QMRF Title: **IFSQSAR HHLT v1; QSAR for whole body total elimination (terminal) half-life in humans, Arnot, Brown and Wania 2014**
Contact: **Trevor N. Brown – trevor.n.brown@gmail.com**
Date: **8 March 2025**

1. QSAR identifier

1.1. QSAR identifier (title):

IFSQSAR HHLT v1; QSAR for whole body total elimination (terminal) half-life in humans, Arnot, Brown, and Wania 2014

1.2. Other related models:

IFSQSAR HHLB v1; QSAR for whole body biotransformation half-life in reference human, Arnot, Brown, and Wania 2014

1.3. Software coding the model:

IFSQSAR python package, HHLT v1 is included in versions 1.0.0+ and up.

Model is coded in python, with openbabel used for chemistry and numpy used for mathematics.

2. General information

2.0. Abstract

From the paper abstract: *The whole body, total elimination half-life (HLT) and the whole body is a key parameter determining the extent of bioaccumulation, biological concentration, and risk from chemical exposure. Approximately 1900 HLs for human adults were collected and reviewed. HLs span approximately 7.5 orders of magnitude from 0.05 h for nitroglycerin to 2 × 10^6 h for 2,3,4,5,2',3',5',6'-octachlorobiphenyl with a median of 7.6 h. The automated Iterative Fragment Selection (IFS) method was applied to develop and evaluate various quantitative structure−activity relationships (QSARs) to predict HLT from chemical structure. The HLT QSAR shows similar statistical performance for training and external validation sets*.

2.1. Date of QMRF

8 March 2025

2.2. QMRF author(s) and contact details

Trevor N. Brown – trevor.n.brown@gmail.com

2.3. Date of QMRF update(s)

NA

2.4. QMRF update(s)

NA

2.5. Model developer(s) and contact details

Jon A. Arnot – ARC Arnot Research & Consulting, Toronto, ON, Canada; Department of Physical and Environmental Science, University of Toronto, ON, Canada; Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON, Canada

Trevor N. Brown – ARC Arnot Research & Consulting, Toronto, ON, Canada
trevor.n.brown@gmail.com

Frank Wania – University of Toronto Scarborough, Department of Physical and Environmental Sciences, Toronto, ON, Canada

2.6. Date of model development and/or publication

2014

2.7. Reference(s) to main scientific papers and/or software package

[1] Arnot, J.A., T.N. Brown, and F. Wania, Estimating screening-level organic chemical half-lives in humans. Environmental Science & Technology, 2014. 48(1): p. 723-30.

[2] Brown, T.N., Predicting hexadecane-air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data. SAR and QSAR in Environmental Research, 2014. 25(1): p. 51-71.

[3] Brown, T.N., J.A. Arnot, and F. Wania, Iterative fragment selection: a group contribution approach to predicting fish biotransformation half-lives. Environmental Science & Technology, 2012. 46(15): p. 8253-60.

[4] Brown, T.N., J.M. Armitage, and J.A. Arnot, Application of an Iterative Fragment Selection (IFS) Method to Estimate Entropies of Fusion and Melting Points of Organic Chemicals. Molecular Informatics, 2019. 38(8-9): p. e1800160.

2.8. Availability of information about the model

The model is non-proprietary: training and external validation datasets are available in the publication supplemental information. The model is also available for use on the online platform EAS-E Suite: https://arnotresearch.com/eas-e-suite/.

2.9. Availability of another QMRF for exactly the same model

NA

3. Defining the endpoint - OECD Principle 1: “A DEFINED ENDPOINT"

3.1. Species

Human

3.2. Endpoint

Whole body total elimination (terminal) half-life (HHLT)

3.3 Comment on endpoint

Details regarding the gender, age, body composition, etc. are variable or unavailable for the half-life data

3.4. Endpoint units

Hours

3.5. Dependent variable

$\log_{10}$ HHLT

3.6. Experimental protocol

NA

3.7. Endpoint data quality and variability

The data were collected from aggregated sources such as online databases and review papers, and from the primary literature, as referenced in [1]. Because the data is collected from diverse sources the data quality is likely to be variable.

4. Defining the algorithm - OECD Principle 2: "AN UNAMBIGUOUS ALGORITHM"

4.1. Type of model

QSAR – multiple linear regression, ordinary least squares

4.2. Explicit algorithm

The predictions are calculated as the linear sum of fragment counts (f) and regression coefficients (a).

$\log_{10}$ HHLT = $a_0 + a_1f_1 + a_2f_2 + \ldots + a_nf_n$

4.3. Descriptors in the model

The descriptors are all molecular fragments specified as SMARTS strings. The inputs used in the algorithm shown in 4.2 are the counts of each fragment in a chemical structure. There are 63 fragments plus the intercept (a_0) in the model.

4.4. Descriptor selection

The preliminary descriptor pool is generated using custom code by recursively fragmenting the training dataset to obtain all possible molecular fragments, which typically number in the tens of thousands, a ratio of up to 100:1 vs. the number of training data. After consolidating fragment with colinear counts and applying other filters such as a cut off for correlation vs. the dependent variable the number of fragments in final descriptor pool is typically at a ratio of about 10:1 vs. the number of training data points.

4.5. Algorithm and descriptor generation

Fragments are selected from several sub-pools of descriptors proceeding iteratively from simple fragments to complex fragments. Fragments in each sub-pool are selected by iterative forward selection and backwards removal, with replacement of fragments in the model with fragments from the sub-pool also considered in the forward selection. Fragment selection and removal are chosen by selecting or removing the fragment which improves the goodness-of-fit (GoF) the most. GoF is the Akaike Information Criteria Corrected for dataset size (AICC), which

uses as input the predictive sum of squares (PRESS) calculated from a k-fold cross validation (typically k = 10).

4.6. Software name and version for descriptor generation

IFSQSAR custom development code, see 4.4 and 4.5.

4.7. Chemicals/Descriptors ratio

Ratio of training data to selected descriptors is 552:63, or approximately 9:1.

5. Defining the applicability domain - OECD Principle 3: "A DEFINED DOMAIN OF APPLICABILITY"

5.1. Description of the applicability domain of the model

Three methods of assessing Applicability Domain (AD) are applied simultaneously in IFSQSAR:

1. The leverage approach [2] which is essentially a metric that quantifies the distance in the model descriptor space between a chemical the training data.
2. Chemical similarity score (CSS) [3] which is an k-nearest neighbours approach (k=5) that also incorporates how well the model fits the training data of the nearest neighbours.
3. Other structural or property alerts [4]:
    a. None of the fragments in the model are in the chemical and the prediction is just the intercept.
    b. A check for atom or bond types not found in the training dataset.
    c. Boundary condition violations, i.e. for HHLT a minimum possible half-life is defined because the elimination of chemicals from the human body is limited by physiological parameters.

The first two methods are applied in a complementary way to define chemicals as in AD, borderline cases which are in AD but have more uncertainty, out of AD, and cases of egregious extrapolation which are very uncertain. These designations are overridden by alerts of the third method which identify very uncertain predictions.

The uncertainty of predictions is quantified as a 95% confidence factor (Cf). For example, Cf=2 means that the prediction is estimated to be within a factor 2 of the real value, at a confidence level of 95%. The uncertainty is estimated by evaluating the accuracy of predictions for chemicals in the external validation dataset.

5.2. Method used to assess the applicability domain

It is observed for this model and others that the uncertainty as quantified using the external validation dataset increases proceeding from: in AD, borderline, out AD, egregious extrapolation, missing atoms or bonds. This pattern of increasing validated uncertainty with increasingly dire AD warnings builds confidence that the AD methods are effective and robust.

5.3. Software name and version for applicability domain assessment

IFSQSAR custom code, see 5.1.

5.4. Limits of applicability

Boundaries for the leverage approach, where p' is the number of model parameters including the intercept, and n is the number of training data:

- Borderline: leverage > 0.174, defined as $1.5 \cdot p'/n$
- Out of AD: leverage > 0.348 defined as $3 \cdot p'/n$
- Egregious extrapolation: leverage > 1

Boundaries of the CSS approach where CSS is in the range 0 (no similarity) to 1 (very similar):

- Out of AD: CSS < 0.242, defined as lower CSS than 95% of the training data.
- Borderline: CSS < 0.379, defined as lower CSS than 75% of the training data.

A chemical is assigned in AD, borderline, out of AD, or egregious extrapolation based on the most conservative (poorest) result from the leverage and CSS approaches, so a chemical is only assigned in AD if both approaches assign a chemical as in AD.

Out of AD due to no fragment overlap: counts for all fragments in the model are zero.

Out of AD due to novel bonds or atoms: at least one bond or atom type not found in the training dataset.

Out of AD due to lower boundary violation: predicted HHLT less than 0.05 hours.

6. Defining goodness-of-fit and robustness (internal validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"

6.1. Availability of the training set

The data come from publicly available sources and are summarized in the supplemental information of the paper [1].

6.2. Available information for the training set

a) Chemical names: Yes

b) CAS numbers: No

c) SMILES: No

d) InChI codes: No

e) MOL files: No

f) Structural formula: No

g) Nanomaterials: No

h) test chemical purity: No

i) Any other structural information: No

6.3. Data for each descriptor variable for the training set

No, descriptor values are not included.

6.4. Data for the dependent variable for the training set

Yes, dependent variable values are included in the supplemental information of the paper [1].

6.5. Other information about the training set

The training dataset contains n=552 chemicals with $\log_{10}$ HHLT in the range -1.3 to 6.3.

6.6. Pre-processing of data before modelling

Data was transformed from HHLT to $\log_{10}$ HHLT.

6.7. Statistics for goodness-of-fit

r^2 = 0.887, RMSE = 0.445

slope= 0.881, intercept = 0.143 for correlation between fitted and expected $\log_{10}$ HHLT values.

6.8. Robustness - Statistics obtained by leave-one-out cross-validation

NA

6.9. Robustness - Statistics obtained by leave-many-out cross-validation

NA

6.10. Robustness - Statistics obtained by Y-scrambling

Not performed for this QSAR, but y-scrambling was done with 50 iterations while developing the algorithm and it was found that the average y-scrambled r^2 and RMSE were 0.086 and 1.10 compared to fitted values of 0.789 and 0.526 [3].

6.11. Robustness - Statistics obtained by bootstrap

NA

6.12. Robustness - Statistics obtained by other methods

AICC = -745.0

7. Defining predictivity (external validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"

7.1. Availability of the external validation set

The data come from publicly available sources and are summarized in the supplemental information of the paper [1].

7.2. Available information for the external validation set

a) Chemical names: Yes

b) CAS numbers: No

c) SMILES: No

d) InChI codes: No

e) MOL files: No

f) Structural formula: No

g) Nanomaterials: No

h) test chemical purity: No

i) Any other structural information: No

7.3. Data for each descriptor variable for the external validation set

No, descriptor values are not included.

7.4. Data for the dependent variable for the external validation set

Yes, dependent variable values are included in the supplemental information of the paper [1].

7.5. Other information about the external validation set

The external validation dataset contains n=552 chemicals with $\log_{10}$ HHLT in the range -1.1 to 5.8.

7.6. Experimental design of test set

The training data is initially seeded with the chemicals that have the highest and lowest expected values. Chemicals are then alternately added to the training and validation datasets until all chemicals are assigned to a dataset. When adding to the training dataset the chemical which is the least similar to chemicals already in the training dataset and most similar to chemicals already in the external validation dataset is selected. When adding to the external validation dataset the chemical which is the least similar to chemicals already in the external validation dataset and most similar to chemicals already in the training dataset is selected. More details about quantifying the similarity are available in [2], but the method is similar to the definition of CSS, see 5.1.

7.7. Predictivity - Statistics obtained by external validation

r^2 = 0.723, RMSE = 0.698

slope= 0.799, intercept = 0.216 for correlation between fitted and expected $\log_{10}$ HHLT values.

7.8. Predictivity - Assessment of the external validation set

The external validation dataset comprises 50% of the available data, and the methodology to assign chemicals to the training and external validation datasets ensures representative dependent variable and fragment count distributions in both datasets.

7.9. Comments on the external validation of the model

Comments on the external validation of the model: Add any other useful comments about the external validation procedure.

8. Providing a mechanistic interpretation - OECD Principle 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE"

8.1. Mechanistic basis of the model

The descriptors are all fragments (sub-structures) of the chemicals in the training dataset and therefore depending on the sign of their regression coefficients are interpretable. Fragment interpretation was done in the paper [1] for both HHLT and HHLB because the values are correlated, and the models contain similar fragments. Fragments can make chemicals more susceptible to whole body elimination (negative regression coefficients, lower HHLT and HHLB) or less susceptible to whole body elimination (positive regression coefficients, higher HHLT and HHLB). General observations from the paper [1] are that non-polar functional groups such as halogens, aromatic and olefinic structures increase HHLT and HHLB, while polar heteroatom functional groups tend to decrease the HHLT and HHLB.

8.2. A priori or a posteriori mechanistic interpretation

A posteriori.

8.3. Other information about the mechanistic interpretation

NA

9. Miscellaneous information

9.1. Comments

The inclusion in the model outputs of a quantitative uncertainty metric is useful for assessing the reliability of the predictions and for propagating uncertainty in applications of the QSAR predictions.

9.2. Bibliography

NA

9.3 Supporting information

NA