Sensitivity and Specificity Assessment of Quantitative Structure-Activity Relationship Tools Applied to Flavor Ingredients

Trudi Denoon, Chastain Anderson, Molly E. Pitegoff, Vanessa Haase, Britt Langston, Wanyoike Kangethe, Donna C. Williams

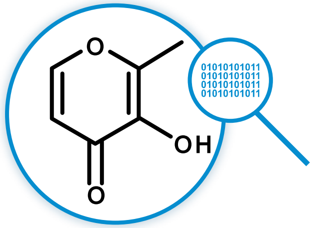
Society for Risk Analysis

12.09.25

Quantitative Structure-Activity Relationship (QSAR)

FDA's Center for Tobacco Products requires applicants to demonstrate that a new tobacco product is appropriate for the protection of public health which may include an evaluation of the carcinogenic potential of the new products and ingredients, including GRAS flavors

Risk assessments must be conducted for these ingredients to identify any potential hazards

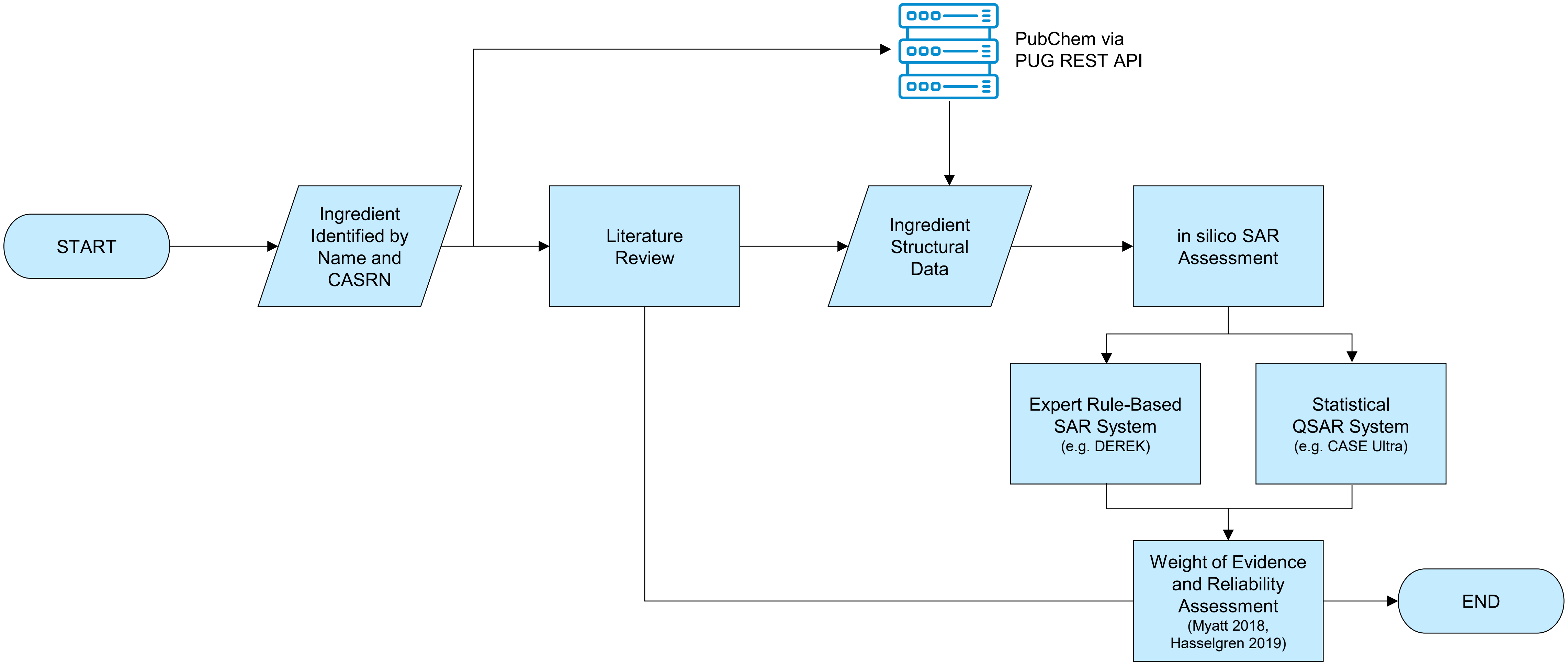


QSAR models are machine-learning-based statistical tools commonly used in predictive toxicology and hazard I.D.

Provides a cost-effective, time-efficient mechanism

- Facilitates rapid screening of chemical structures to evaluate mutagenic, genotoxic, and/or carcinogenic potential, even in cases where there is little to no data
- Prioritizes compounds for review and reduces the need for animal testing

In Silico Hazard Identification Framework



Different Statistical QSAR Tools

There are several QSAR models commonly used for hazard identification:

- **CASE Ultra**
- **Leadscope**
- **OECD QSAR Toolbox**

Most have been validated for chemicals and chemical classes in the pharmaceutical space, and CTP has previously validated some of these tools in the combusted cigarette space

Here, we performed a validation study for flavor ingredients associated with novel potentially reduced-harm tobacco products

TOOL DIFFERENCES

Each tool is unique in its reference databases; resulting in slightly different outcomes

VARIATIONS IN PREDICTIONS

Variations in predictions may impede risk management decisions concerning potential toxicity endpoints

FALSE PREDICTION CALLS

There is a risk of generating false positives and false negatives, which can produce differences in sensitivity, specificity, and overall accuracy

PREDICTION ACCURACY

There is a need to validate the prediction accuracy of different QSAR models to improve confidence in the model outputs for the flavor ingredient space

Statistics-based QSAR: CASE Ultra

Case Ultra

Statistics-based (Q)SAR software platform

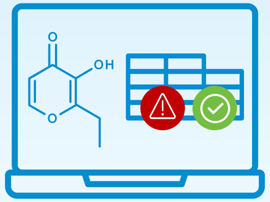
developed by MultiCASE Inc. in the United States in collaboration with the U.S. Food and Drug Administration

The platform utilizes a machine-learning algorithm, in combination with molecular descriptors generated by fragmenting the training set structures

Any fragments identified as statistically associated with active molecules in the training set are **defined as structural alerts**

Other fragments are identified as deactivating features that decrease the potency of the alerts

MODEL



GENERATES A CONFIDENCE SCORE BETWEEN 0 AND 1

to predict the likelihood of a test chemical being positive based on alerts and deactivating features

Statistics-based QSAR: Leadscope Model Applier; LSMA (1 of 2)

LSMA

Statistics-based QSAR tool

developed by Leadscope Inc. in collaboration with the U.S. Food and Drug Administration

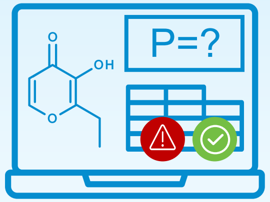
LSMA is a data mining software, consisting of models that use training set chemicals and publicly available toxicological data

The models are designed from several calculated descriptors such as:

- Pre-defined structural features
- Physicochemical properties
- Active structural features from published data

A partial logistic regression model is used to establish a relationship between the descriptors and a toxicological response and generate a prediction

MODEL



DISPLAYS THE PROBABILITY OF OBTAINING A POSITIVE RESULT and structural characteristics contributing to activation and deactivation

GENERATES A CONFIDENCE SCORE BETWEEN 0 AND 1 **to predict the likelihood of a test chemical being positive** based on available alerts and deactivating features

Statistics-based QSAR: Leadscope Model Applier; LSMA (2 of 2)

LSMA

Statistics-based QSAR tool

developed by Leadscope Inc. in collaboration with the U.S. Food and Drug Administration

For most models, a test chemical is evaluated as:

ACTIVE
FOR A SET OF MODELS
IF AVERAGE PROBABILITY
>0.5

INACTIVE
FOR A SET OF MODELS
IF AVERAGE PROBABILITY
<0.5

The exceptions are the Genetox *Salmonella* and *E. coli* AT models

INACTIVE/ NEGATIVE
PROBABILITY
≤0.4

ACTIVE/ POSITIVE
PROBABILITY
≥0.6

INTERMEDIATE
MUTAGENICITY PREDICTIONS
BETWEEN
0.4 and 0.6

DISPLAYS DETAIL TEST INFORMATION
related to the chemical substances in the training set

This enables experts to conduct detailed reviews of the training set

OECD Toolbox

OECD QSAR Toolbox

Consists of databases (such as Danish QSAR Database (Danish Q.D.)), that contain chemical, toxicological and ecotoxicological data and estimated values

derived from libraries that contain third party QSAR models such as CASE Ultra, Leadscope and SciQSAR

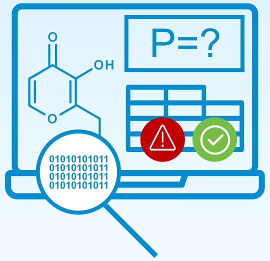
Danish Q.D. was developed by the Technical University of Denmark

- Contains estimates/predictions from over 200 QSAR models
- Database consists of QSAR predictions for over 600,000 substances and nearly all 80,000 organic single constituents that are registered under REACH

SciQSAR is a statistics-based software developed by SciMatics to build models for mutagenicity and genotoxicity using FDA alerts

- SciQSAR contains over 400 built in molecular descriptors

MODEL



PREDICTIVE MODEL
is built on a regression analysis

STRUCTURAL ALERTS
are used to create a profile of the query compound

QUERY COMPOUND
is profiled, and the generated profile is used to screen for compounds in the databases with the same or similar profile

Validation and Analysis Framework – 166 Oral Flavor Ingredients

Query Compound

Case Ultra
Mutagenicity and Genetox models

LS
Mutagenicity and Genetox models

OECD SciQSAR
Mutagenicity and Genetox models

Review publicly available mutagenicity and genetox data and calculate Sensitivity, Specificity and Accuracy

SENSITIVITY
(True Positives) ÷
(True Positives +False Negatives)

SPECIFICITY
(True Negatives) ÷
(True Negatives +False Positives)

ACCURACY
(True Negatives + True Positives) ÷
(True Negatives +False Positives +False Negatives +True Positives)

Compare between all three software programs

Results – Bacterial Mutagenicity

166

Commonly used flavor ingredients were assessed for

- **Bacterial mutagenicity and**
- **Mammalian genotoxicity**

using three previously mentioned QSAR tools

Sensitivity, Specificity and Accuracy were calculated

Bacterial Mutagenicity

QSAR Tools

Parameter (%)	Case Ultra		Leadscope						OECD Toolbox Danish SciQSAR	
	GT1_BMUT-OECD 471 Bacterial Mutagenicity	Out of Domain Calls	Bacterial Mutation v2	Out of Domain Calls	*E Coli - Sal 102* A-T Mut v2	Out of Domain Calls	*Salmonella* Mut v4	Out of Domain Calls	Ames test in *S.typhimurium* (in vitro)	Out of Domain Calls
Specificity	94		95	1	97		96		99	
Accuracy	93	5	95		96	14	95	2	96	18
Sensitivity	80		90		88		80		38	

Results – Mouse Lymphoma

166

Commonly used flavor ingredients were assessed for

- **Bacterial mutagenicity and**
- **Mammalian genotoxicity**

using three previously mentioned QSAR tools

Sensitivity, Specificity and Accuracy were calculated

Mouse Lymphoma Assay

QSAR Tools

Parameter (%)	Case Ultra				Leadscope				OECD Toolbox Danish SciQSAR	
	In Silico Mouse Lymphoma, activated	Out of Domain Calls	In Silico Mouse Lymphoma, unactivated	Out of Domain Calls	In Silico Mouse Lymphoma, activated	Out of Domain Calls	In Silico Mouse Lymphoma, unactivated	Out of Domain Calls	Mutations in TK Locus in Mouse Lymphoma cells	Out of Domain Calls
Specificity	93		76		83		52		40	
Accuracy	93	9	77	10	83	23	54	16	42	82
Sensitivity	100*		100*		100*		100*		100*	

Results – In Vivo Micronucleus

In Vivo Mouse Micronucleus

QSAR Tools

Parameter (%)	Case Ultra		Leadscope		OECD Toolbox Danish SciQSAR	
	GT3_MNT_MOUSE-Micronucleus test, in vivo, mouse	Out of Domain Calls	Leadscope In Vivo Micronuc Mouse v2	Out of Domain Calls	Micronucleus test in Mouse Erythrocytes	Out of Domain Calls
Specificity	83	15	80	15	95	32
Accuracy	84		80		94	
Sensitivity	100*		100*		50	

Conclusion

CASE Ultra

Leadscope

OECD QSAR Toolbox

Bacterial Mutagenicity

All three systems shared **similar specificity and accuracy**.
However, **OECD had the lowest sensitivity**

Mammalian Mutagenicity (MLA)

CASEUltra had greatest specificity and accuracy, followed by Leadscope and OECD toolbox
*Sensitivity for all three tools were **the same;** artifact of low false negatives

In Vivo Genotoxicity (Mouse MN)

OECD QSAR toolbox has the **highest specificity**, but the **lowest sensitivity**

OECD QSAR Had the Highest Number of "Out of Domain" Calls

which can **affect the available data set**

Overall, OECD QSAR toolbox appeared to be the least sensitive tool

Most limited data processing capabilities

Limitations and Future Directions

EXPANDING

Data Set

could help inform the sensitivity calculations by providing the ability to capture more false negatives, rather than provide a higher N

Positive Control Data Set

is especially important as **a low true positive data set can limit the capabilities of some tests**

These tools don't provide the same coverage

and can generate a varied number of "out of domain" calls

which can greatly affect the prediction outcomes and affect:

Sensitivity | Specificity | Accuracy