linear molecules to completely connected graph
2D fingerprints

Nodes do not represent atoms but features such as functionally important groups or whole ring system.

---

Figure 1.12. Examples of reduced graphs. Nodes corresponding to aromatic rings (Ar), aliphatic rings (R), functional groups (F) and linking groups (L) are shown (adapted from Gillet et al. 2003)
DESCRIPTORS BASED ON 3D REPRESENTATIONS

• Require the generation of 3D conformations
  – Can be computationally time consuming with large data sets
  – Usually must take into account conformational flexibility
  – 3D fragment screens encode spatial relationships between atoms, ring centroids, and planes

• Toxicophores
  – Based on atoms or substructures thought to be relevant for toxicity
  – Expert driven
DRAGON

- Dragon - [http://www.vcclab.org/lab/edragon/](http://www.vcclab.org/lab/edragon/)
  - an application for the calculation of molecular descriptors
  - 1,600 molecular descriptors that are divided into 20 logical blocks.

- ACD Labs - [http://www.acdlabs.com](http://www.acdlabs.com)
  - ACD Labs values now incorporated into the CAS Registry File for millions of compounds
    - Name generation
    - NMR prediction
    - Physical property prediction
Open Tox Descriptor Algorithms

ToxDesc

http://opentox-dev.informatik.tu-muenchen.de:8080/ToxDesc

11 descriptor calculation algorithms

• Existing methods
  – OpenBabel, JOELib2, The Chemistry Development Kit (CDK) and gSpan’,
• New methods
  – FMiner, MakeMNA, and MakeQNA.
• AMBIT
  – The single descriptor calculation web service – developed by OpenTox partner IDEA – offers descriptors calculated by several packages
  – http://www.opentox.org/dev/documentation/components/ambit
  – http://ambit.sourceforge.net/
Open Tox Descriptor Algorithms

![Screenshot of Ambit website with molecular structures and data]

The image shows a screenshot of the Ambit website, which is a tool for generating molecular descriptors and analyzing chemical compounds. The website provides functionalities such as searching for substructures and properties, and displaying information about compounds in the EU's ECHA database. The screenshot includes two compounds with their respective registration dates, CAS numbers, and ECHA identifiers. The compounds are represented with molecular structures, and their names are shown as linalyl cinnamate and trimethoxycinnamic acid. The site mentions that some services are under development.
## Open Tox Descriptor Algorithms

<table>
<thead>
<tr>
<th>Find</th>
<th>Name</th>
<th>Units</th>
<th>Origin (Dataset, Model or Algorithm)</th>
<th>Nominal values</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="Structural Alert for genotoxic carcinogenicity" /></td>
<td>Structural Alert for nongenotoxic carcinogenicity</td>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" /></td>
<td>Benign / Bossa rulebase (for mutagenicity and carcinogenicity)</td>
<td>NO</td>
</tr>
<tr>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="Potential S. typhimurium TA100 mutagen based on QSAR" /></td>
<td>Unlikely to be a S. typhimurium TA100 mutagen based on QSAR</td>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" /></td>
<td>Benign / Bossa rulebase (for mutagenicity and carcinogenicity)</td>
<td>NO</td>
</tr>
<tr>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="" /></td>
<td>Potential carcinogen based on QSAR</td>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" /></td>
<td>Benign / Bossa rulebase (for mutagenicity and carcinogenicity)</td>
<td>NO</td>
</tr>
<tr>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="Unlikely to be a carcinogen based on QSAR" /></td>
<td>For a better assessment a QSAR calculation could be applied.</td>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" /></td>
<td>Benign / Bossa rulebase (for mutagenicity and carcinogenicity)</td>
<td>NO</td>
</tr>
<tr>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="Negative for genotoxic carcinogenicity" /></td>
<td>Negative for nongenotoxic carcinogenicity</td>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" /></td>
<td>Benign / Bossa rulebase (for mutagenicity and carcinogenicity)</td>
<td>NO</td>
</tr>
<tr>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="Error when applying the decision tree" /></td>
<td></td>
<td><img src="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" alt="http://www.opentox.org/echaEndpoints.owl#Carcinogenicity" /></td>
<td>Benign / Bossa rulebase (for mutagenicity and carcinogenicity)</td>
<td>NO</td>
</tr>
</tbody>
</table>
Open Tox Descriptor Algorithms

The available models are listed under the “Predictions” tab. Click on the “Calculate” link next to each model to obtain predictions.
Open Tox Descriptor Algorithms

The available models are listed under the “Predictions” tab. Click on the “Calculate” link next to each model to obtain predictions. Chemical structure and prediction results can be downloaded via the “Download” link.
Open Tox Data Sets

- Textual databases (e.g., IARC [35], NTP [36]);
- Machine readable files (e.g., .sdf)
  - DSSTox, ISSCAN, AMBIT, RepDose
- Public Databases
- PubChem, ACToR, CPDBAS, DBPCAN, EPAFHM, KIERBL, IRISTR, FDAMDD, ECETOC skin irritation, LLNA skin sensitization and the Bioconcentration factor (BCF)
  Gold Standard Database
Open Tox Dataset Selection

The toxicity data for the selected compound could be accessed at the “Datasets” tab. The top level categories show the endpoint name and the number of datasets (in brackets) containing toxicity information for the selected compounds. Click the “Show” link to view the list of the datasets.
Open Tox Dataset Selection

- Pressing the “Show” link next to the endpoint name list the available datasets.
- The “Show” link next to the dataset title displays the content for the selected compound.
- To browse the entire dataset, use “Browse dataset” link. Note this will bring you to another page, but you could access the current page either via browser “Back” button, or the ToxPredict “View Results” menu on the right.
• Click the “Browse dataset” link next to the ISSCAN dataset. This will show the “Datasets” page.
Chemistry Space of the NCTR estrogen binding Dataset

NCTR dataset:
- Strong
- Weak
- inactive

EPA dataset (58K)

- Competitive binding assays for the estrogen nuclear receptor proteins
- **NCTR dataset cover a significant part of the chemistry space defined by the EPA dataset**
- Most active compounds cluster together, whereas inactives scatter over the space.
Data sets

NCTR Dataset
(130 active and 100 inactive)

Relative Binding Activity (RBA)

Over six orders of magnitude of RBA range

A wide range of chemical classes
Algorithms

• Linear QSAR Models
• Supervised classification
• Unsupervised classification
  – Clustering approaches
• Machine Learning methods
  – Genetic Algorithms
  – Neural Networks
  – Self Organizing Maps (SOM)
• Toxicophore search
Open Tox Algorithms

• Feature Selection Algorithms - algorithms for the reduction of the dimensionality of a dataset, by selecting only a subset of a full set of descriptors included in the dataset
  – including a partial least-squares filter, principle components analysis (PCA), chi-squared attribute evaluation, information-gain attribute evaluation, a scaling filter, and a missing value replacer;

• Un Supervised Clustering Algorithms
  – Standard algorithms - unsupervised learning algorithms that group objects of similar kind into respective categories.
  – TUM’s StructuralClustering procedure

• Supervised Classification and Regression Algorithm
  – several common (Q)SAR modelling algorithms (Linear Regression, Multiple Linear Regression, PLS, Gaussian Processes, Neural Networks)
  – standard machine learning methods (SVMs, KNN, M5’, J48, Naïve Bayes)
  – in house algorithms from partners (lazar, ToxTree, BBRC, LastPM, LoMoGraph, MaxTox, iSAR, Three Conditional Density Estimators, MakeSCR)

• Applicability domain estimation algorithms – is the chemical space adequately assessed
Linear QSAR models

- **Simple linear model**
  - \( y_{pred} \) is the model output,
  - \( x_i \) and \( w_i \) the \( i \)th descriptor (out of a total \( M \)) (weight) respectively, and
  - \( w_0 \) an offset term

- **Multiple linear Regression models**
- **Principal component regression models**
- **Partial Least squares regression**
- **Linear Discriminant analysis**
- **Naïve Bayes**

Aromatic amines mutagenicity model

\[
\log \text{TA100} = 0.92 \text{Kow} + 1.17 \text{HOMO} - 1.18 \text{LUMO} + 7.35
\]

\( \log \text{TA100} \) is the mutagenic potency (revertants/Nmol)
\( \text{Kow} \) hydrophobicity
HOMO is the energy of the highest occupied molecular orbital
LUMO is the energy of the lowest unoccupied molecular orbital.
Supervised Classification Approach

- Pattern recognition methods:
  - Classification and regression tree (CART)
  - K-nearest neighbor (KNN)
  - Artificial neural network (ANN)
  - SIMCA
- Similar model results obtained using various pattern recognition methods
- Overall correct prediction rate is between 70 - 80%
- The nature of descriptors are more critical
CART Model

Classification and Regression Tree (CART)

- Unique contribution to the overall prediction
- Descriptors are selected based on genetic algorithm
  - \( \text{clogP (hydrophobicity)} \)
  - \( \text{Phenolic ring indicator} \)
  - \( \text{Shadow-XY (length of structure)} \)
  - \( \text{RPCS (relative positive charge surface area)} \)
  - \( \text{PNSA (total charge weighted negative surface area)} \)

CART uses a decision tree to display how chemical may be classified or predicted through a series of rules. These rules are expressed as “if ... then...”.
Similarity Search

Cluster Identification

Test Compound

Safety Data on the Chemical and Similar Substances

Similarity Index
1.0000
Triazolam
Unsupervised clustering

Compound Classification and Selection

Agglomerative methods start at the bottom and merge similar clusters (bottom-up)

Divisive hierarchical clustering starts with all compounds in a single cluster and partitions the data (top-down)
Machine Learning Methods – Genetic algorithms

- Different parameters and model solutions to given problems are encoded in a chromosome and subjected to iterative random variation, thus generating a population.
- Solutions provided by these chromosomes are evaluated by a fitness function that assigns high scores to desired results.
- Chromosomes yielding the best intermediate solutions are subjected to mutation and crossover operation that correspond to random genetic mutations and gene recombination events.
- The resulting modified chromosomes represent the next generation and the process is continued until the obtained results meet a satisfactory convergence criterion.
Toxicophore Searching

Template Structure

Query

Searching ...

Hits

Database with multiple conformers

Match
Estimating Toxic Potential Using toxicophore searching

Reduce test molecule to all possible 2 - 10 atom fragments
Compare molecular fragments to active and inactive fragments in the control toxicity database

Fragments Exclusively Associated with Toxicity (~200, e.g.)
Fragments NOT Exclusively Associated with Toxicity (~500,000, e.g.)

Identity of Chemicals with Structural Alert Fragments
Toxic Potency
Number of Compounds / SA
Presence and Identity of Modulators
QSAR Properties

Estimate Toxic Potential
Open Tox Model/Algorithm Selection

- Select “Models” from ToxPredict menu on the right.
- The list of models appears. Only selected models will be displayed/used when running predictions on other ToxPredict pages.
- Select/Unselect models one by one or by filtered subsets.
  - Use the Endpoints drop down box to filter models by endpoint.
  - The “Select/Unselect” checkbox on the toolbar is applied only on currently visible models.
- Click the “Show” link next to each model to view more information about a model

When ready, click on “Search structure” menu on the right.
Open Tox Model/Algorithm Selection

Model details includes information about the algorithm, used to derive the model, independent and dependent variables, training dataset, access rights, and links to validation performed and resulted validation reports.

Models are OpenTox web services, derived by learning algorithms, also OpenTox web services. Validations and validation reports are created by the OpenTox Validation service.
The selected models are listed under the “Predictions” tab. Click on the “Calculate” link next to each model to obtain predictions. Select the “Dataset” tab to verify if there is toxicity data for a compound.

As an alternative to inspecting structures one by one, click on the “Browse All” link on the left (just below the selected compound diagram).
Assignment for next week

• Apply ToxPredict on your set of chemicals
  – Focus on algorithms/models which use QSAR/SAR based predictions
    • Compare QSAR predictions with Toxicophore based predictions
  – Modify your compound in similar ways as last week
    • Run QSAR predictions again
    • Compare with Toxicophore based predictions
Validation

False Positives
Actual Negative
Predicted Positive

Predictive Value
Predicted Positive
Actual Positive

Specificity
Actual Negative
Predicted Negative

Sensitivity
Actual Positive
Predicted Positive
ToxPredict: Step 1 (behind the scenes)

Find structure by name, registry number, SMILES, InChI, structure, substructure, similarity...

OT Dataset API HTTP GET

text/uri-list, application/rdf+xml, chemical/x-daylight-smiles chemical/x-mdl-sdf, ...

Here is the list of structures as URI links, RDF, MOL or SMILES.

ToxPredict Web Application

OT Ontology Service
<table>
<thead>
<tr>
<th>Model</th>
<th>Endpoint</th>
<th>Algorithm</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Weight</td>
<td>Acute toxicity to fish (lethality)</td>
<td>Molecular Weight</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Verhaar scheme for predicting toxicity mode of action</td>
<td>Acute toxicity to fish (lethality)</td>
<td>ToxTree: Verhaar scheme for predicting toxicity mode of action</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Benigni/Bossa rules for carcinogenicity and mutagenicity</td>
<td>Carcinogenicity</td>
<td>ToxTree: Benigni/Bossa rules for carcinogenicity and mutagenicity</td>
<td></td>
</tr>
<tr>
<td>pKa</td>
<td>Dissociation constant (pKa)</td>
<td>pKa</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Structure Alerts for the in vivo micronucleus assay in rodents</td>
<td>Endpoints</td>
<td>ToxTree: Structure Alerts for the in vivo micronucleus assay in rodents</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Michael acceptors</td>
<td>Endpoints</td>
<td>ToxTree: Michael acceptors</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Eye irritation</td>
<td>Eye irritation/corrosion</td>
<td>ToxTree: Eye irritation</td>
<td></td>
</tr>
<tr>
<td>Caco-2 Cell Permeability</td>
<td>Gastrointestinal absorption</td>
<td>Regression: Linear regression</td>
<td>Model validation report</td>
</tr>
<tr>
<td>OpenTox model created with TUM's PL regression model learning web service.</td>
<td>Gastrointestinal absorption</td>
<td><a href="http://open.tox.informatik.tu-muenchen.de:3080/OpenTox-dev/algorithm/PL">http://open.tox.informatik.tu-muenchen.de:3080/OpenTox-dev/algorithm/PL</a> Regression</td>
<td></td>
</tr>
<tr>
<td>OpenTox model created with TUM's kNN regression model learning web service.</td>
<td>Gastrointestinal absorption</td>
<td><a href="http://open.tox.informatik.tu-muenchen.de:3080/OpenTox-dev/algorithm/kNN">http://open.tox.informatik.tu-muenchen.de:3080/OpenTox-dev/algorithm/kNN</a> Regression</td>
<td></td>
</tr>
<tr>
<td><a href="http://open.tox.informatics.ru:3000/model/675a80e6-b2d9-45c1-ba42-44b9085b5898">http://open.tox.informatics.ru:3000/model/675a80e6-b2d9-45c1-ba42-44b9085b5898</a></td>
<td>Gastrointestinal absorption</td>
<td>Multiple Linear Regression Training Algorithm</td>
<td></td>
</tr>
<tr>
<td>Lipinski Rule of Five</td>
<td>Human health effects</td>
<td>Lipinski Rule of Five</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Cramer rules</td>
<td>Human health effects</td>
<td>ToxTree: Cramer rules</td>
<td></td>
</tr>
<tr>
<td>XLogP</td>
<td>Octanol-water partition coefficient (Kow)</td>
<td>XLogP</td>
<td></td>
</tr>
<tr>
<td>START biodegradation and persistence plug-in</td>
<td>Persistence, Biodegradation</td>
<td>START biodegradation and persistence plug-in</td>
<td></td>
</tr>
<tr>
<td>SmartCYP: Cytochrome P450 Mediated Drug Metabolism</td>
<td>Protein-binding</td>
<td>SmartCYP: Cytochrome P450 Mediated Drug Metabolism</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Skin irritation</td>
<td>Skin irritation/corrosion</td>
<td>ToxTree: Skin irritation</td>
<td></td>
</tr>
<tr>
<td>ToxTree: Skin sensitisation alerts (M. Cronin)</td>
<td>Skin sensitisation</td>
<td>ToxTree: Skin sensitisation alerts (M. Cronin)</td>
<td></td>
</tr>
</tbody>
</table>
ToxPredict: Step 2 (Verify structure(s))

Select and/or edit structure(s)

OT Dataset API HTTP GET
(HTTP POST for structure editing)

text/uri-list,
application/rdf+xml,
chemical/x-daylight-smiles
chemical/x-mdl-sdfile,
image/png

Here is the list of structures as URI links, RDF, MOL, SMILES or images.
**CASN 71-43-2**

**Synonym(s)**: benzene; (6)annulene; benzene; Benzol; Benzene; bicarbonate of hydrogen; carbon oil; Coal naphtha; cyclohexadiene; mineral naphtha; motor benzol; nitrated benzene; Phenol; Phenyldimethyl; pyrobenzol; Phenol; BENZENE; BICYCLO[5.5.0]DECANE

**EINECS**: 200-733-7

**IUPAC name**: benzene

**InChI Key**: \texttt{UHOVZQYV3NRB-UHFFFAOYSA-N}

**InChI std**: \texttt{InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H}

**REACH Registration Date**: 30.11.2010

**SMILES**: \texttt{ccccc1C=CC=CC=C1}

**OpenTox model created with TUM's PLS regression model learning web service.**


**Tetrahedron**

**Lipinski Failures**

**Acute toxicity to fish, lethality**

**ToxTree: Verhaer scheme for predicting toxicity mode of action**

**Verhaer scheme**

Class 1 (narcosis or baseline toxicity)

**Carcinogenicity**

**ToxTree: Benigni/Bossa rules for carcinogenicity and mutagenicity**

- Structural Alert for genotoxic carcinogenicity
- Structural Alert for nongenotoxic carcinogenicity
- Potentially S. typhimurium TA100 mutagen based on QSAR
- Unlikely to be a S. typhimurium TA100 mutagen based on QSAR
- Potential carcinogen based on QSAR
- Unlikely to be a carcinogen based on QSAR
- For a better assessment a QSAR calculation could be applied.
- Negative for genotoxic carcinogenicity
- Negative for nongenotoxic carcinogenicity
- Structural Alert for genotoxic carcinogenicity explaining

**QSAR**

- Acyl halides: No
- Alkyl (C<5) or benzyl ester of sulphonic or phosphonic acid: No
- N-methylol derivatives: No
- Nonephalolone: No
- S or N mustard: No
- Propalectones and propolulones: No
- Promoters and aldehydes: No
- Other: No

**Conclusion**

No indication of carcinogenicity or mutagenicity.
What prediction models are available? Is there a model for endpoint X?

ToxPredict Web Application

HTTP GET SPARQL query

OT Ontology Service

application/sparql-results+xml

Here is the list of model URIs and related endpoints and algorithms in SPARQL format.