Overview of T.E.S.T.
(Toxicity Estimation Software Tool)

Goal

- Develop user friendly software that can estimate toxicity and physical properties from molecular structure
  - For applications such as hazard comparison or alternatives assessment
  - Can screen hypothetical/new chemicals and faster and cheaper than conducting experiments

OECD* Principles for QSAR Models

- An unambiguous algorithm (QSAR methods)
- A defined endpoint (what is modeled)
- A defined domain of applicability (when to trust predictions)
- Appropriate measures of goodness-of fit, robustness and predictivity (training/test set statistics)
- A mechanistic interpretation, if possible (analysis of descriptors appearing in the models)

*Organisation for Economic Co-operation and Development

Model variables

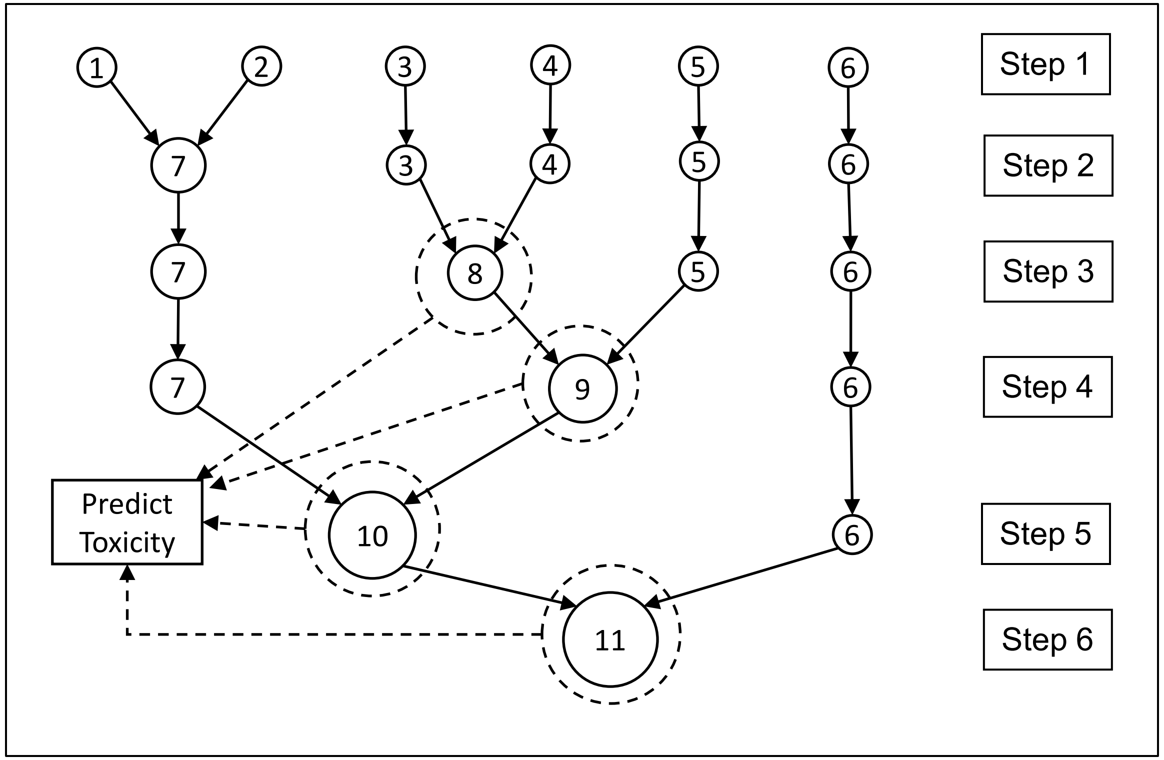
- T.E.S.T. Descriptors are used for model building
  - Combination of whole molecule descriptors (continuous) and molecular fragment counts (integer)
  - Descriptors do not use x-y-z coordinates (3d descriptors omitted)
- Classes of descriptors:
  - E-state
  - Constitutional descriptors
  - Topological descriptors
  - Walk and path counts
  - Connectivity
  - Information content
  - 2d autocorrelation
  - Burden eigenvalue
  - Molecular property
  - H bond acceptor/donor
  - Molecular distance edge
  - Molecular fragment counts

QSAR Methods

- QSAR methods:
  - Hierarchical clustering
  - Single Model
  - Group contribution
  - Nearest neighbor
  - Consensus
- See the TEST User's guide for more information

Hierarchical clustering

- Similar chemicals are grouped using Ward’s method
- Uses information from entire data set



Prediction = weighted average of best model from each step:

$$Tox = \sum_{i=1}^{k} w_i \times Tox_i \Big/ \sum_{i=1}^{k} w_i \qquad Tox_i = \sum_{i=1}^{\#descriptors} a_i x_i + a_0$$

Single model

- Predictions is made using multilinear regression model fit to entire training set:

$$Tox = \sum a_i x_i + a_0$$

- Descriptors, x_i, are 2d molecular descriptors
- Example: 48 hr *Daphnia magna* LC_{50} model
  - Toxicity = 1.2157×(xc4) + 0.1341×(StN) + 0.6974×(SsSH) - 1.3213×(SsOH_acnt) + 0.8605×(Hmax) + 1.4685×(ssi) - 0.9197×(MDEN33) + 0.2238×(BEHm1) + 1.4502×(BEHp1) + 2.4060×(Mv) + 1.9085×(MATS1m) - 2.4036×(MATS1e) - 0.3463×(GATS3m) + 0.0255×(AMR) - 1.4215×(-C(=S)- [2 nitrogen attach]) - 0.7185×(AN) - 1.0232×(-N< [attached to P]) - 1.5228×(-S(=O)(=O)- [aromatic attach]) - 6.5594

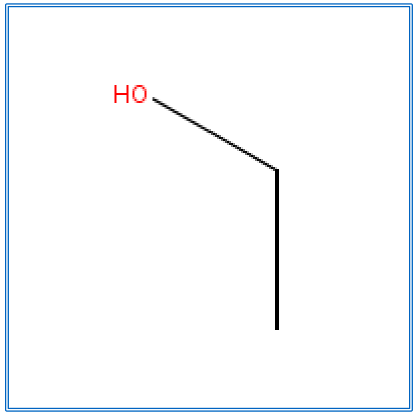
Group contribution

- Predictions is made using multilinear regression model fit to entire training set:

$$Tox = \sum a_i x_i + a_0$$

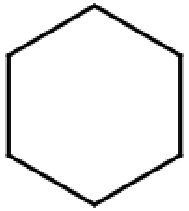
- Descriptors, x_i, are molecular fragment counts

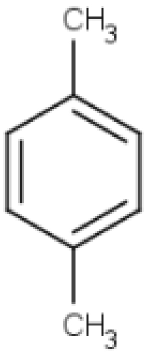
Descriptor	x_i	a_i	$a_i \times x_i$
-CH3 [aliphatic attach]	1	0.23	0.23
-CH2- [aliphatic attach]	1	0.27	0.27
-OH [aliphatic attach]	1	-0.58	-0.58
Model intercept (a_0)	1	1.96	1.96
Tox (-Log10(LC_{50} mol/L))			1.88

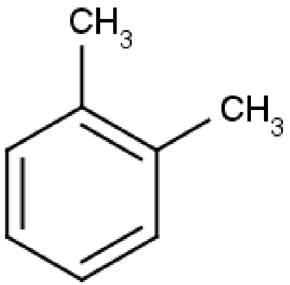


Nearest Neighbor

- Predicted toxicity is simply the average of the three nearest neighbors (i.e. analogous to read-across)
- All neighbors must exceed a minimum similarity
- For example, the toxicity for benzene is obtained by averaging the experimental values for the following analogs:







Consensus model

- The consensus prediction is simply the average predicted value for all the models that have predictions inside their applicability domain
- A prediction is made if at least two models have a valid prediction in terms of their respective applicability domain
- Using multiple models minimizes bad predictions and maximizes prediction accuracy
- Using different applicability domains maximizes prediction coverage
- Recommended method

Available endpoints

Toxicity endpoints

- 96 hour fathead minnow LC50
- 48 hour *D. magna* LC50
- 48 hour *T. pyriformis* IGC50
- Oral rat LD50
- Bioaccumulation factor
- Developmental toxicity
- Ames mutagenicity

Physchem properties

- Normal boiling point
- Vapor pressure
- Melting point
- Flash point
- Density
- Surface tension
- Thermal conductivity
- Viscosity
- Water solubility

Applicability Domain (AD)

AD measures for regression-based models in T.E.S.T.:

- Rmax (all descriptors)
  - Distance from the test chemical to the centroid is less than the maximum distance for any chemical to the centroid of the cluster
- Model ellipsoid (model descriptors)
  - Leverage of test compound must be less than leverage of all compounds included in the model
- Fragments constraint
  - Cluster must contain one example of each fragment in the test chemical

Mechanistic interpretation

- Descriptors in models can be examined *a posteriori* for mechanistic plausibility
  - LogP descriptors (ALOGP, XLOGP) show up in models for aquatic toxicity (narcosis mechanism)
  - Molecular fragment counts modulate toxicity (+/-)
  - Whole molecule descriptors are related to features such as molecular size, polarizability, or hydrophobicity

Test set statistics*

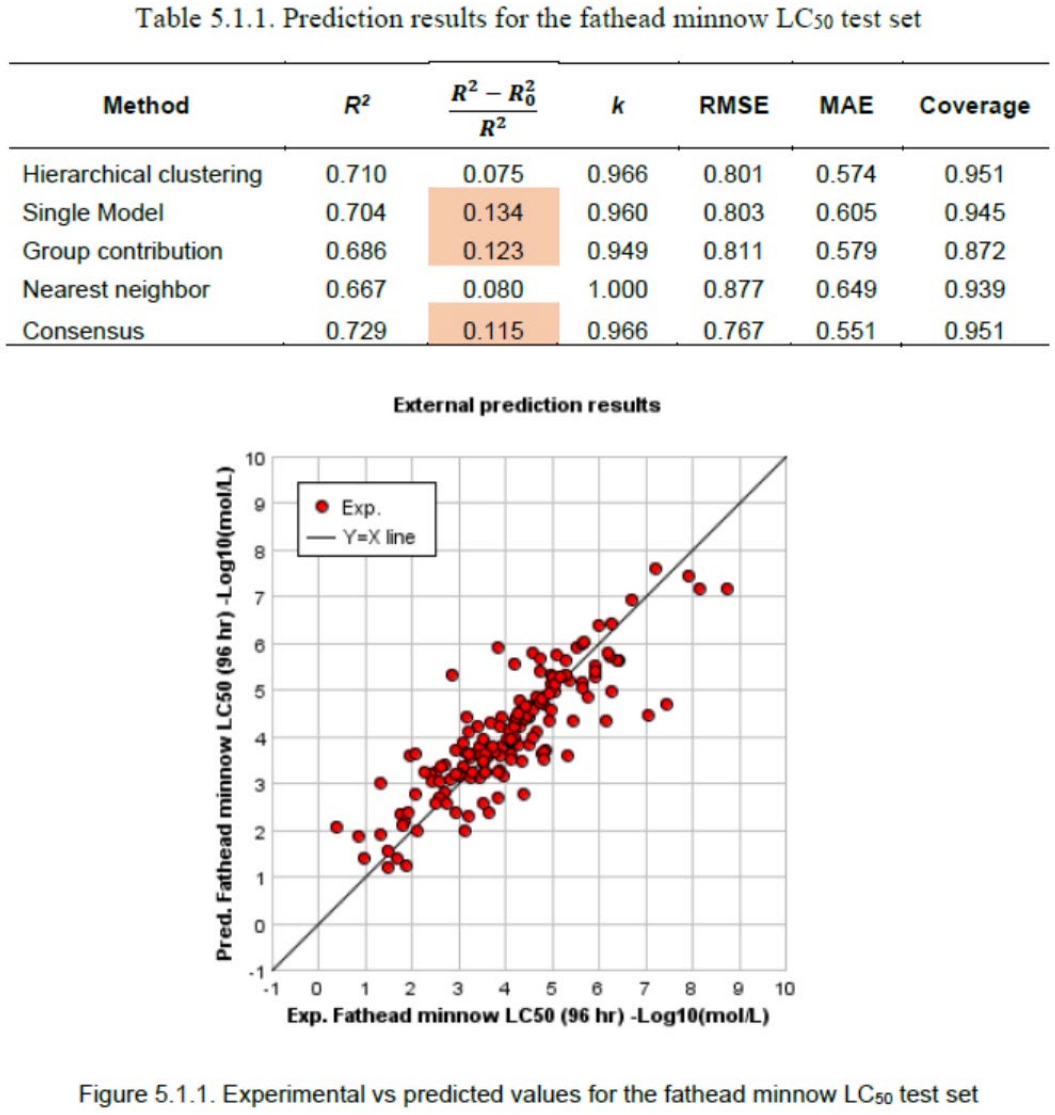


Table 5.1.1. Prediction results for the fathead minnow LC_{50} test set

Method	R^2	$\frac{R^2 - R_0^2}{R^2}$	k	RMSE	MAE	Coverage
Hierarchical clustering	0.710	0.075	0.966	0.801	0.574	0.951
Single Model	0.704	0.134	0.960	0.803	0.605	0.945
Group contribution	0.686	0.123	0.949	0.811	0.579	0.872
Nearest neighbor	0.667	0.080	1.000	0.877	0.649	0.939
Consensus	0.729	0.115	0.966	0.767	0.551	0.951

External prediction results

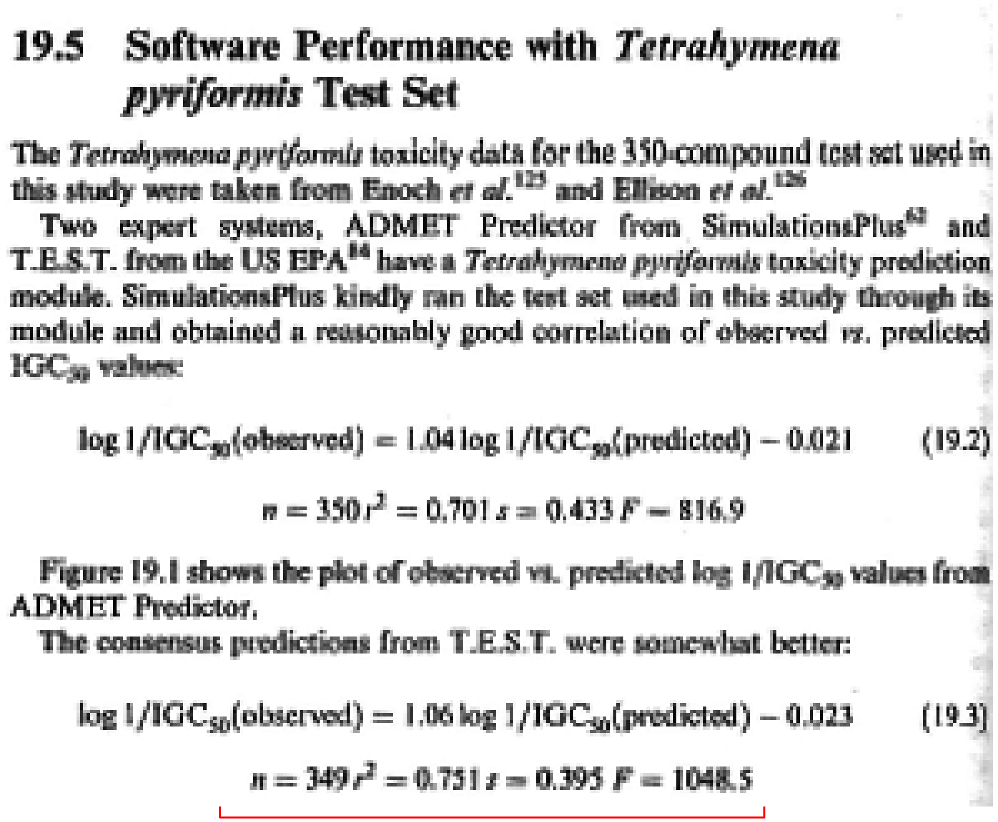
Exp.
Y=X line
Pred. Fathead minnow LC50 (96 hr) -Log10(mol/L)
Exp. Fathead minnow LC50 (96 hr) -Log10(mol/L)
-1
0
1
2
3
4
5
6
7
8
9
10

Figure 5.1.1. Experimental vs predicted values for the fathead minnow LC_{50} test set

* See T.E.S.T. User’s Guide, Chapter 5

Comparison to other tools

IGC$_{50}$ performance*



19.5 Software Performance with *Tetrahymena pyriformis* Test Set

The *Tetrahymena pyriformis* toxicity data for the 350-compound test set used in this study were taken from Enoch *et al.*[125] and Ellison *et al.*[126]

Two expert systems, ADMET Predictor from SimulationsPlus[83] and T.E.S.T. from the US EPA[84] have a *Tetrahymena pyriformis* toxicity prediction module. SimulationsPlus kindly ran the test set used in this study through its module and obtained a reasonably good correlation of observed *vs.* predicted IGC$_{50}$ values:

$$\log 1/\mathrm{IGC}_{50}(\mathrm{observed}) = 1.04 \log 1/\mathrm{IGC}_{50}(\mathrm{predicted}) - 0.021 \qquad (19.2)$$

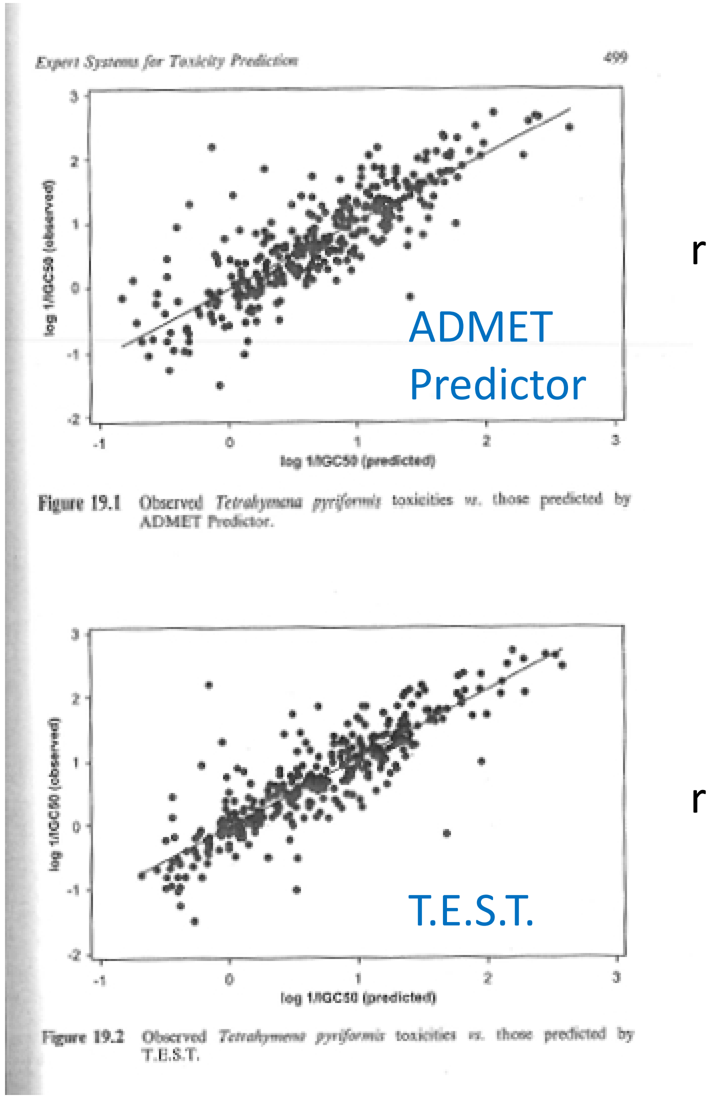
$$n = 350\ r^2 = 0.701\ s = 0.433\ F = 816.9$$

Figure 19.1 shows the plot of observed vs. predicted log 1/IGC$_{50}$ values from ADMET Predictor.

The consensus predictions from T.E.S.T. were somewhat better:

$$\log 1/\mathrm{IGC}_{50}(\mathrm{observed}) = 1.06 \log 1/\mathrm{IGC}_{50}(\mathrm{predicted}) - 0.023 \qquad (19.3)$$

$$\underbrace{n = 349\ r^2 = 0.751\ s = 0.395\ F = 1048.5}$$



Expert Systems for Toxicity Prediction 499

r^2=0.70

Figure 19.1 Observed *Tetrahymena pyriformis* toxicities *vs.* those predicted by ADMET Predictor.

r^2=0.75

Figure 19.2 Observed *Tetrahymena pyriformis* toxicities *vs.* those predicted by T.E.S.T.

*Dearden, 2010

Mutagenicity performance*

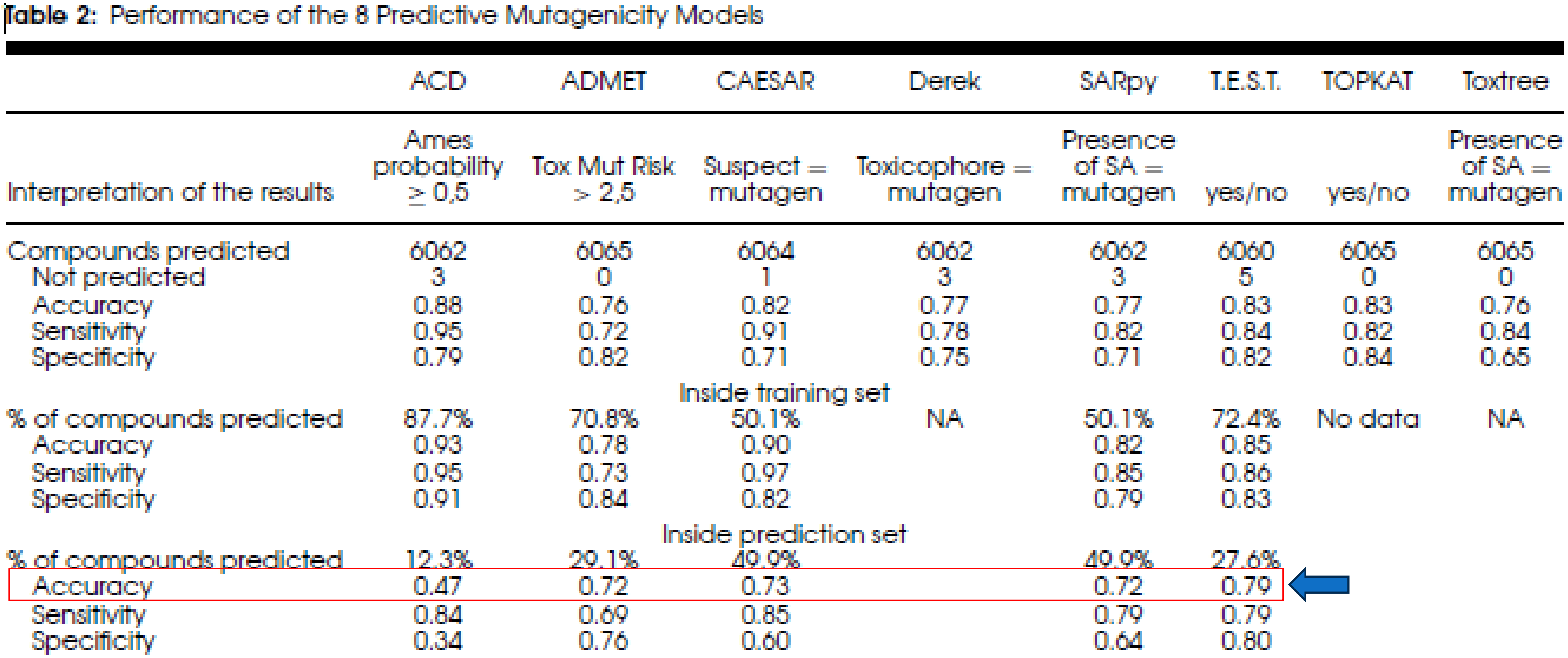
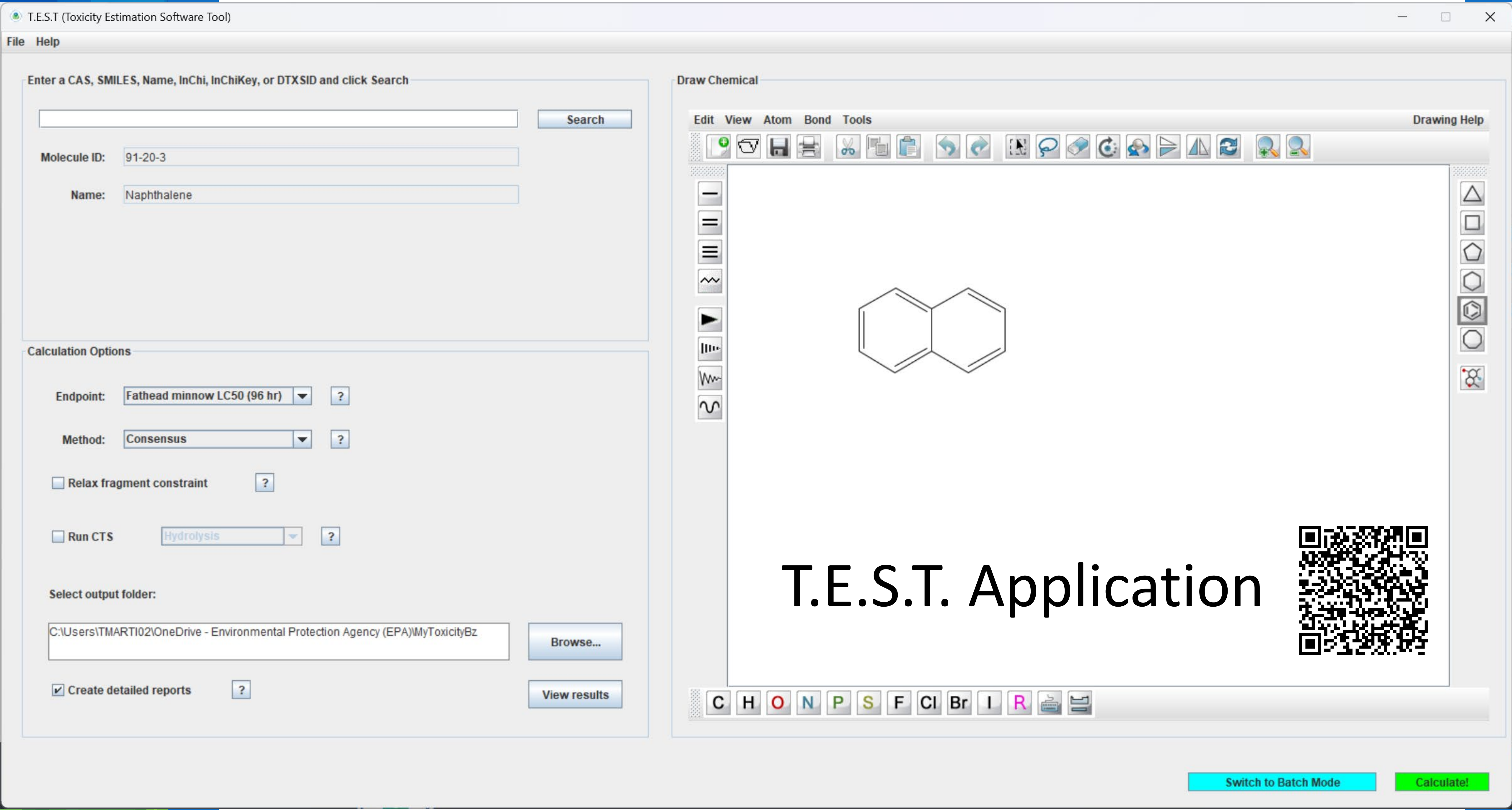


Table 2: Performance of the 8 Predictive Mutagenicity Models

	ACD	ADMET	CAESAR	Derek	SARpy	T.E.S.T.	TOPKAT	Toxtree
Interpretation of the results	Ames probability ≥ 0,5	Tox Mut Risk > 2,5	Suspect = mutagen	Toxicophore = mutagen	Presence of SA = mutagen	yes/no	yes/no	Presence of SA = mutagen
Compounds predicted	6062	6065	6064	6062	6062	6060	6065	6065
Not predicted	3	0	1	3	3	5	0	0
Accuracy	0.88	0.76	0.82	0.77	0.77	0.83	0.83	0.76
Sensitivity	0.95	0.72	0.91	0.78	0.82	0.84	0.82	0.84
Specificity	0.79	0.82	0.71	0.75	0.71	0.82	0.84	0.65
			Inside training set					
% of compounds predicted	87.7%	70.8%	50.1%	NA	50.1%	72.4%	No data	NA
Accuracy	0.93	0.78	0.90		0.82	0.85		
Sensitivity	0.95	0.73	0.97		0.85	0.86		
Specificity	0.91	0.84	0.82		0.79	0.83		
			Inside prediction set					
% of compounds predicted	12.3%	29.1%	49.9%		49.9%	27.6%		
Accuracy	0.47	0.72	0.73		0.72	0.79		
Sensitivity	0.84	0.69	0.85		0.79	0.79		
Specificity	0.34	0.76	0.60		0.64	0.80		

- T.E.S.T. achieved highest prediction accuracy for external set

*Bakhtyari et al., 2013



T.E.S.T. Application

Enter a CAS, SMILES, Name, InChi, InChiKey, or DTXSID and click Search

Molecule ID: 91-20-3

Name: Naphthalene

Calculation Options

Endpoint: Fathead minnow LC50 (96 hr) ?

Method: Consensus ?

☐ Relax fragment constraint ?

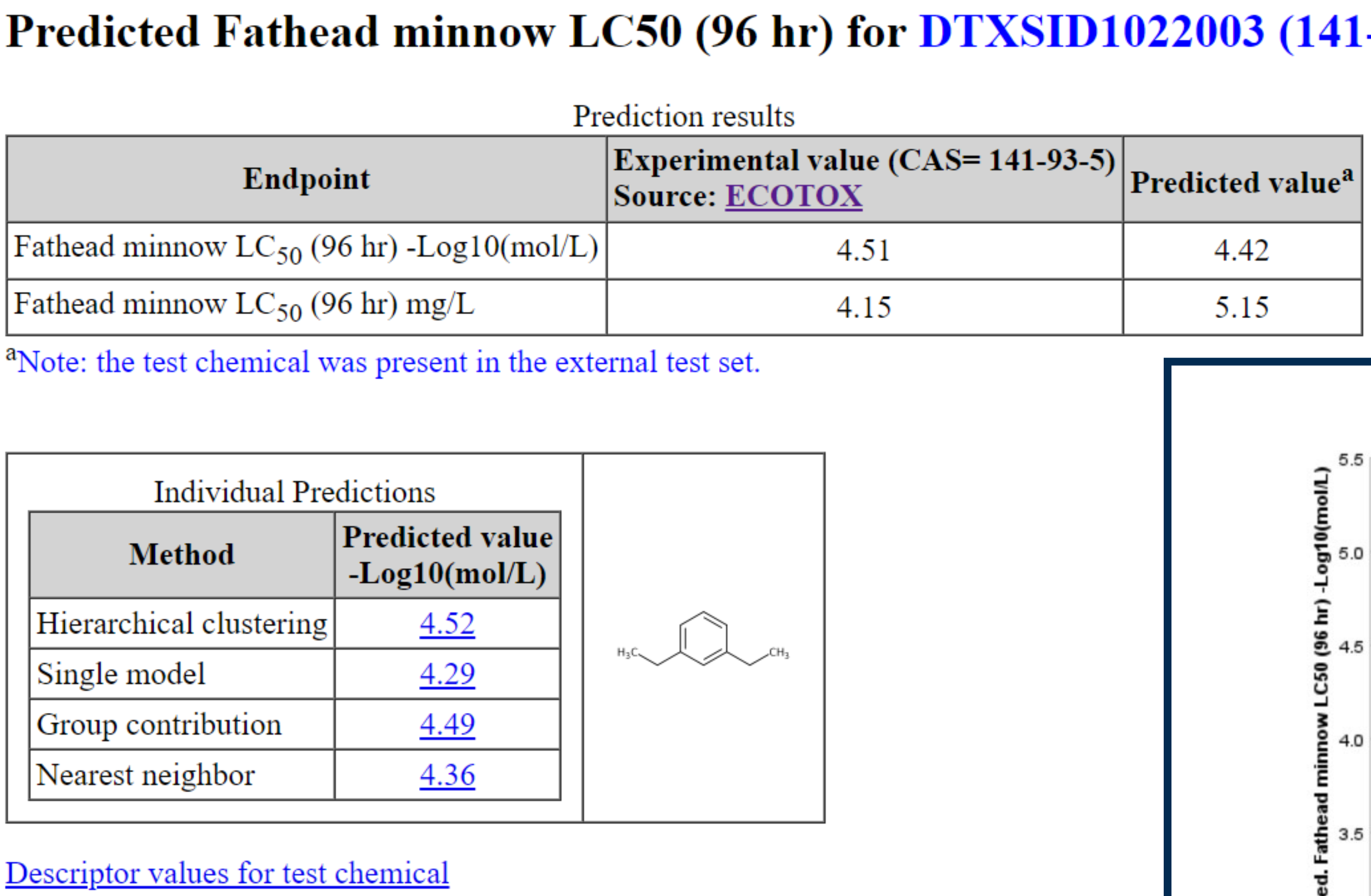
☐ Run CTS Hydrolysis ?

Select output folder:

C:\Users\TMARTI02\OneDrive - Environmental Protection Agency (EPA)\MyToxicityBz

☑ Create detailed reports ?

Sample output: well predicted chemical



Predicted Fathead minnow LC50 (96 hr) for DTXSID1022003 (141-93-5) from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 141-93-5) Source: ECOTOX	Predicted value[a]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	4.51	4.42
Fathead minnow LC_{50} (96 hr) mg/L	4.15	5.15

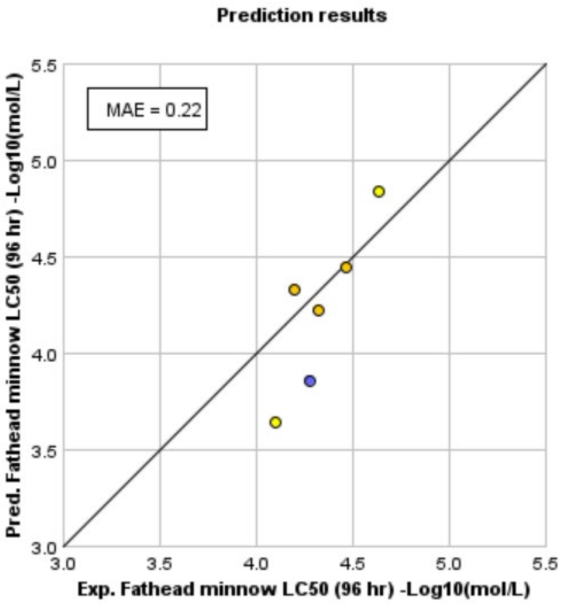
[a]Note: the test chemical was present in the external test set.

Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	4.52
Single model	4.29
Group contribution	4.49
Nearest neighbor	4.36

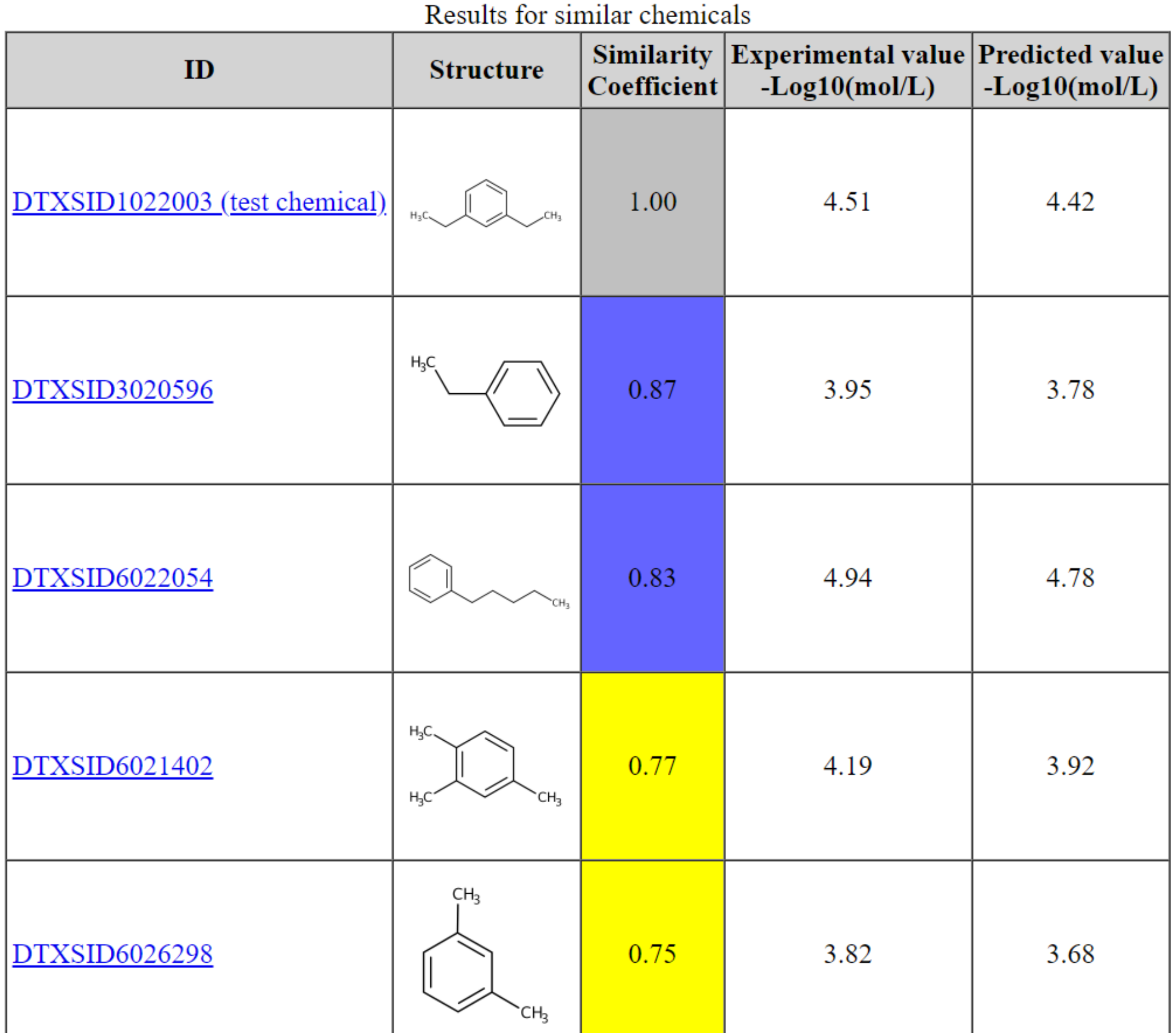
Descriptor values for test chemical

- Predictions are consistent



- Similar test set chemicals are predicted well

Well predicted chemical, cont.

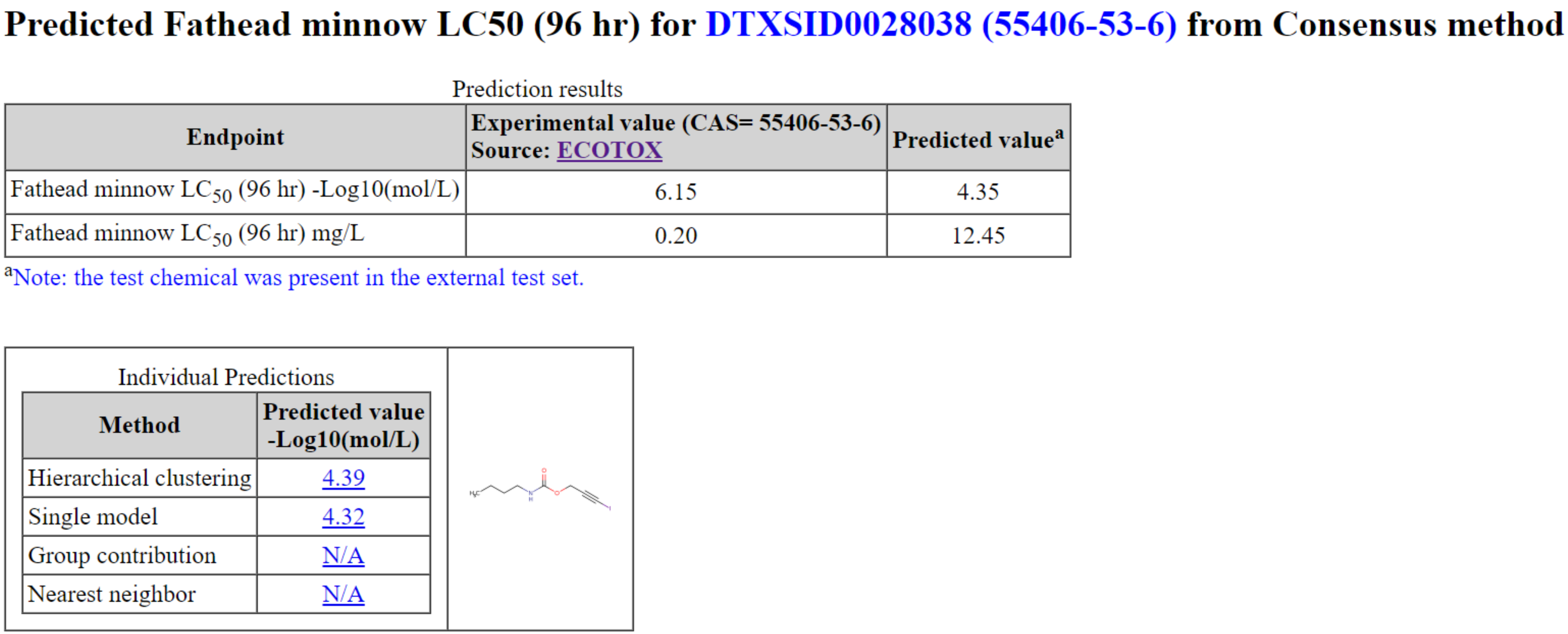


Results for similar chemicals

ID	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
DTXSID1022003 (test chemical)		1.00	4.51	4.42
DTXSID3020596		0.87	3.95	3.78
DTXSID6022054		0.83	4.94	4.78
DTXSID6021402		0.77	4.19	3.92
DTXSID6026298		0.75	3.82	3.68

- Similar chemicals are present in the training set

Poorly predicted chemical



Predicted Fathead minnow LC50 (96 hr) for DTXSID0028038 (55406-53-6) from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 55406-53-6) Source: ECOTOX	Predicted value[a]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	6.15	4.35
Fathead minnow LC_{50} (96 hr) mg/L	0.20	12.45

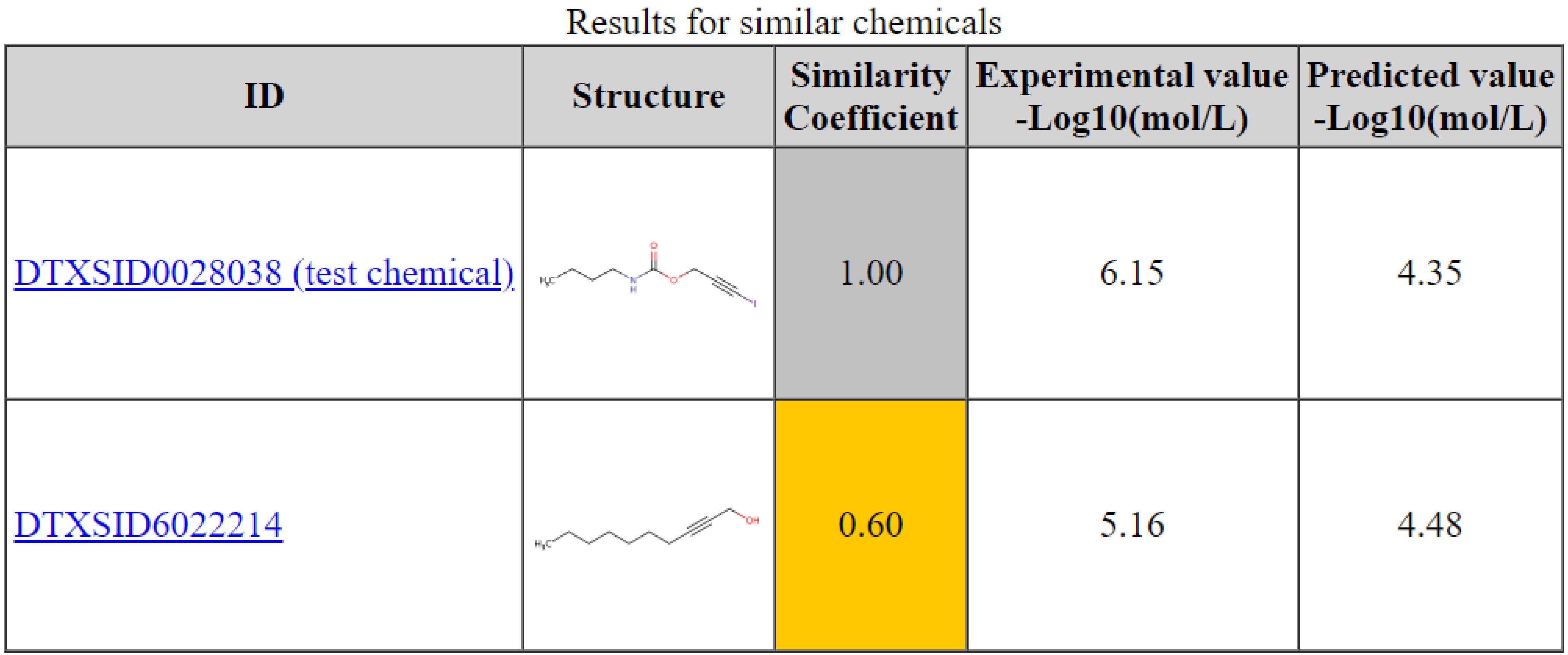
[a]Note: the test chemical was present in the external test set.

Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	4.39
Single model	4.32
Group contribution	N/A
Nearest neighbor	N/A

- For poorly predicted chemicals:
  - The predictions are not consistent between models
  - Some models are outside their applicability domain

Poorly predicted chemical,cont.

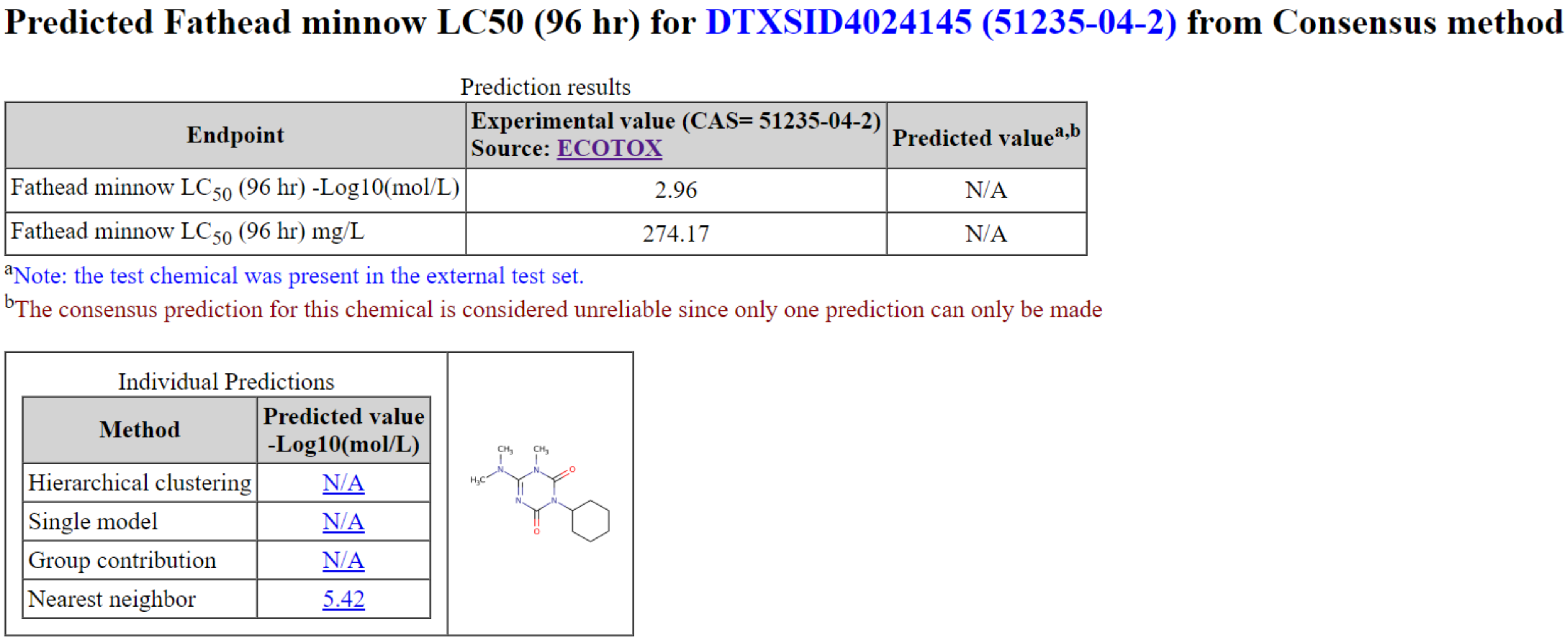


Results for similar chemicals

ID	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
DTXSID0028038 (test chemical)		1.00	6.15	4.35
DTXSID6022214		0.60	5.16	4.48

- There are no sufficiently similar chemicals in the test set
- In this example there is only one similar chemical in the training set and it doesn’t have the same functional groups

Chemical which can’t be predicted



Predicted Fathead minnow LC50 (96 hr) for DTXSID4024145 (51235-04-2) from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 51235-04-2) Source: ECOTOX	Predicted value[a,b]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	2.96	N/A
Fathead minnow LC_{50} (96 hr) mg/L	274.17	N/A

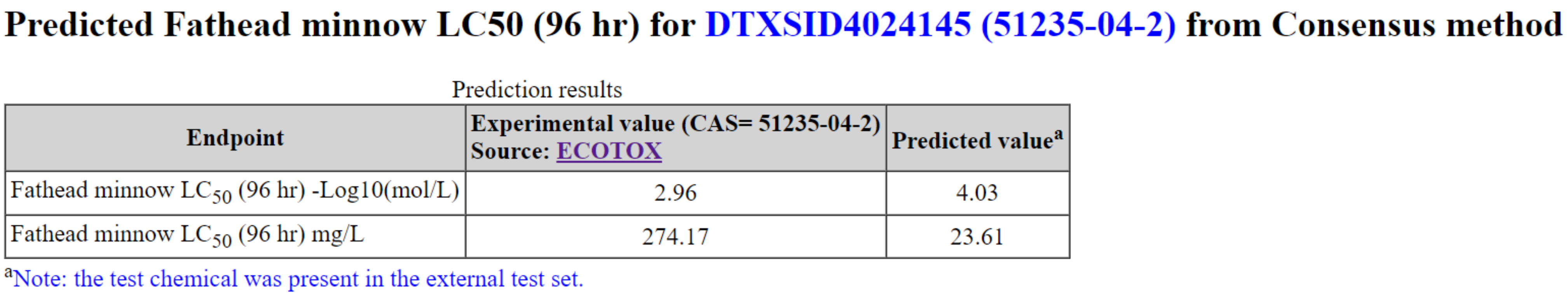
[a] Note: the test chemical was present in the external test set.

[b] The consensus prediction for this chemical is considered unreliable since only one prediction can only be made

Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	N/A
Single model	N/A
Group contribution	N/A
Nearest neighbor	5.42

After relaxing fragment constraint

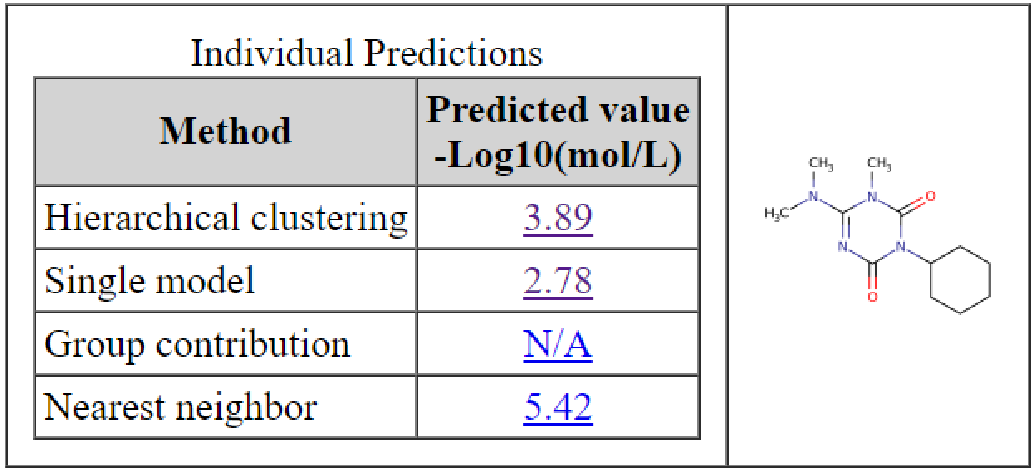


Predicted Fathead minnow LC50 (96 hr) for DTXSID4024145 (51235-04-2) from Consensus method

Prediction results

Endpoint	Experimental value (CAS= 51235-04-2) Source: ECOTOX	Predicted value[a]
Fathead minnow LC_{50} (96 hr) -Log10(mol/L)	2.96	4.03
Fathead minnow LC_{50} (96 hr) mg/L	274.17	23.61

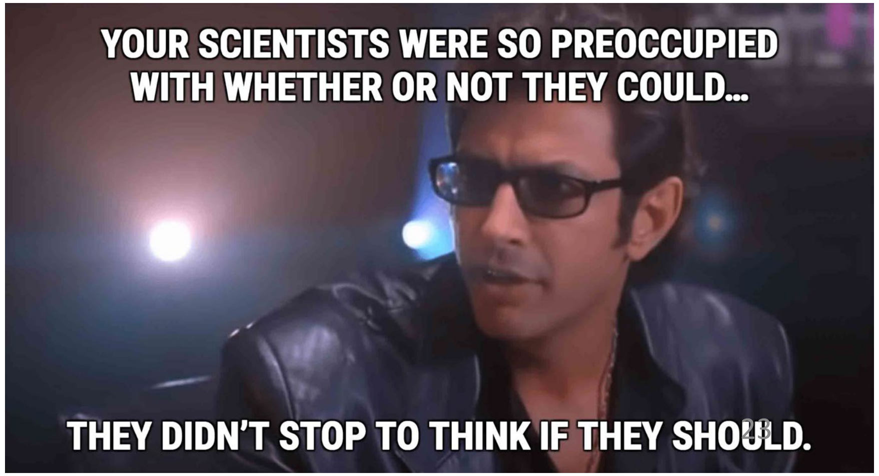
[a]Note: the test chemical was present in the external test set.



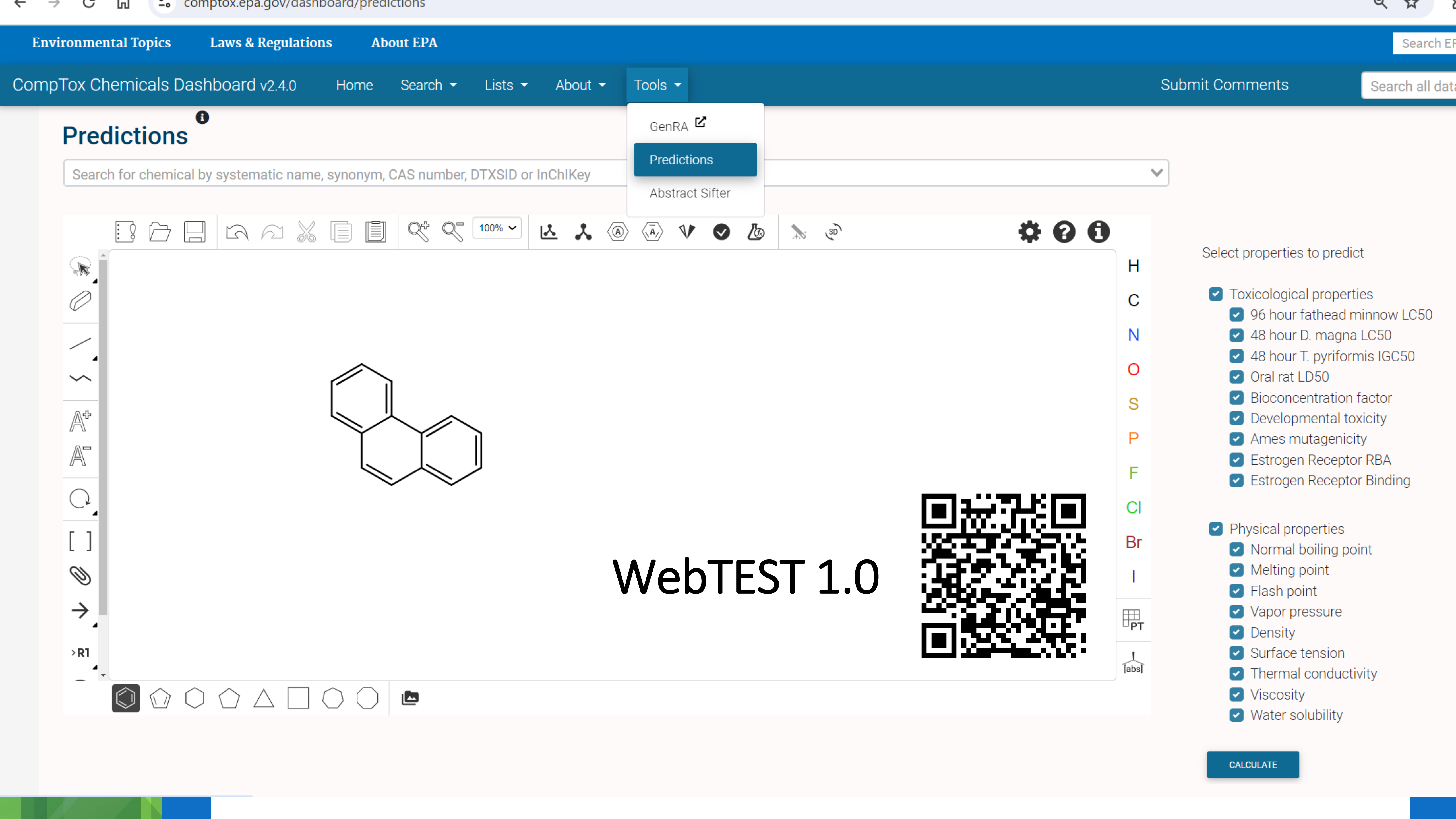
Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	3.89
Single model	2.78
Group contribution	N/A
Nearest neighbor	5.42

- A prediction can be made but it's not reliable (applicability domain worked properly)



YOUR SCIENTISTS WERE SO PREOCCUPIED WITH WHETHER OR NOT THEY COULD...
THEY DIDN'T STOP TO THINK IF THEY SHOULD.



Predictions

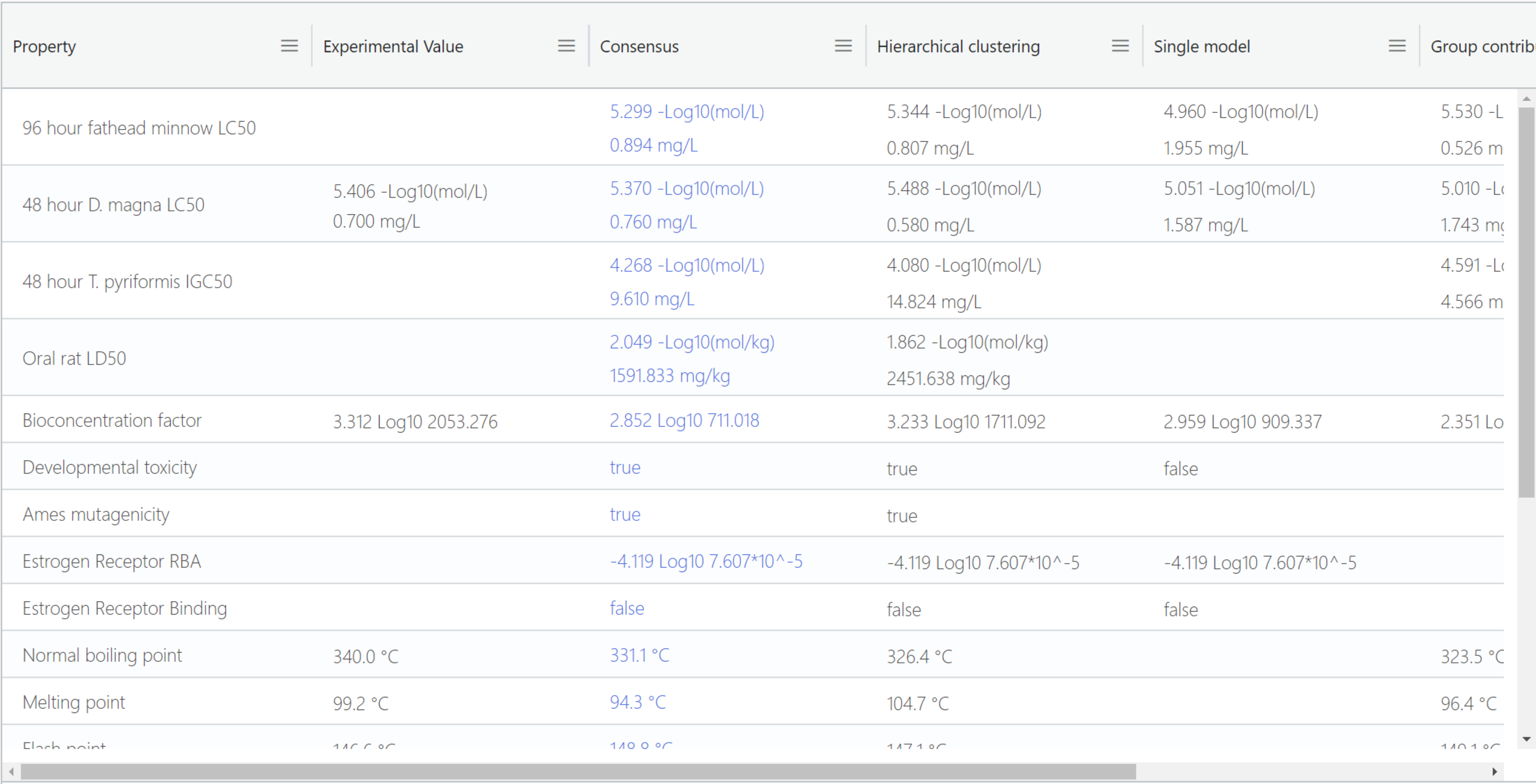
Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey

WebTEST 1.0

Select properties to predict

- ☑ Toxicological properties
  - ☑ 96 hour fathead minnow LC50
  - ☑ 48 hour D. magna LC50
  - ☑ 48 hour T. pyriformis IGC50
  - ☑ Oral rat LD50
  - ☑ Bioconcentration factor
  - ☑ Developmental toxicity
  - ☑ Ames mutagenicity
  - ☑ Estrogen Receptor RBA
  - ☑ Estrogen Receptor Binding
- ☑ Physical properties
  - ☑ Normal boiling point
  - ☑ Melting point
  - ☑ Flash point
  - ☑ Vapor pressure
  - ☑ Density
  - ☑ Surface tension
  - ☑ Thermal conductivity
  - ☑ Viscosity
  - ☑ Water solubility

WebTEST1.0 output



Property	Experimental Value	Consensus	Hierarchical clustering	Single model	Group contrib
96 hour fathead minnow LC50		5.299 -Log10(mol/L) 0.894 mg/L	5.344 -Log10(mol/L) 0.807 mg/L	4.960 -Log10(mol/L) 1.955 mg/L	5.530 -L 0.526 m
48 hour D. magna LC50	5.406 -Log10(mol/L) 0.700 mg/L	5.370 -Log10(mol/L) 0.760 mg/L	5.488 -Log10(mol/L) 0.580 mg/L	5.051 -Log10(mol/L) 1.587 mg/L	5.010 -Lo 1.743 mg
48 hour T. pyriformis IGC50		4.268 -Log10(mol/L) 9.610 mg/L	4.080 -Log10(mol/L) 14.824 mg/L		4.591 -Lo 4.566 m
Oral rat LD50		2.049 -Log10(mol/kg) 1591.833 mg/kg	1.862 -Log10(mol/kg) 2451.638 mg/kg		
Bioconcentration factor	3.312 Log10 2053.276	2.852 Log10 711.018	3.233 Log10 1711.092	2.959 Log10 909.337	2.351 Lo
Developmental toxicity		true	true	false	
Ames mutagenicity		true	true		
Estrogen Receptor RBA		-4.119 Log10 7.607*10^-5	-4.119 Log10 7.607*10^-5	-4.119 Log10 7.607*10^-5	
Estrogen Receptor Binding		false	false	false	
Normal boiling point	340.0 °C	331.1 °C	326.4 °C		323.5 °C
Melting point	99.2 °C	94.3 °C	104.7 °C		96.4 °C
Flash point	146.6 °C	148.8 °C	147.1 °C		149.1 °C

“GET” API Call

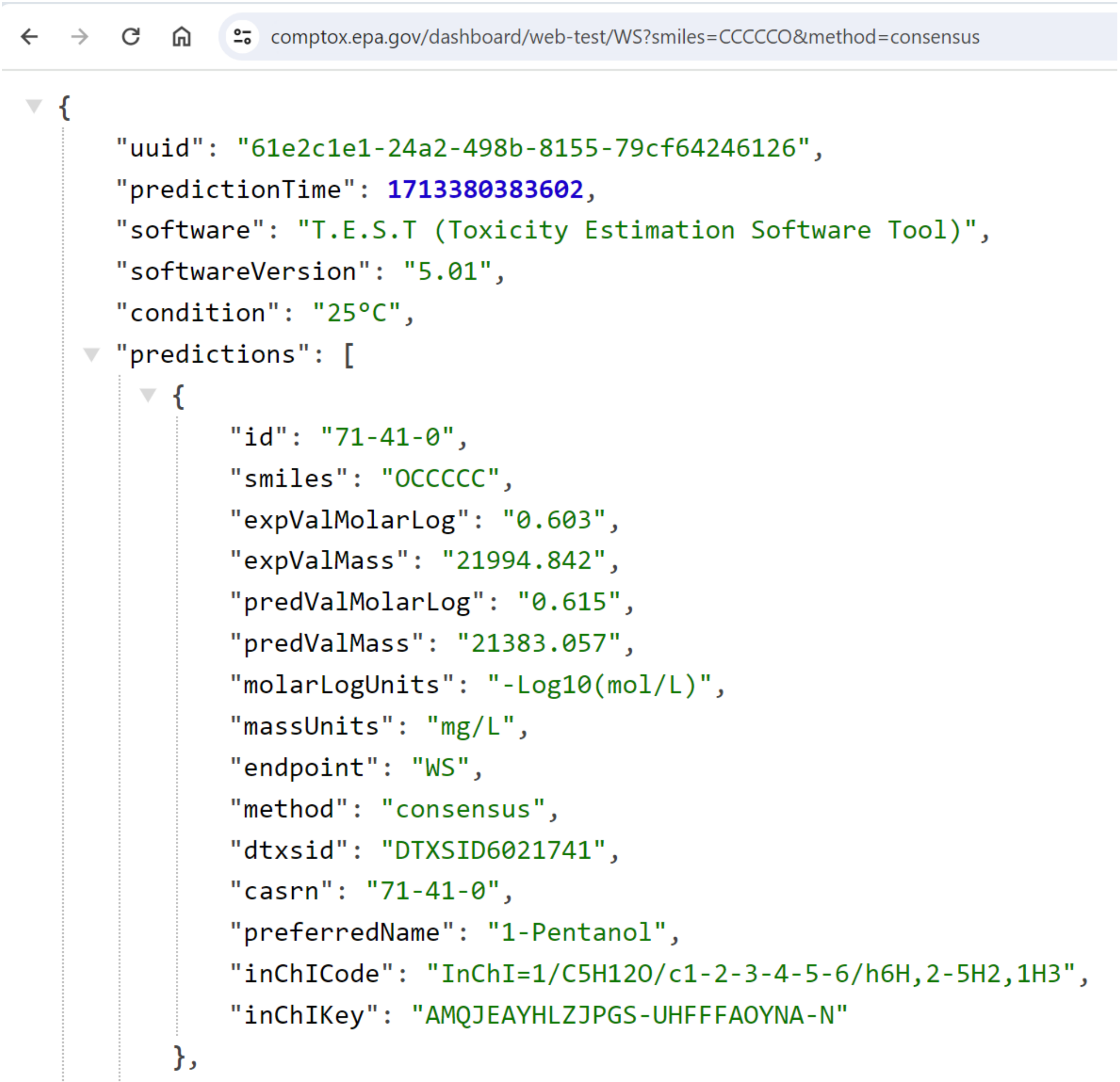
URL/endpointAbbreviation?smiles=desiredSmiles&method=methodAbbreviation

where URL = https://comptox.epa.gov/dashboard/web-test/

Endpoint	Abbreviation
Fathead minnow LC50 (96 hr)	LC50
Daphnia magna LC50 (48 hr)	LC50DM
T. pyriformis IGC50 (48 hr)	IGC50
Oral rat LD50	LD50
Bioaccumulation factor	BCF
Developmental Toxicity	DevTox
Mutagenicity	Mutagenicity
Normal boiling point	BP
Vapor pressure at 25°C	VP
Melting point	MP
Flash point	Density
Density	FP
Surface tension at 25°C	ST
Thermal conductivity at 25°C	TC
Viscosity at 25°C	Viscosity
Water solubility at 25°C	WS

Method	Abbreviation
Hierarchical clustering	hc
Single model	sm
Nearest neighbor	nn
Group contribution	gc
Consensus	consensus (default)

Example “GET” Call



comptox.epa.gov/dashboard/web-test/WS?smiles=CCCCCO&method=consensus

```
{
    "uuid": "61e2c1e1-24a2-498b-8155-79cf64246126",
    "predictionTime": 1713380383602,
    "software": "T.E.S.T (Toxicity Estimation Software Tool)",
    "softwareVersion": "5.01",
    "condition": "25°C",
    "predictions": [
        {
            "id": "71-41-0",
            "smiles": "OCCCCC",
            "expValMolarLog": "0.603",
            "expValMass": "21994.842",
            "predValMolarLog": "0.615",
            "predValMass": "21383.057",
            "molarLogUnits": "-Log10(mol/L)",
            "massUnits": "mg/L",
            "endpoint": "WS",
            "method": "consensus",
            "dtxsid": "DTXSID6021741",
            "casrn": "71-41-0",
            "preferredName": "1-Pentanol",
            "inChICode": "InChI=1/C5H12O/c1-2-3-4-5-6/h6H,2-5H2,1H3",
            "inChIKey": "AMQJEAYHLZJPGS-UHFFFAOYNA-N"
        },
```

WebTEST 2.0: A database centered modeling platform for building and deploying QSAR models

Features of WebTEST2.0

- Central location for real time predictions for EPA models
- Datasets, molecular descriptors, and QSAR methodologies can be versioned in the database
- Utilizes R/python machine learning libraries (e.g. scikit-learn) to build models and make predictions
- Ability to generate WebTEST, PaDEL, Mordred, ToxPrints, and RDKit descriptors

Features of WebTEST2.0, cont.

- Working on adding functionality to deploy models not stored in the database (i.e. “third party” models)
  - Third party models sometimes use special descriptors such as experimental or predicted property values
- Models can be added to the webtool without redeploying the application
- Predictions and molecular descriptors accessible via API calls
- Full documentation of models via Excel summary or QMRF pdf

QSAR Methods in WebTEST2.0

- A variety of QSAR methods can be utilized:
  - MLR – Multilinear Regression
  - RF - Random Forest
  - XGBoost – Extreme Gradient Boosting
  - SVM – Support Vector Machine
  - kNN – k Nearest Neighbors
  - Consensus – average of selected models
- Easily implementable as web services for both model building and model prediction

Physicochemical properties are needed to evaluate environmental and exposure pathways of PFAS

Property	PFAS Experimental Data	All Chemical Experimental Data
HLC	32	1908
VP	101	3440
BP	260	6903
WS	81	9241
LogP	53	14545
MP	195	29052

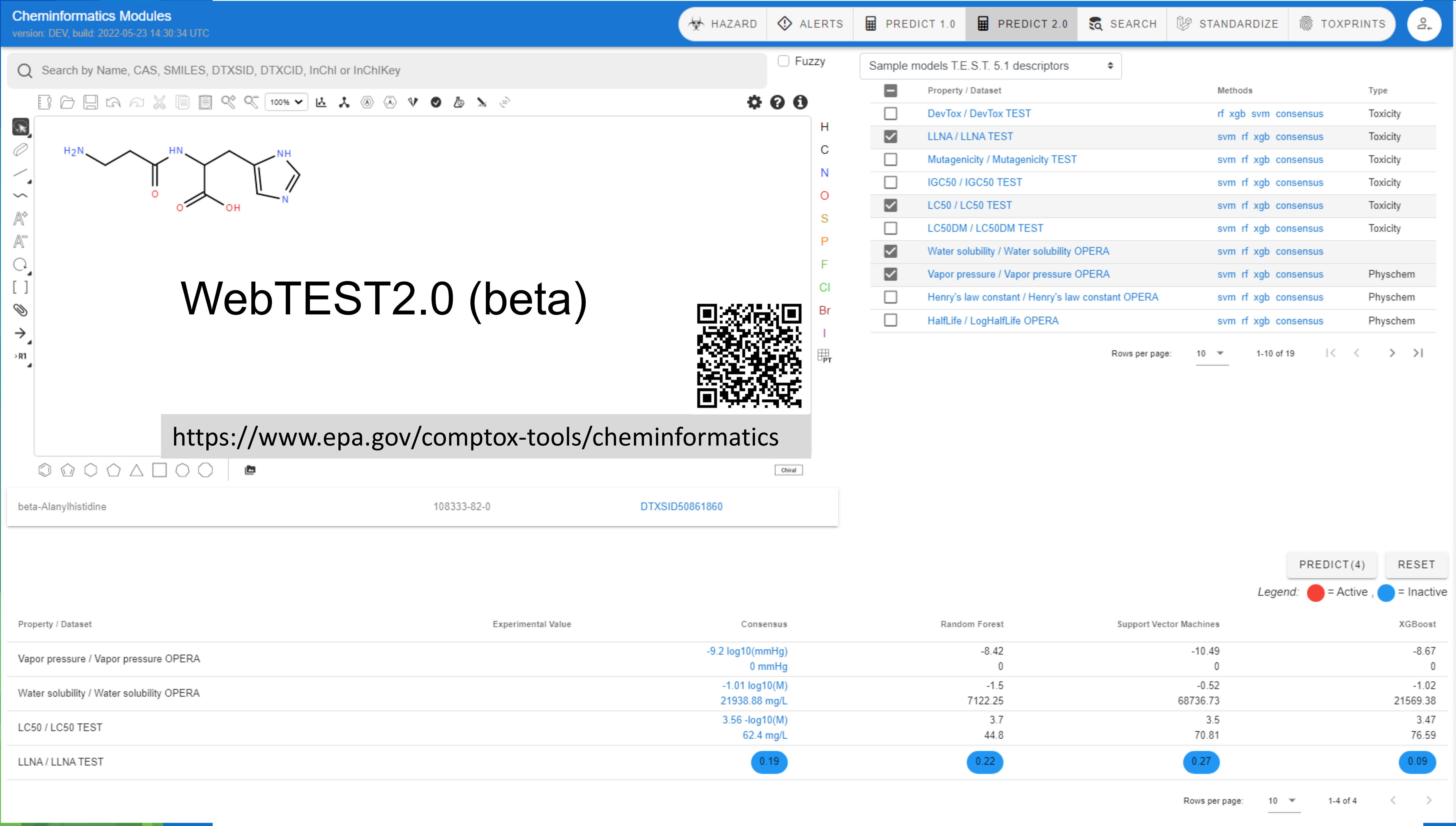
- Curated experimental data for PFAS have been limited.
- Physical chemical data was compiled from multiple public sources and QCd to the data source.
- Two sets of consensus QSAR models were developed for each of the six physical chemical properties
  - PFAS only model
  - All chemicals model
- Consensus model averaged the predictions from XGBoost and Random forest models

Development of Updated QSAR Models to Predict PFAS Physical Chemical Properties

Test set results for PFAS

	Trained to All Chemicals		Trained to PFAS	
Property	R^2	MAE	R^2	MAE
HLC	0.68	1.18	0.85	1.13
VP	0.97	0.55	0.93	0.63
BP	0.88	20.1	0.87	19.4
WS	0.65	0.83	0.57	0.83
LogP	0.59	0.91	0.47	1.07
MP	0.84	40.2	0.77	46.1

- Consensus QSAR models trained on all chemical classes gave slighter better results for predicting physical chemical properties of PFAS.
- Model performance for PFAS is similar to CADASTER models, which were trained on PFAS substances.
- Additional QC of experimental data is on-going.



WebTEST2.0 (beta)

https://www.epa.gov/comptox-tools/cheminformatics

beta-Alanylhistidine	108333-82-0	DTXSID50861860

Sample models T.E.S.T. 5.1 descriptors

☐	Property / Dataset	Methods	Type
☐	DevTox / DevTox TEST	rf xgb svm consensus	Toxicity
☑	LLNA / LLNA TEST	svm rf xgb consensus	Toxicity
☐	Mutagenicity / Mutagenicity TEST	svm rf xgb consensus	Toxicity
☐	IGC50 / IGC50 TEST	svm rf xgb consensus	Toxicity
☑	LC50 / LC50 TEST	svm rf xgb consensus	Toxicity
☐	LC50DM / LC50DM TEST	svm rf xgb consensus	Toxicity
☑	Water solubility / Water solubility OPERA	svm rf xgb consensus	
☑	Vapor pressure / Vapor pressure OPERA	svm rf xgb consensus	Physchem
☐	Henry's law constant / Henry's law constant OPERA	svm rf xgb consensus	Physchem
☐	HalfLife / LogHalfLife OPERA	svm rf xgb consensus	Physchem

Rows per page: 10 1-10 of 19

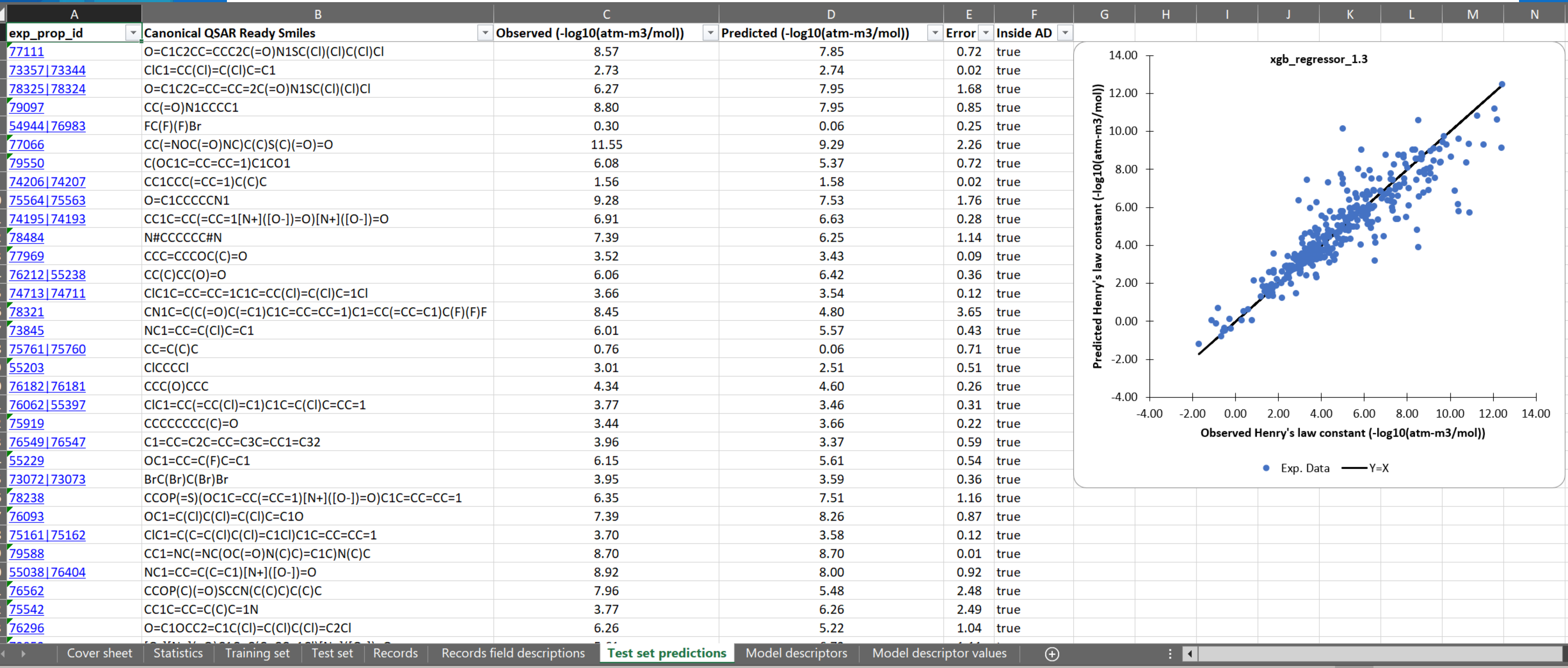
PREDICT (4) RESET

Legend: 🟥 = Active , 🟦 = Inactive

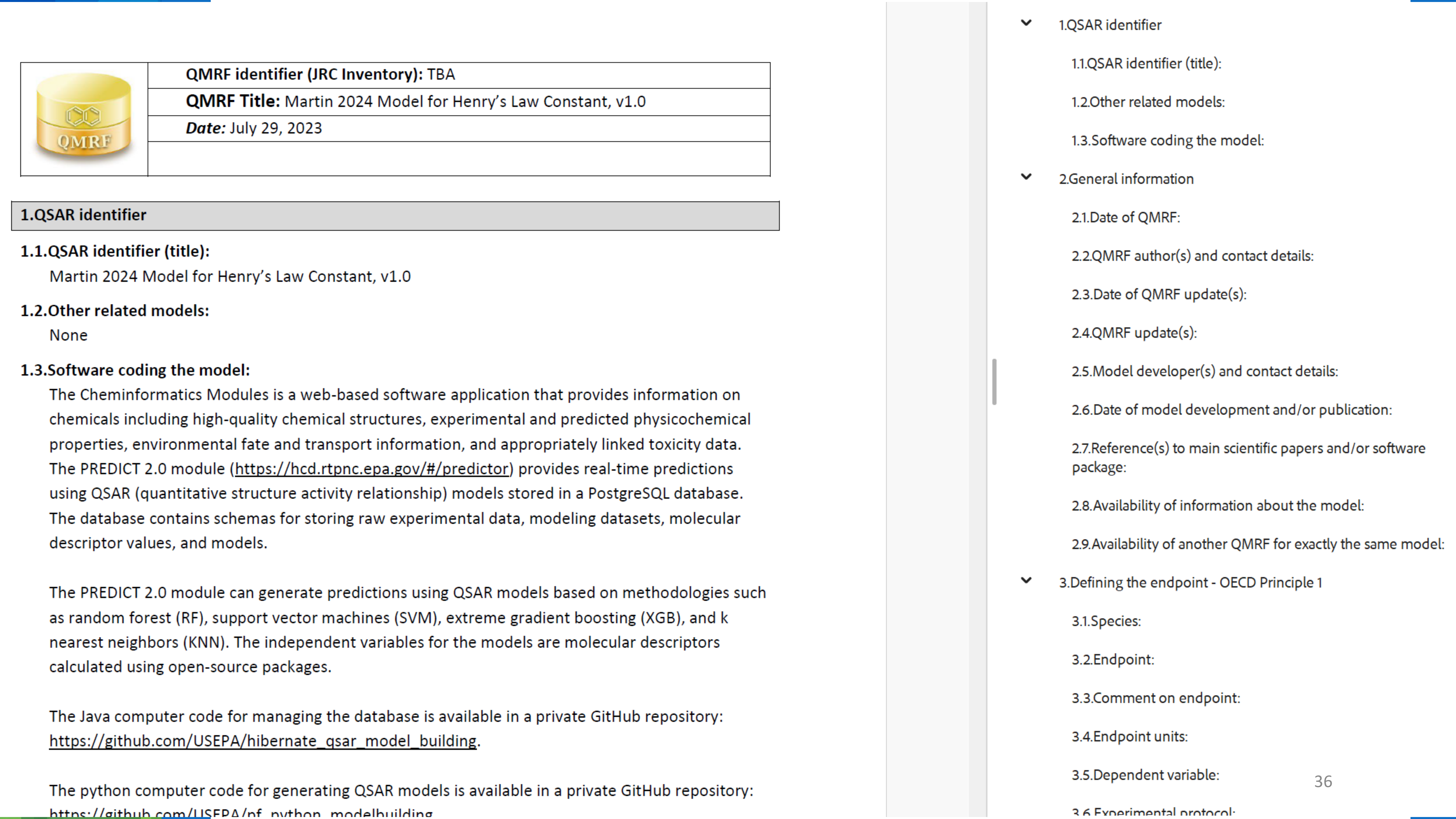
Property / Dataset	Experimental Value	Consensus	Random Forest	Support Vector Machines	XGBoost
Vapor pressure / Vapor pressure OPERA		-9.2 log10(mmHg) 0 mmHg	-8.42 0	-10.49 0	-8.67 0
Water solubility / Water solubility OPERA		-1.01 log10(M) 21938.88 mg/L	-1.5 7122.25	-0.52 68736.73	-1.02 21569.38
LC50 / LC50 TEST		3.56 -log10(M) 62.4 mg/L	3.7 44.8	3.5 70.81	3.47 76.59
LLNA / LLNA TEST		0.19	0.22	0.27	0.09

Rows per page: 10 1-4 of 4

Excel summary



exp_prop_id	Canonical QSAR Ready Smiles	Observed (-log10(atm-m3/mol))	Predicted (-log10(atm-m3/mol))	Error	Inside AD
77111	O=C1C2CC=CCC2C(=O)N1SC(Cl)(Cl)C(Cl)Cl	8.57	7.85	0.72	true
73357|73344	ClC1=CC(Cl)=C(Cl)C=C1	2.73	2.74	0.02	true
78325|78324	O=C1C2C=CC=CC=2C(=O)N1SC(Cl)(Cl)Cl	6.27	7.95	1.68	true
79097	CC(=O)N1CCCC1	8.80	7.95	0.85	true
54944|76983	FC(F)(F)Br	0.30	0.06	0.25	true
77066	CC(=NOC(=O)NC)C(C)S(C)(=O)=O	11.55	9.29	2.26	true
79550	C(OC1C=CC=CC=1)C1CO1	6.08	5.37	0.72	true
74206|74207	CC1CCC(=CC=1)C(C)C	1.56	1.58	0.02	true
75564|75563	O=C1CCCCCN1	9.28	7.53	1.76	true
74195|74193	CC1C=CC(=CC=1[N+]([O-])=O)[N+]([O-])=O	6.91	6.63	0.28	true
78484	N#CCCCCC#N	7.39	6.25	1.14	true
77969	CCC=CCCOC(C)=O	3.52	3.43	0.09	true
76212|55238	CC(C)CC(O)=O	6.06	6.42	0.36	true
74713|74711	ClC1C=CC=CC=1C1C=CC(Cl)=C(Cl)C=1Cl	3.66	3.54	0.12	true
78321	CN1C=C(C(=O)C(=C1)C1C=CC=CC=1)C1=CC(=CC=C1)C(F)(F)F	8.45	4.80	3.65	true
73845	NC1=CC=C(Cl)C=C1	6.01	5.57	0.43	true
75761|75760	CC=C(C)C	0.76	0.06	0.71	true
55203	ClCCCCl	3.01	2.51	0.51	true
76182|76181	CCC(O)CCC	4.34	4.60	0.26	true
76062|55397	ClC1=CC(=CC(Cl)=C1)C1C=C(Cl)C=CC=1	3.77	3.46	0.31	true
75919	CCCCCCCC(C)=O	3.44	3.66	0.22	true
76549|76547	C1=CC=C2C=CC=C3C=CC1=C32	3.96	3.37	0.59	true
55229	OC1=CC=C(F)C=C1	6.15	5.61	0.54	true
73072|73073	BrC(Br)C(Br)Br	3.95	3.59	0.36	true
78238	CCOP(=S)(OC1C=CC(=CC=1)[N+]([O-])=O)C1C=CC=CC=1	6.35	7.51	1.16	true
76093	OC1=C(Cl)C(Cl)=C(Cl)C=C1O	7.39	8.26	0.87	true
75161|75162	ClC1=C(C=C(Cl)C(Cl)=C1Cl)C1C=CC=CC=1	3.70	3.58	0.12	true
79588	CC1=NC(=NC(OC(=O)N(C)C)=C1C)N(C)C	8.70	8.70	0.01	true
55038|76404	NC1=CC=C(C=C1)[N+]([O-])=O	8.92	8.00	0.92	true
76562	CCOP(C)(=O)SCCN(C(C)C)C(C)C	7.96	5.48	2.48	true
75542	CC1C=CC=C(C)C=1N	3.77	6.26	2.49	true
76296	O=C1OCC2=C1C(Cl)=C(Cl)C(Cl)=C2Cl	6.26	5.22	1.04	true



QMRF identifier (JRC Inventory): TBA

QMRF Title: Martin 2024 Model for Henry’s Law Constant, v1.0

Date: July 29, 2023

1.QSAR identifier

1.1.QSAR identifier (title):

Martin 2024 Model for Henry’s Law Constant, v1.0

1.2.Other related models:

None

1.3.Software coding the model:

The Cheminformatics Modules is a web-based software application that provides information on chemicals including high-quality chemical structures, experimental and predicted physicochemical properties, environmental fate and transport information, and appropriately linked toxicity data. The PREDICT 2.0 module (https://hcd.rtpnc.epa.gov/#/predictor) provides real-time predictions using QSAR (quantitative structure activity relationship) models stored in a PostgreSQL database. The database contains schemas for storing raw experimental data, modeling datasets, molecular descriptor values, and models.

The PREDICT 2.0 module can generate predictions using QSAR models based on methodologies such as random forest (RF), support vector machines (SVM), extreme gradient boosting (XGB), and k nearest neighbors (KNN). The independent variables for the models are molecular descriptors calculated using open-source packages.

The Java computer code for managing the database is available in a private GitHub repository: https://github.com/USEPA/hibernate_qsar_model_building.

The python computer code for generating QSAR models is available in a private GitHub repository: https://github.com/USEPA/pf_python_modelbuilding

Demo

Questions???

The views expressed in this presentation are those of the author and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency