Neural Network Programs for the Estimation of Toxicity and Receptor Binding

39th IUPAC Congress & 86th Conference
of the Canadian Society for Chemistry

August 2003

by

Klaus L.E. Kaiser, Ph.D., FCIC, C.Chem.

Director of Research

TerraBase Inc.

TerraBase Inc.

TerraQSAR™

stand-alone,

probabilistic neural network (PNN) - based

toxicity/effect computation programs for PCs

TerraBase Inc.

Principle of Method

TerraQSAR™ toxicity estimation software is based on the probabilistic neural network (PNN) methodology, using the molecular structure of the substance under investigation as the sole input.

Purpose

TerraQSAR™ - FHM is a specialized, neural network-based software program, designed and optimized solely for the computation of acute (96-hr) median lethal concentrations (LC50) of organic (carbon-containing) substances with a defined chemical structure to the fish fathead minnow (*Pimephales promelas).*

Purpose

TerraQSAR™ - RMIV is a specialized, neural network-based software program, designed and optimized solely for the computation of the intravenous lethal dose (LD50) of organic (carbon-containing) substances with a defined chemical structure to rat and mouse.

Purpose

TerraQSAR™ - E2-RBA is a specialized, neural network-based software program, designed and optimized solely for the computation of estrogen receptor binding affinity (E2-RBA) of organic (carbon-containing) substances with a defined chemical structure relative to that of 17*beta*-estradiol (E2).

Purpose

TerraQSAR™ - HIV is a specialized, neural network-based software program, designed and optimized solely for the computation of human immunodeficiency (HIV-1) effectice concentrations (EC50) of organic (carbon-containing) substances with a defined chemical structure to the HIV-1 virus [fall 2003].

Input

TerraQSAR modules use as input only the chemical's SMILES code (2-D or 3-D), which is an international code for the representation of chemical structures and amenable to computer analysis.

SMILES = acronym for
Simplified **M**olecular **L**ine **E**ntry **S**ystem

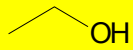
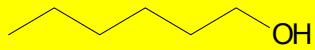
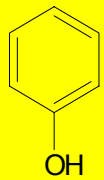
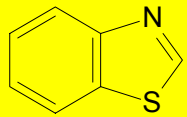
Output

The **TerraQSAR™** programs typically compute the **LC50, LD50, or EC50** in both mg/L or mg/kg b.w. (body weight) and in pT (log [L/mmole] or log ([kg b.w.]/mmole) units, as well as the molecular weight (MW) of substances entered, directly from the compound's structure.

Data Set

The **TerraQSAR™ - FHM** fathead minnow toxicity estimation program is based on a data set of measured 96-hr LC50 values for fathead minnow for 886 organic (carbon-containing) compounds. These data are widely available, both from non-commercial sources, such as the US Environmental Protection Agency's Ecotox database, or from commercial sources, such as TerraBase Inc.'s **TerraTox™ - Explorer** database.

SMILES string examples and structures (hydrogen atoms omitted)

ethanol	**CCO**	
n-hexanol	**CCCCCCO**	
phenol	**c1ccccc1O**	
benzothiazole	**c1cccc2ncsc12**	

Note regarding the SMILES notation

TerraBase Inc. uses the Accelrys Corp. system of the SMILES notation, which is not fully compatible with the Daylight Corp. system.

This is not to be considered as a value or preference statement, but merely an acknowledgement of the use of Accelrys software.

Table 2. Valid and invalid examples of SMILES code.

Substance	SMILES not valid	SMILES valid
cyclopentadiene [a]	c1cccC1	C1=CC=CC1
coumarin [a]	c1cc2OC(=O)ccc2cc1	c1cc2OC(=O)C=Cc2cc1

[a] The SMILES strings shown as "not valid" are valid *per se*, however, the interpretation of these codes are different between Daylight and Accelrys chemistry engines, the latter interpreting these as the hydrogen-saturated compounds cyclopentane and 3,4-dihydrocoumarin, respectively.

Fragments

Fragments used in the **TerraQSAR** modules have been described in detail in several publications listed in the literature, especially in the works by Kaiser *et al.* An overview of basic fragment types considered is given in **Table 1**.

Table 1. Fragment types used in TerraQSAR

Fragment Types	Examples
Acidity fragment	C(=O)O, S(=O)(=O)O
Aliphatic ring fragment	C1CCCCC1, C1CCCC1
Aromatic ring fragment	c1ccccc1, c1ccccn1
Atom fragment	C, H, N, O
Bond fragment	C-C, C=C, C#C
Group fragment	C-O-H, C-O-C, O=C-O-C
Hydrophobicity fragment	C(C)(C)C, CCCC
Ionization fragment	[O-], [Na+]
Polarity fragment	O=N(=O)CC(O)
Reactivity fragment	C=CC=O
Stereo fragment	Cl[C@H](C)N, Cl[C@@H](C)N
Weight fragment	molecular weight

Computation

The computer evaluates the number and types of bonds and fragments present in the compound and computes the toxicity estimate on the basis of the same types of bonds and fragments present in the training data set.

Computation time varies with the complexity of the query structure and speed of the computer. Typically, for compounds without chiral centers, and molecular weights of <200, computation time on a 2 GHz machine takes <5 seconds.

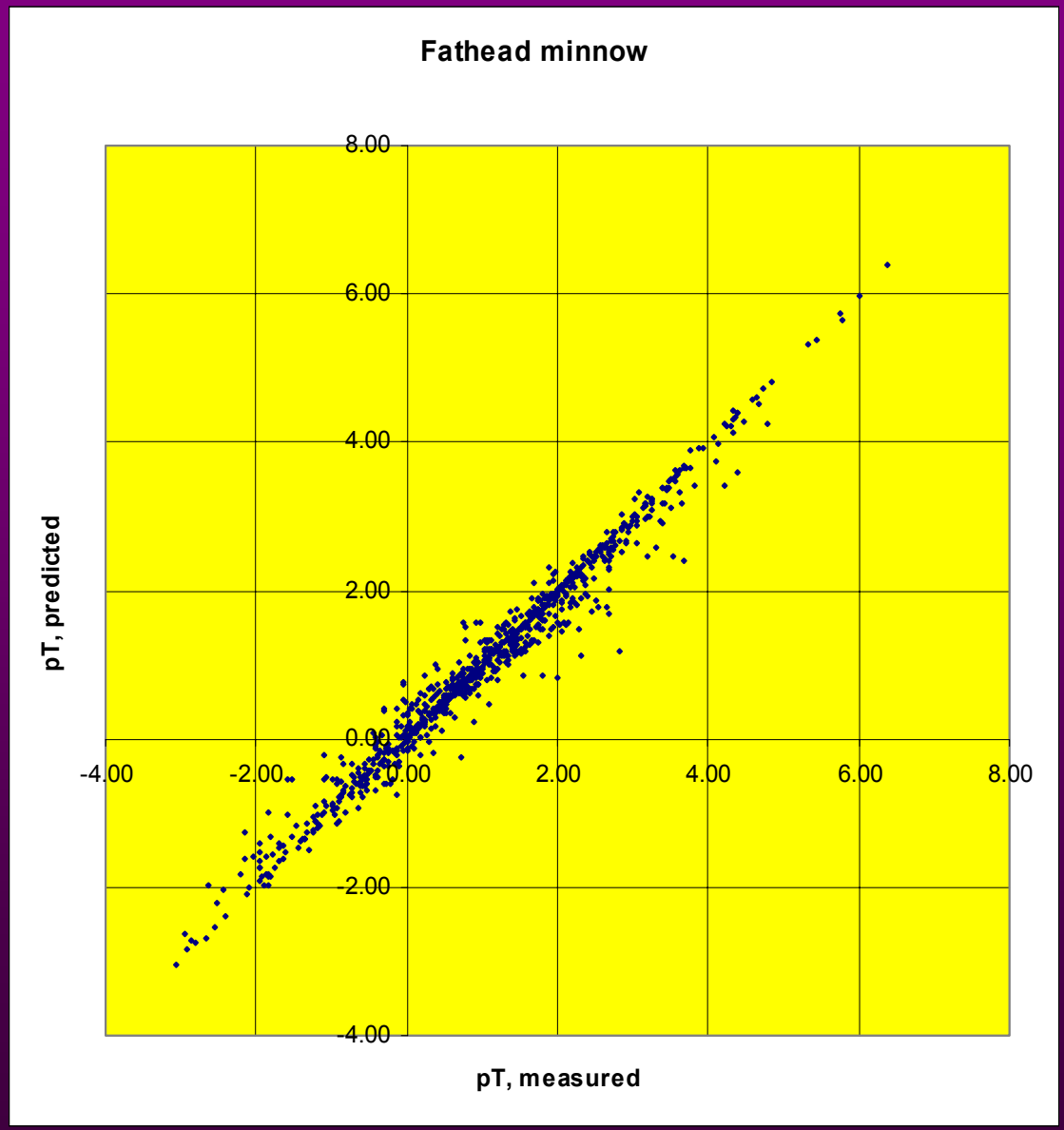
Results

Figure 1 shows the measured vs. predicted values for all 886 compounds fathead minnow values used in the development of the **TerraQSAR - FHM** estimation program, as obtained from the program.

The data cover approximately ten orders of magnitude, ranging from pT= -3 to pT= 7, where pT is the negative logarithm of the millimolar LC50 concentration, pT = log (L/mmol).

Figure 1

FHM



TerraBase Inc.

Table 3. Statistcal indicators for TerraQSAR - FHM

Program Module	Indicator	Variable
FHM	Number of compounds	886
	Range (log units)	10.5
	Correl. coeff. (r^2)	0.979
	Slope	1.064
	Intercept	-0.061
	RMS* error (leave-out 33%)	0.192
	RMS error (full training set)	0.063

* RMS: root mean square error

TerraBase Inc.

Figure 2

RMIV

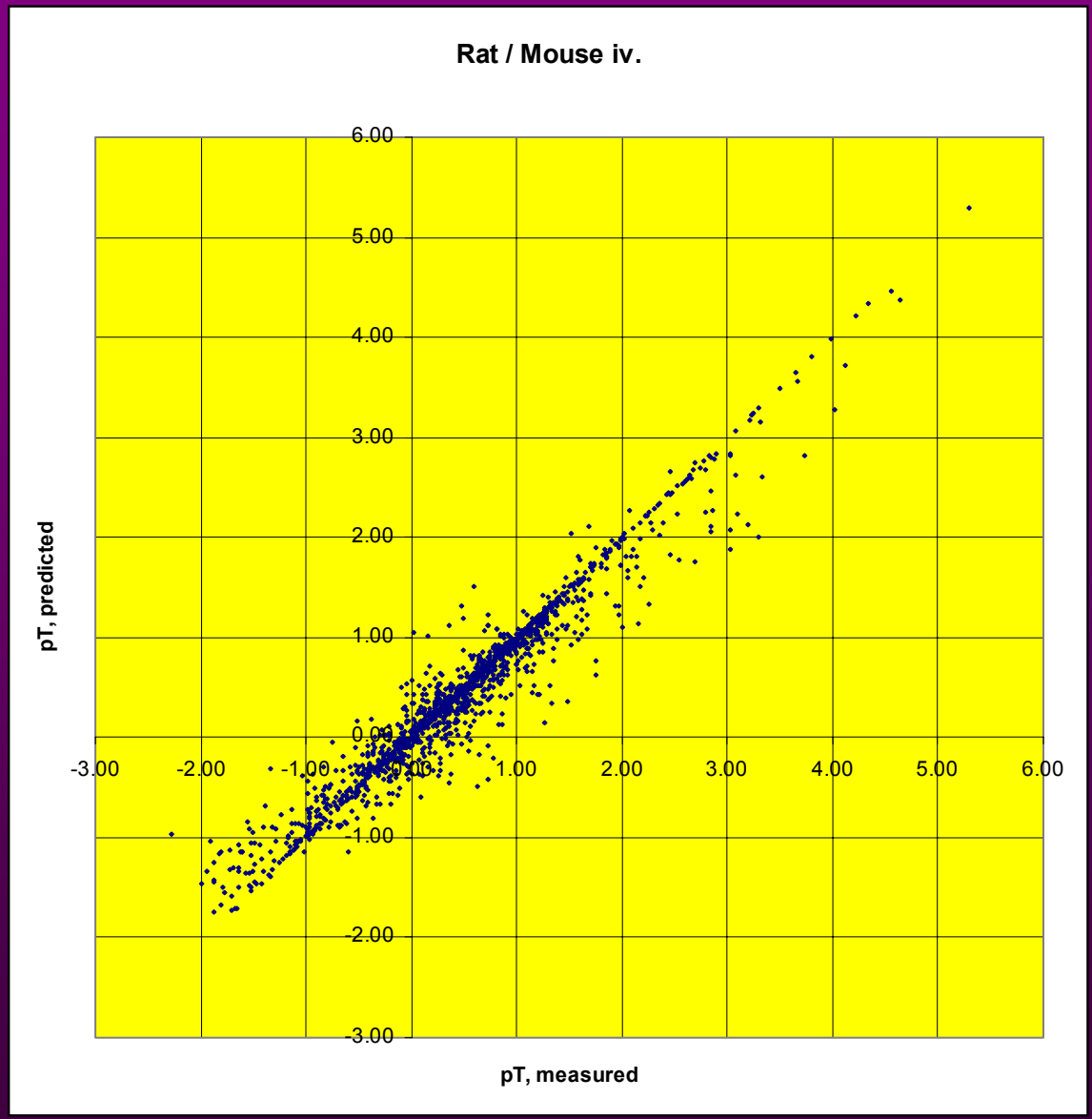


Table 4. Statistical indicators for TerraQSAR - RMIV

Program Module	Indicator	Variable
RMIV	Number of compounds	1497
	Range (log units)	7.5
	Correl. coeff. (r^2)	0.969
	Slope	1.057
	Intercept	0.005
	RMS* error (leave-out 33%)	0.199
	RMS error (full training set)	0.076

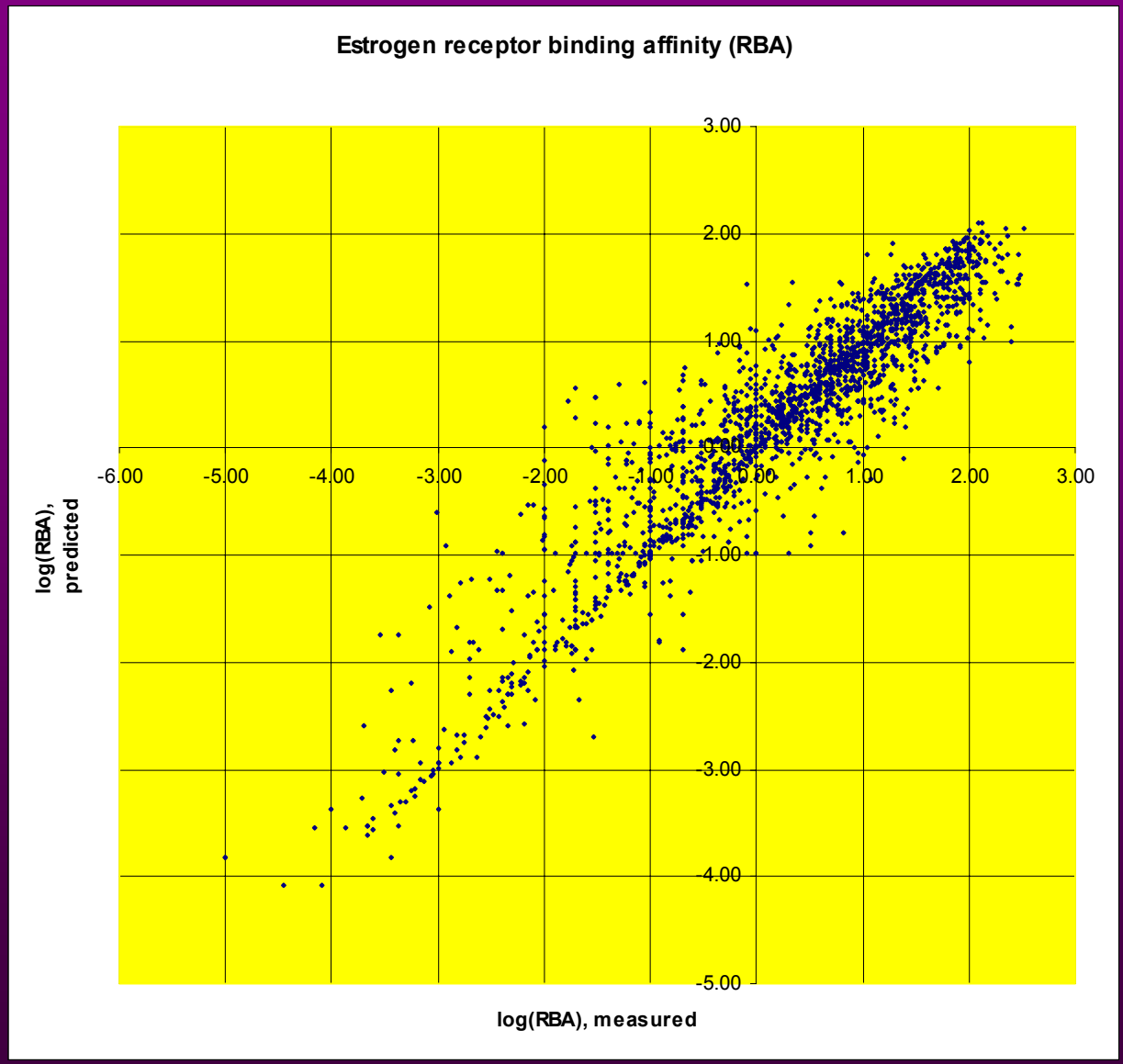
* RMS: root mean square error

Note regarding the E2-RBA data

The measured data available for this program (>2000 compounds) show a non-normal distribution with more than twice the numbers of compounds in the log(RBA) range 0 to +3 range than in the -4 to 0 range.

Figure 3

E2-RBA



TerraBase Inc.

Table 5. Statistical indicators for TerraQSAR – E2-RBA

Program Module	Indicator	Variable
E2-RBA	Number of compounds	> 2000
	Range (log units)	7.5
	Correl. coeff. (r^2)	0.930
	Slope	1.082
	Intercept	-0.068
	RMS* error (leave-out 33%)	0.250
	RMS error (full training set)	0.120

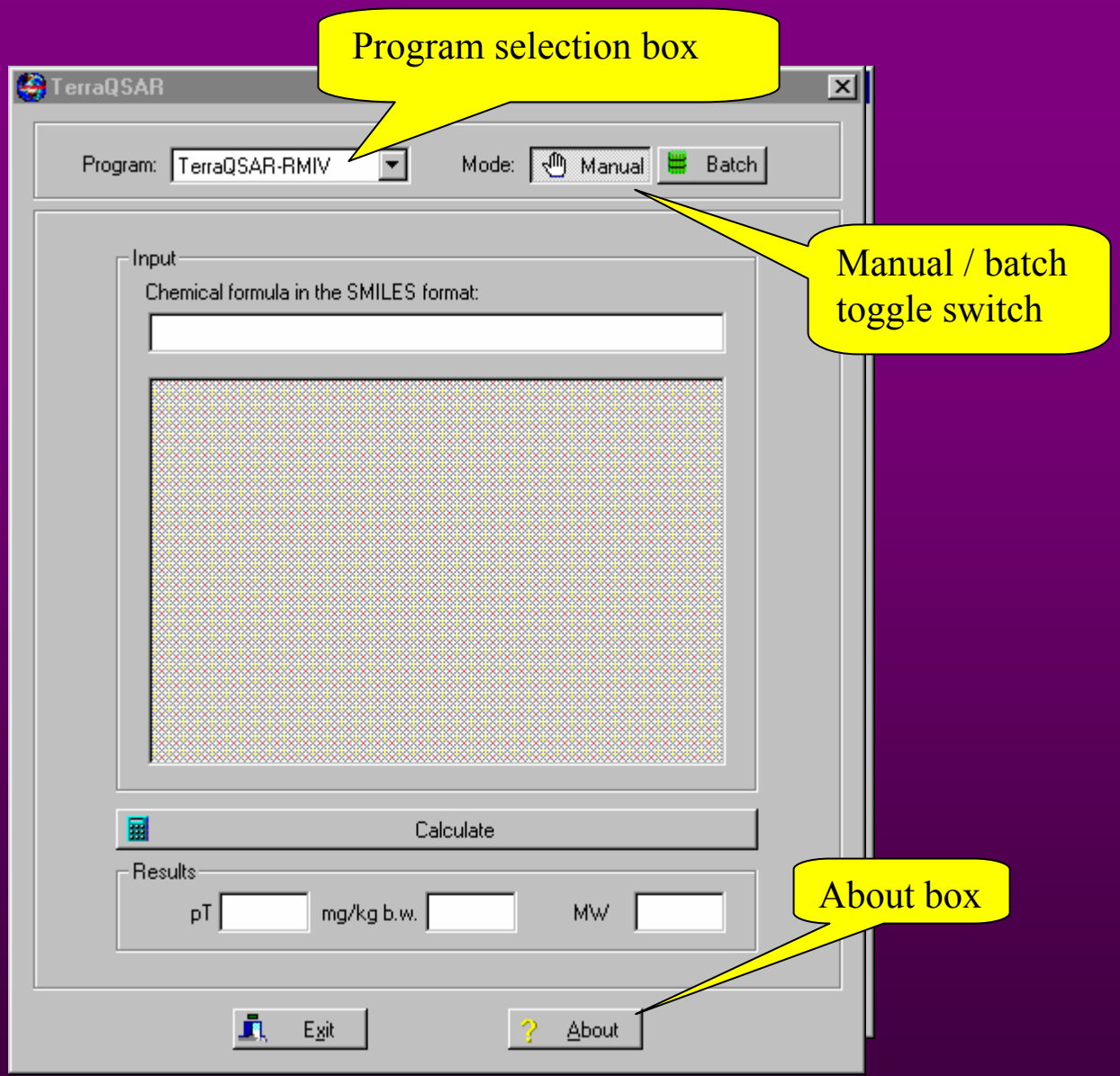
* RMS: root mean square error

TerraBase Inc.

Program Interface

The program interface of the **TerraQSAR** toxicity prediction modules is shown in the next slide.

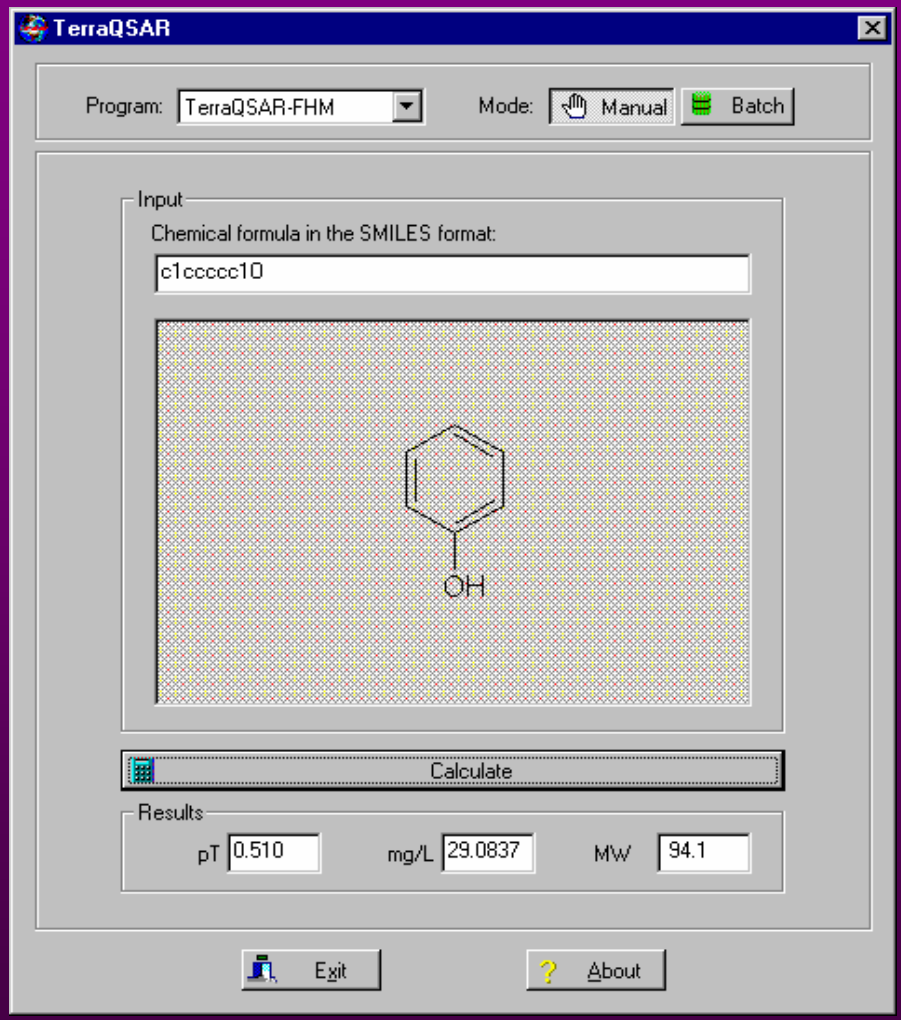
It is simple,
intuitive, and
highly functional.



TerraBase Inc.

Example 1:

Phenol has the SMILES string "c1ccccc1O". Copying this string into memory, for example from this text (making sure the quotation marks are omitted), and pasting it into the Input field will result in the appearance of the chemical structure of phenol in the shaded, rectangular field below, as shown in the next slide.



TerraBase Inc.

Once the user has ascertained that the structure of the compound is that of the desired chemical, a simple click of the Calculate bar below the structure field will result in the three fields below the bar to be filled with the predicted values for the compound, as shown in the next slide.

Field 1 (pT) is the - log LC50 [mmol/L];

Field 2 (mg/L) is the LC50 in mg/L (FHM module);

Field 3 (MW) shows the molecular weight of the compound.

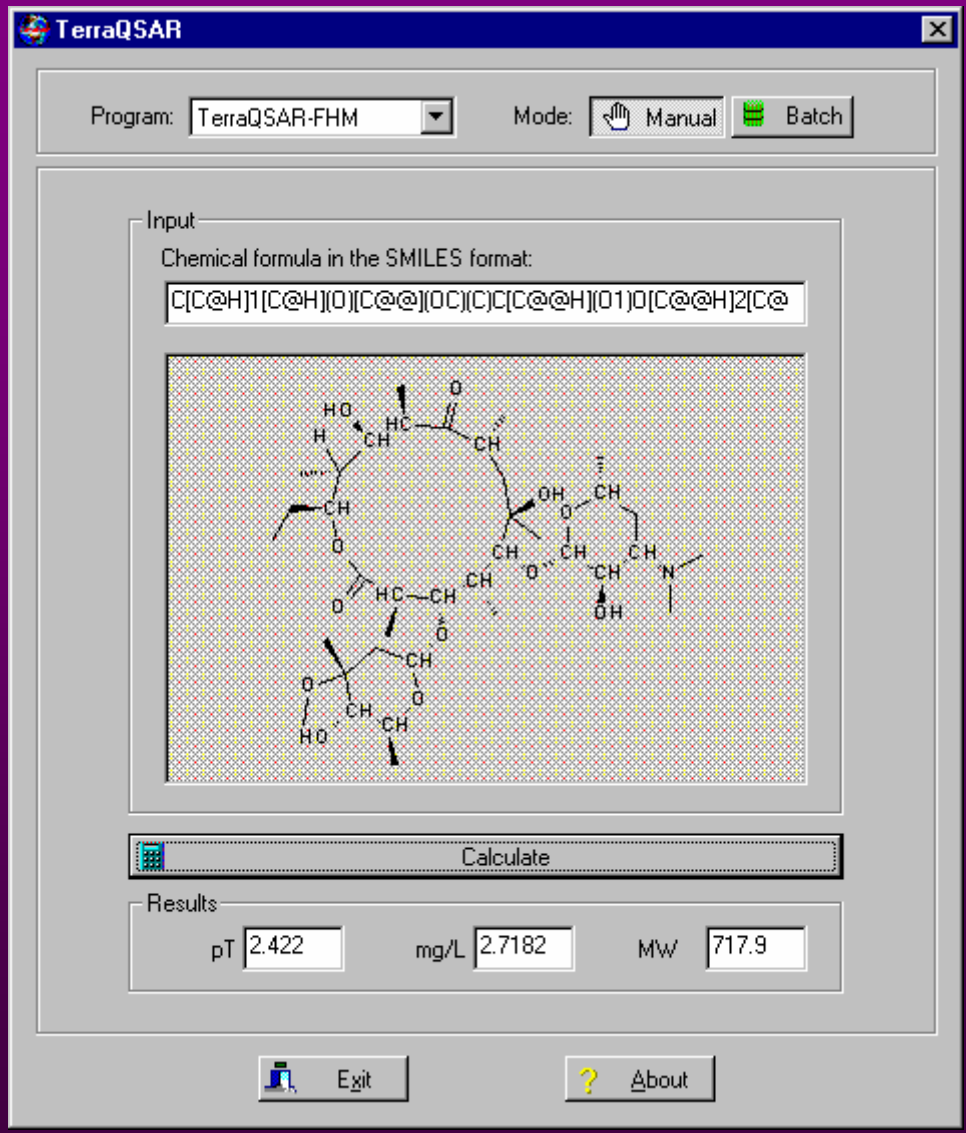
Example 1 (continued): phenol “c1cccc1O”

The fields show the following values:

- field 1 (pT) is the LC50 value in log(L/mmol): 0.51
- field 2 (mg/L) is the LC50 value in mg/L: 29.08
- field 3 shows the mol. weight of the compound: 94.1

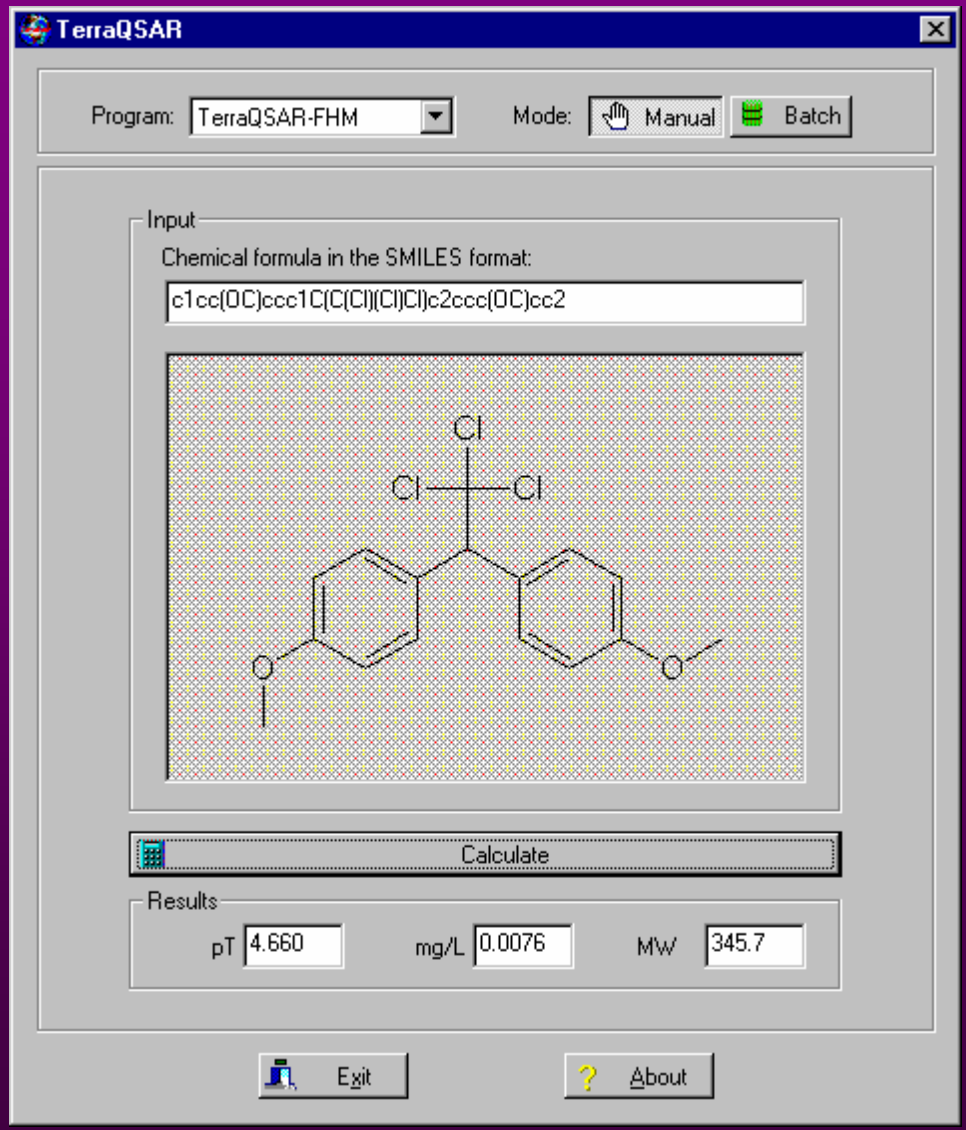
Computation time ~ 2 sec at 2 GHz

FHM:
Erythromycin



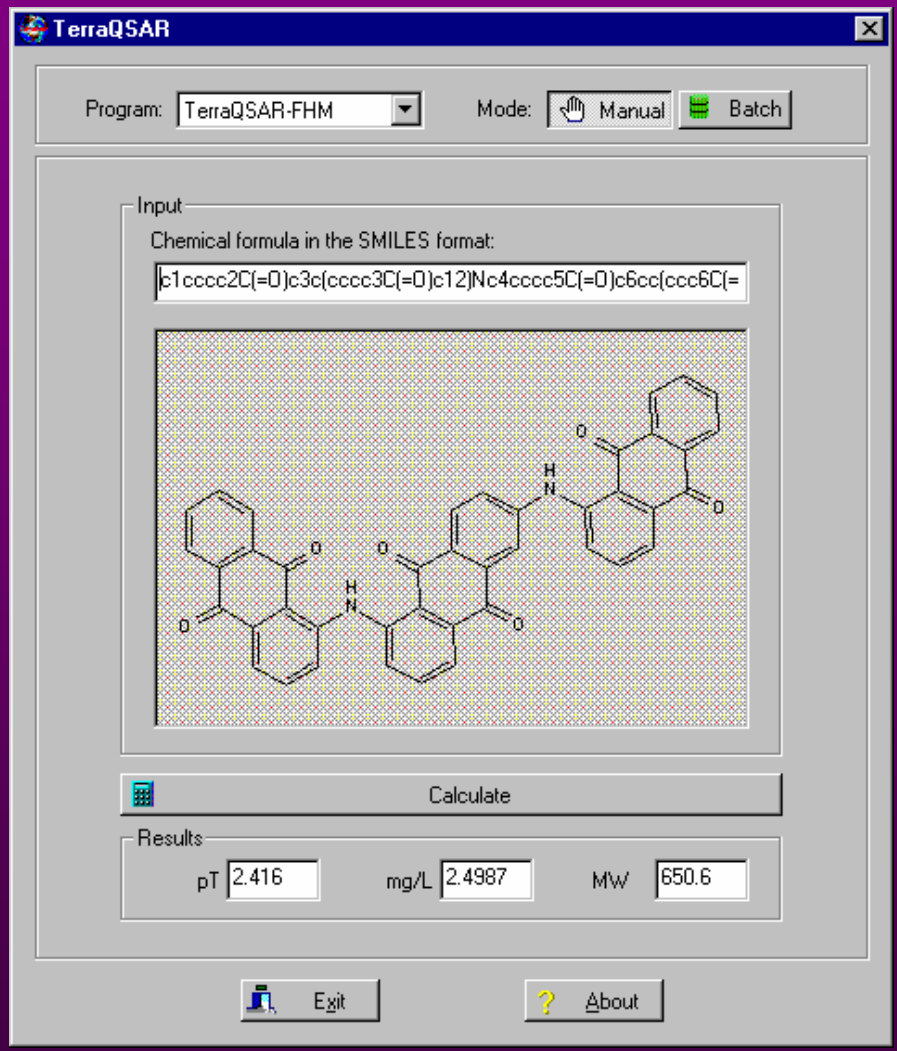
TerraBase Inc.

FHM:
Methoxychlor



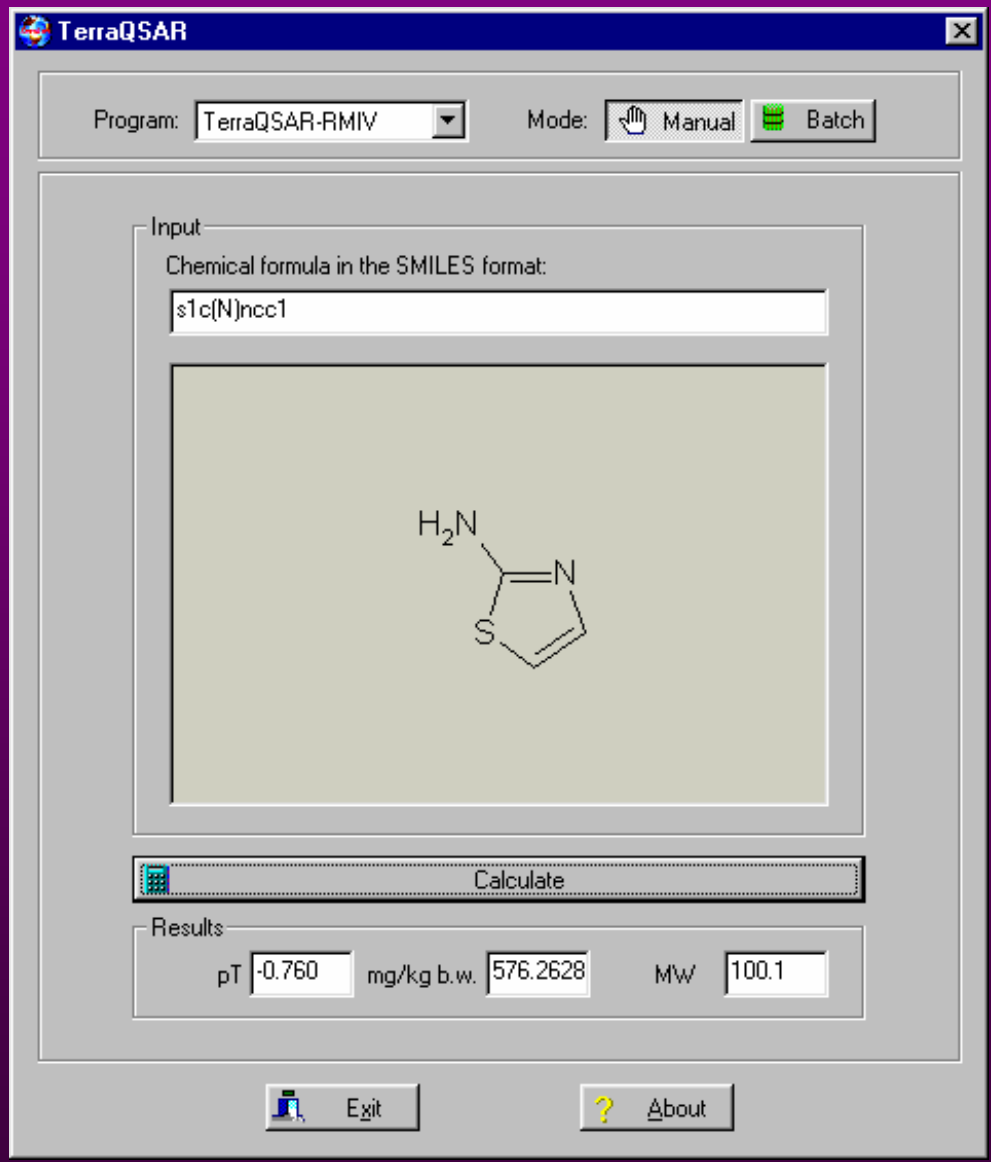
TerraBase Inc.

FHM:
anthraquinone dye



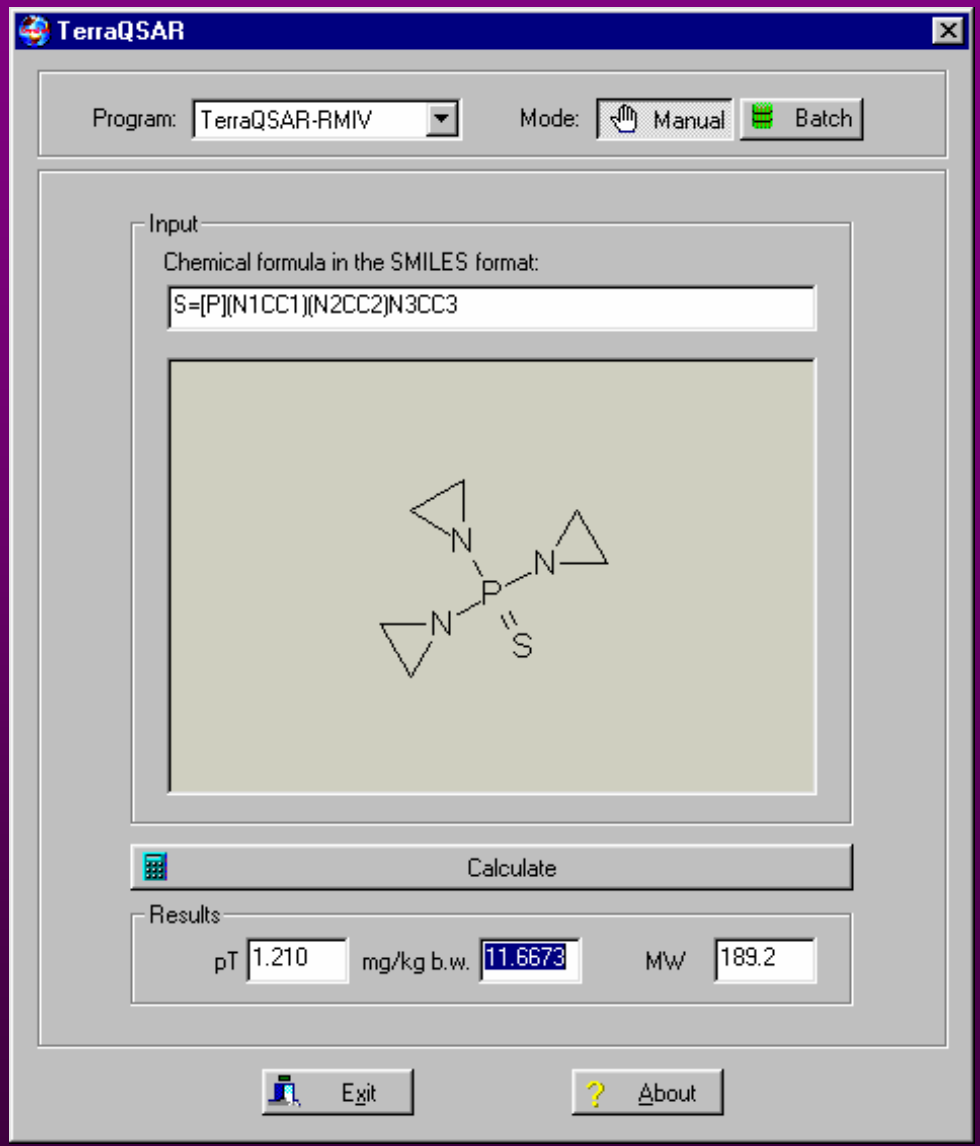
TerraBase Inc.

RMIV:
2-aminothiazole



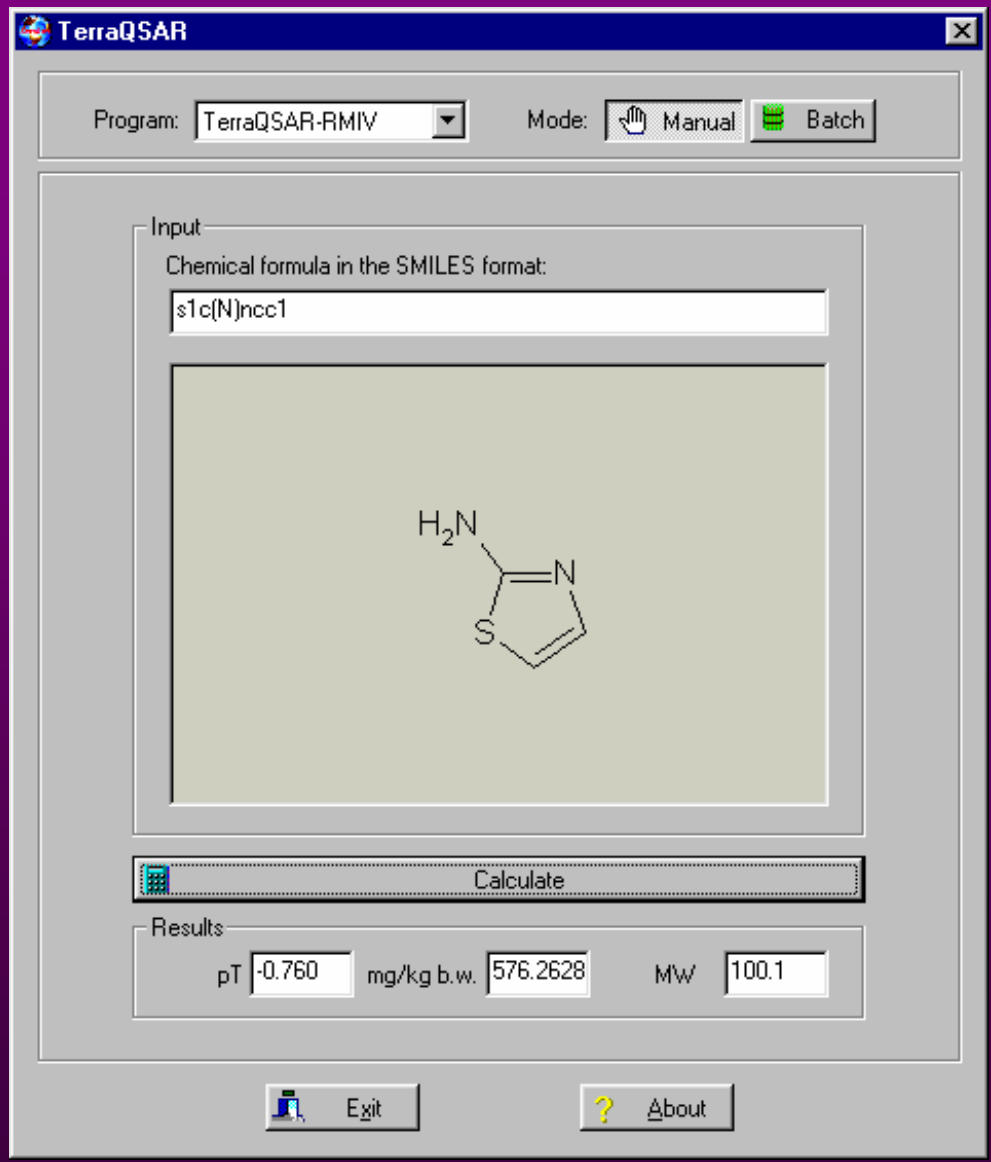
TerraBase Inc.

RMIV:
2-aminothiazole



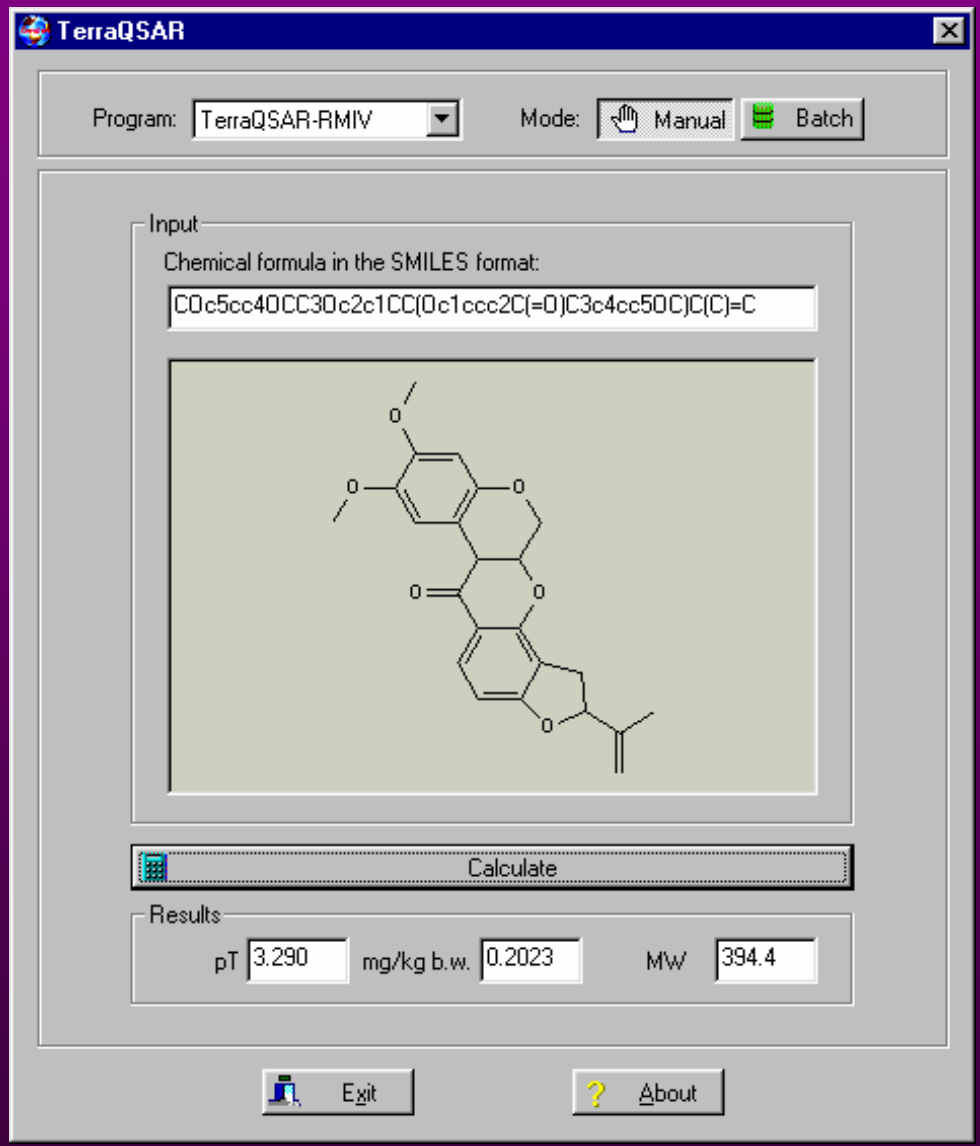
TerraBase Inc.

RMIV:
2-aminothiazole



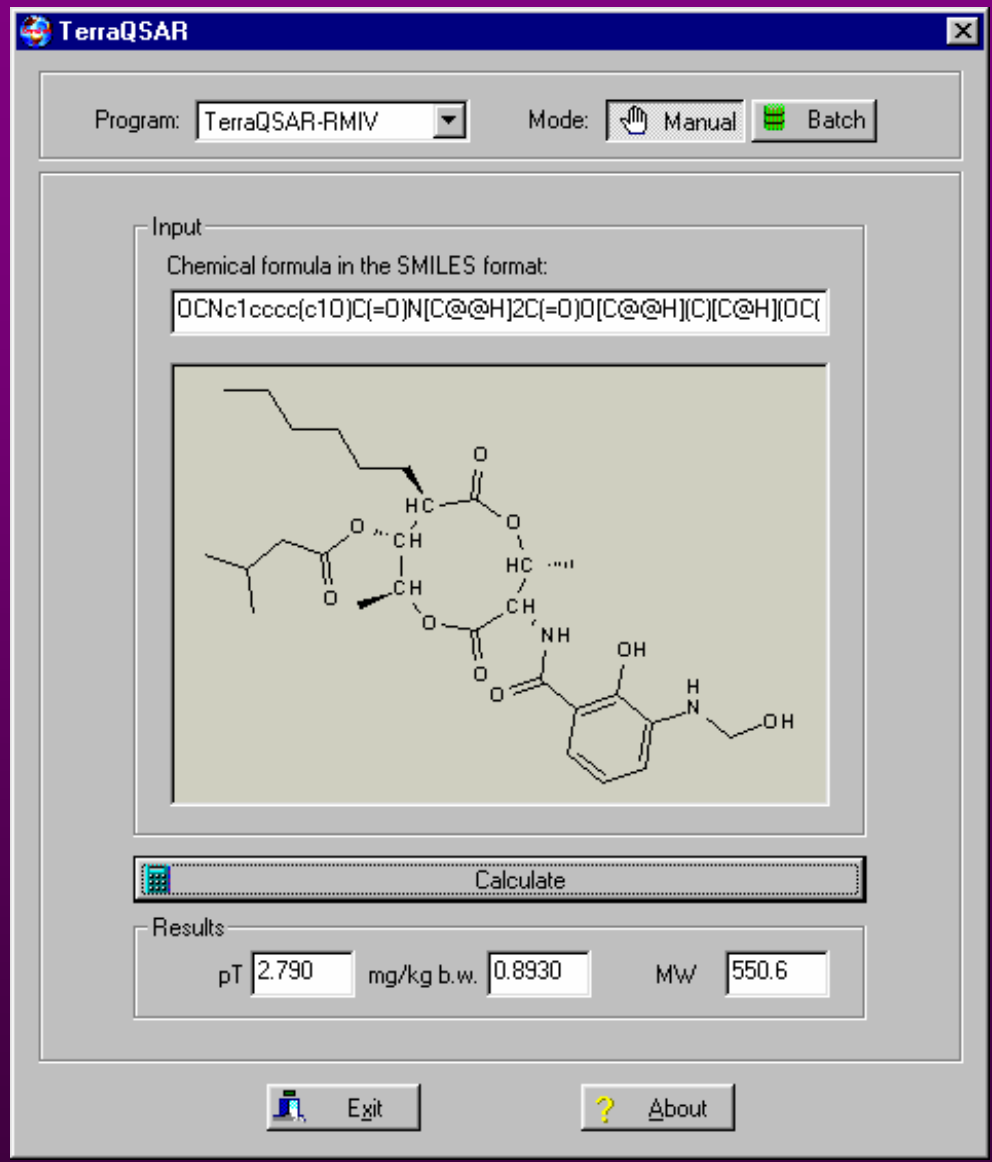
TerraBase Inc.

RMIV:
Rotenone



TerraBase Inc.

RMIV:
Antimycin A



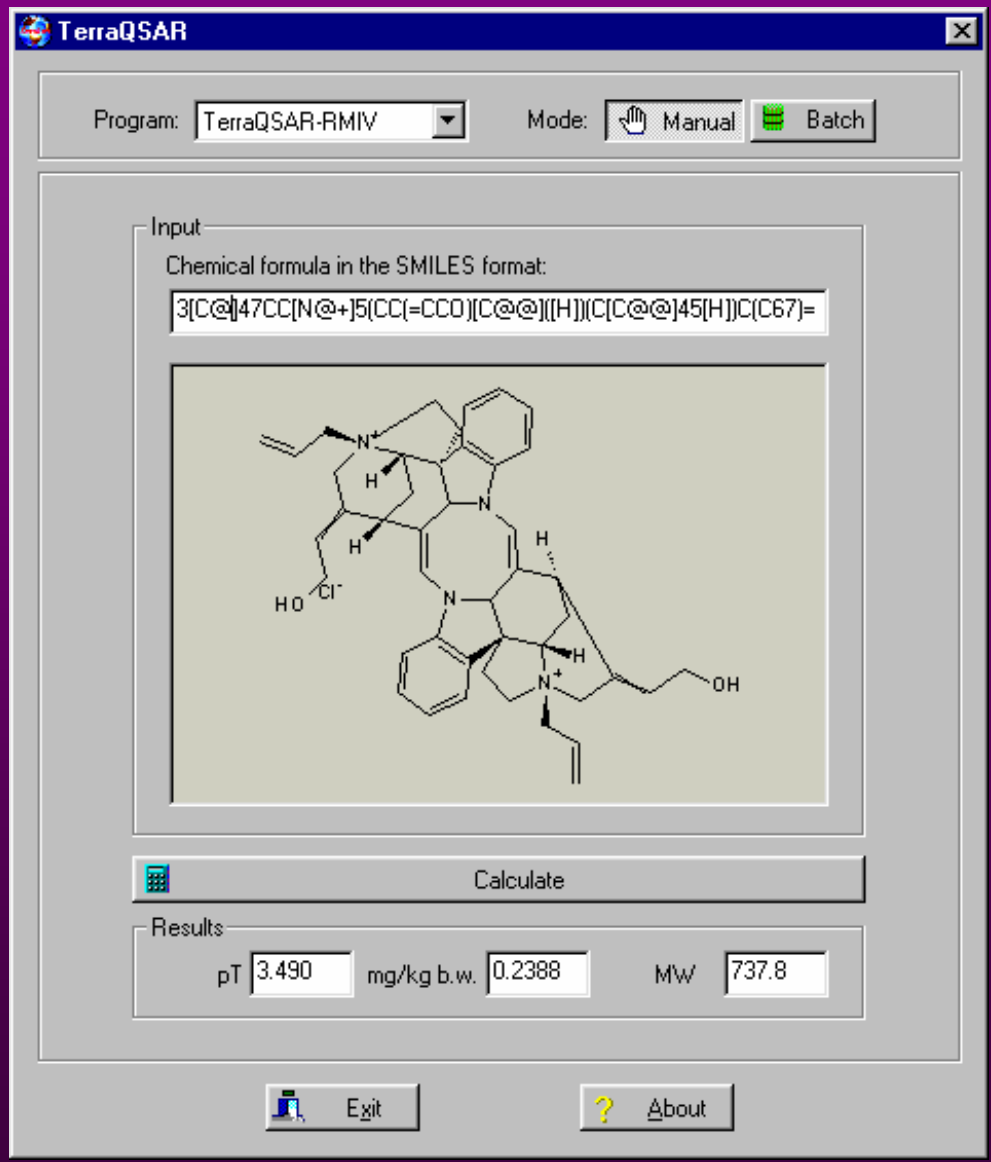
TerraBase Inc.

Examples 5-8:

The following examples of more complex molecules demonstrate the versatility of the TerraQSAR - FHM program:

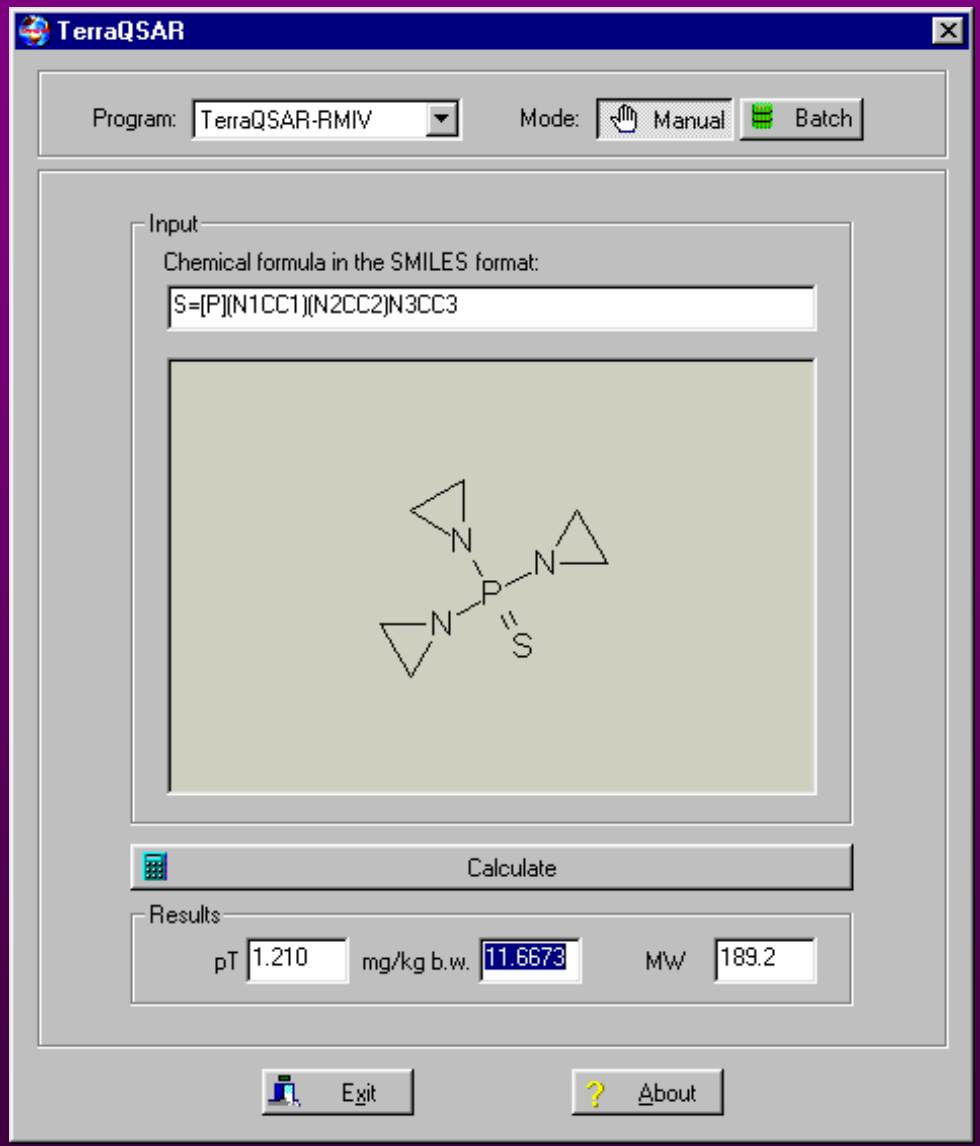
- methoxychlor (pesticide),
- hexadecylphenol deriv. (surfactant),
- synthetic steroid with chiral centres (drug),
- diazo compound sulfonic acid salt (azo-dye).

RMIV: Alcuronium dichloride



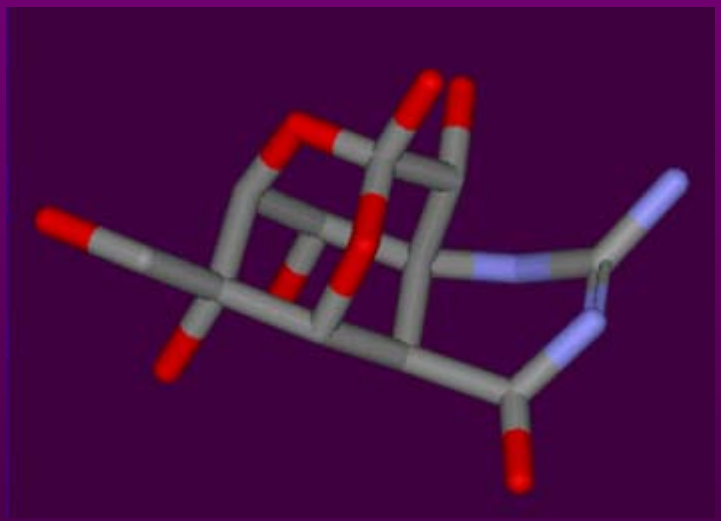
TerraBase Inc.

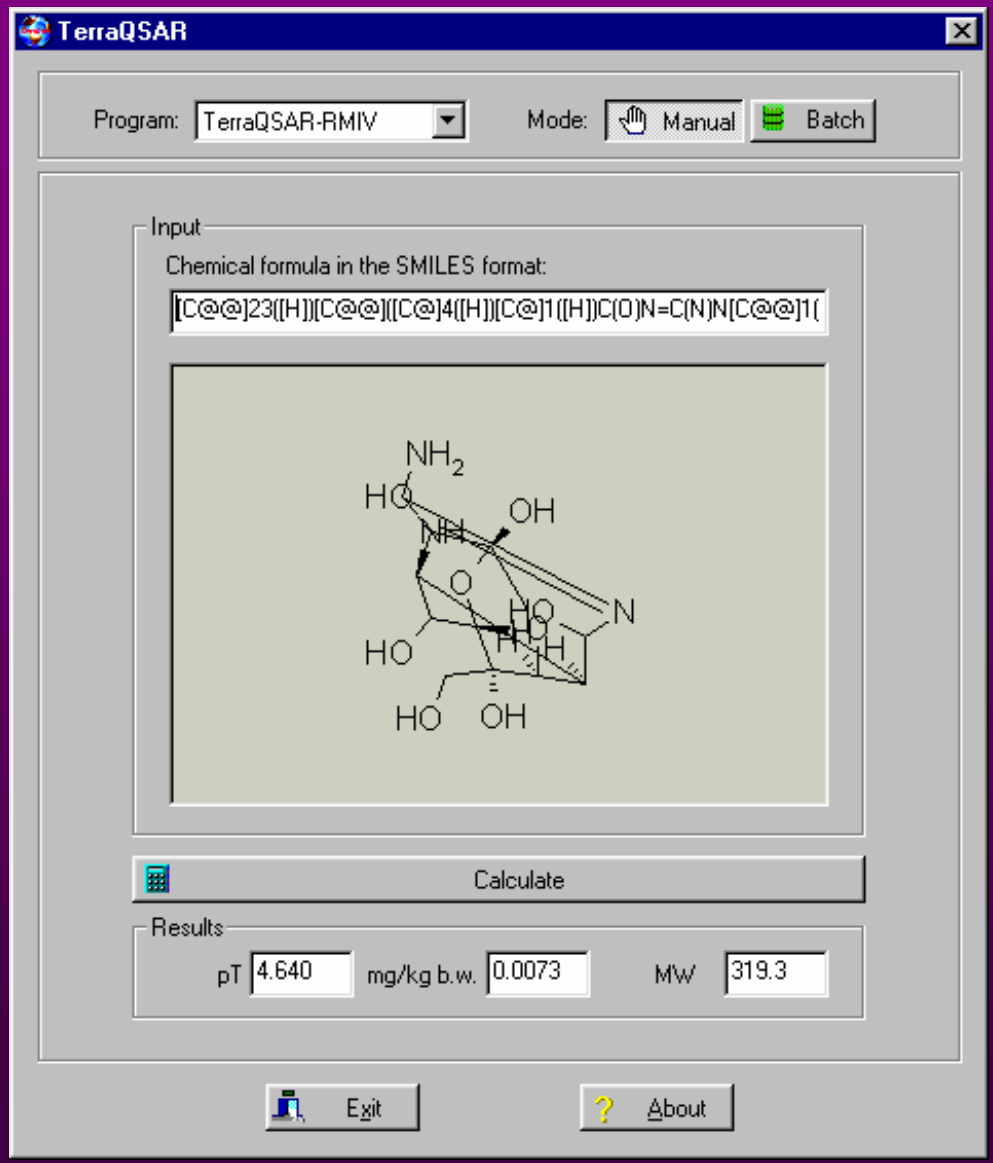
RMIV: 2-aminothiazole



TerraBase Inc.

RMIV:
Tetrodotoxin





TerraBase Inc.

Limitations and Rules

1. Metal ions

TerraQSAR considers all metal ions equal, that is as non-toxic components of the substance. This is correct for sodium, potassium, calcium, etc. salts, but incorrect for salts of copper, cadmium, etc., The latter have their own (metal ion) toxicity, in addition to the toxicity of the organic moiety.

Therefore, the toxicity of salts of organic anions with toxic metal ions will be underestimated by the toxicity contribution of the metal ion and corrections will have to applied for such.

Limitations and Rules

2. Organo-metal compounds with covalent bonds

TerraQSAR computes the correct values for compounds with carbon-metal covalent bonds, such as tin and silicon organo-metal compounds (e.g., diethyl-dibutyltin, CCCC[Sn](CC)(CC)CCC) with high accuracy.

Therefore, no corrections needs to be applied for these types of substances.

Limitations and Rules

3. Organo-metal compounds with pi bonds

Due to limitations of the Accelrys software in correctly representing organo-metal pi bond complexes at this time (such as in ferrocene), TerraQSAR cannot compute values for such compounds.

Limitations and Rules

4. Tautomers

TerraQSAR will compute the correct value for each tautomer, but does not average or proportion their relative abundance.

Therefore, the user must decide which tautomer is the most important one for his purpose.

Example: 2-pyridone (FHM module)

keto-tautomer: pT = -0.11

enol-tautomer: pT = -0.61

Technical Requirements

Operating system:

PC with W95, 98, NT, 2000, or ME OS operating systems; ready for Windows XP, subject to Accelrys software update for XP.

Central processor unit (CPU):

No specific requirement, speed of computations will increase with decrease in CPU speed; 2 GHz or higher recommended.

Mouse, or other pointing device: required.

Screen setting: Variable, 800 x 640, or higher.

Other: CD-ROM drive.

Literature – Neural Network Theory

Grabec, I. Self-organization of neurons described by the maximum-entropy principle. **Biol. Cybern., 63:** 403-409 (1990).

Grabec, I. Optimization of kernel-type density estimator by the principle of maximal self-consistency. **Neural Parallel & Scientific Computations, 1:** 83-92 (1993).

Hecht-Nielsen, R. Nearest matched filter classification of spatiotemporal patterns. **Appl. Optics, 26:** 1892-1899 (1987).

Hecht-Nielsen, R. **Neurocomputing**. Addison-Wesley Publishing Co., Inc. (1990).

Hertz, J., A. Rogh, and R. Palmer. **Introduction to the Theory of Neural Computation**. Addison-Wesley, Redwood City, California (1991). Prechelt, L. (Ed.). Frequently asked questions (FAQ) on neural networks. (1995); http://www.ipd.uka.de/~prechelt/FAQ/nn7.html.

Ramirez, M.R., and D. Arghya. A faster learning algorithm for back-propagation neural networks in NDE applications. **Proc. 2nd Int. Conf. on AI**, pp. 275-283 (1991).

Samaad, T. Backpropagation improvements based on heuristic arguments, theory track, neural and cognitive sciences. **Track. Int. Joint Conf. Neural Networks**, Vol. 1, Washington, D.C., (1990).

Shanno, D.F. Conjugate gradient methods with inexact searches. **Math. Oper. Res., 3:** 244-256 (1978).

Specht, D.F. Probabilistic neural networks for classification, mapping or associative memory. **ICNN, Conference Proc.** (1988).

Yeh, Y., Y. Kuo, and D. Hsu**.** Building KBES for diagnosis PC pile with artificial neural network. **J. Comp. Civil Eng., 7:** 71-93 (1993).

Literature – Neural Network Applications

Kaiser, K.L.E. Neural networks for effect prediction in environmental and health issues using large datasets. **QSAR Comb. Sci., 22:** 185-190 (2003); http://www.wiley-vch.de:80/contents/jc_2022/200302.html .

Kaiser, K.L.E. The use of neural networks in QSARs for acute aquatic toxicological endpoints. **J. Mol. Struc. (Theochem), 622:** 85-95 (2003); http://dx.doi.org/10.1016/S0166-1280(02)00620-6 .

Kaiser, K.L.E., S.P. Niculescu, and G. Schüürmann. Feed forward backpropagation neural networks and their use in predicting the acute toxicity of chemicals to the fathead minnow. **Water Quality Res. J. Canada, 32**: 637-657 (1997); http://www.cciw.ca/wqrjc/32-3/32-3-637.htm .

Kaiser, K.L.E., S.P. Niculescu, and K.M. Gough. Neural network modeling of *Vibrio fischeri* and fathead minnow acute toxicity data with molecular indicator variables and physico-chemical bulk parameters. Poster, **Workshop on Computational Methods in Toxicology**, Dayton, OH, April 20-22, (1998), http://www.ccl.net/ccl/toxicology/abstracts/abs14.html .

Kaiser, K.L.E., and S.P. Niculescu. Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales promelas*): A study based on 865 compounds. **Chemosphere**, **38:** 3237-3245 (1999). http://dx.doi.org/10.1016/S0045-6535(99)00553-6 .

Kaiser, K.L.E., S.P. Niculescu, and T.W. Schultz. Probabilistic neural network modeling of the toxicity of chemicals to *Tetrahymena pyriformis* with molecular fragment descriptors. **SAR & QSAR Envir. Res., 13:** 57-67 (2002); http://taylorandfrancis.metapress.com/link.asp?id=ak11jrbmqqkrnp76 .

Kaiser, K.L.E., and S.P. Niculescu. On the PNN modeling of estrogen receptor binding data for carboxylic acid esters and organochlorine compounds. **Water Qual. Res. J. Canada**, **36:** 619-630 (2001); http://www.cciw.ca/wqrjc/36-3/36-3-619.htm .

Literature – SMILES code

Daylight Chemical Information Systems, Inc. (2002);
http://www.daylight.com/smiles/f_smiles.html

TerraBase Inc.

1063 King St. West, Suite 130
Hamilton, ON, L8S 4S3, Canada
Phone: 905-802-0154
Fax: 905-527-0263
E-mail: mail@terrabase-inc.com

Internet: http://www.terrabase-inc.com/

Future developments:

TerraQSAR modules for other endpoints:

HIV-1

BIODEG

logP

TerraBase Inc. profile

Founded in 1996, **TerraBase Inc.**, a Canadian company, produces specialty toxicity estimation programs and unique databases for scientific research in the health, pharmaceutical, pesticide, environmental risk assessment and waste management fields. TerraBase Inc. also provides toxicological and physico-chemical property estimation, data retrieval and substance classification services for scientific research and industry.

TerraBase Inc. is a leader in desktop computer software for structure-activity, structure-property and inter-species relationships. Our TerraTox™ and TerraQSAR™ products are unrivalled. They provide provide fast access to toxicity and physico-chemical data for thousands of chemicals of common interest, including pesticides, drugs, dyes, chemical intermediates, and their metabolites.

TerraBase Inc. products are used by government agencies, universities, and leading industrial R&D departments in America, Europe, Asia.