PROTOREACH

DOSSIER **QMRF**

www.protoqsar.com

Centro Europeo de Empresas Innovadoras (CEEI).
Parque Tecnológico de Valencia.
Avda. Benjamín Franklin 12, despacho 8.
46980 Paterna (Valencia, Spain)

info@protoqsar.com

+34 96 088 06 58

ProtoQSAR
Computational toxicology:
fast, economical and ethical

ProtoQSAR model for melting point QMRF

1. QSAR identifier

1.1. **QSAR identifier (title):**

ProtoQSAR model for melting point

1.2. **Other related models:**

None

1.3. **Software coding the model:**

ProtoQSAR proprietary software

https://protoqsar.com

2 General information

2.1. **Date of QMRF:**

March 2020

2.2. **QMRF author(s) and contact details:**

ProtoQSAR SL

+34 960880658

info@protoqsar.com

2.3. **Model developer(s) and contact details:**

[1] Sergi Gómez-Ganau

[2] Joel Roca-Martínez

[3] Stephen Jones Barigye

[4] Eva Serrano-Candelas

[5] Rafael Gozalbes

Contact e-mail: info@protoqsar.com

2.4. **Date of model development and/or publication:**

March 2020

2.5. **Reference(s) to main scientific papers and/or software package:**

PHYSPROP database from EPI suite (https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface).

2.6. **Availability of information about the model:**

All the information on ProtoQSAR melting point model is owned by ProtoQSAR SL.

2.7. **Availability of another QMRF for exactly the same model:**

None to date

3 Defining the endpoint - OECD Principle 1

3.1. Species:

N/A

3.2. Endpoint:

Physico-chemical properties: Melting point temperature.

3.3. Comment on endpoint:

Endpoint following the OECD Test No. 102: Melting Point/ Melting Range.

3.4. Endpoint units:

Melting point: ºC

3.5. Dependent variable:

Melting point

3.6. Experimental protocol:

The melting point is defined as the temperature at which the phase transition from the solid to the liquid state at atmospheric pressure takes place.

3.7. Endpoint data quality and variability:

The data for developing the model was extracted from the PHYSPROP public database. However, all the information related to the model is owned by ProtoQSAR.

4 Defining the algorithm - OECD Principle 2

4.1. Type of model:

Light Gradient Boosting Machine (LGBM) Regressor

4.2. Explicit algorithm:

The LGBM is a gradient boosting framework that uses tree-based learning algorithms.

4.3. Descriptors in the model:

- *nN*
- *ATS3m*
- *EEig02x*
- *nCp*
- *nArCOOH*
- *H-050*
- *N-072*
- *nF*
- *MATS1v*
- *EEig03d*
- *nCq*
- *nOHs*
- *O-057*
- *ATS2m*
- *MATS1p*
- *EEig05d*
- *nR=Cp*
- *C-025*
- *O-059*

4.4. Descriptor selection:

The descriptor selection is performed by eliminating non-variant descriptors, as well as filtering collinear descriptors (R2 > 0.9). Afterwards, by Recursive Feature Elimination (RFE) based on ridge regression (alpha 1000), the number of descriptors was reduced based on their correlation with the values of the independent variable.

4.5. Algorithm and descriptor generation:

Descriptors are calculated by an in-house software module in which these are implemented as described in: R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009.

4.6. Software name and version for descriptor generation:

ProtoQSAR proprietary software v1.0

4.7. Chemicals/Descriptors ratio:

Ratio: 6899 /19 = 363.11

5 Defining the applicability domain - OECD Principle 3

5.1. Method used to assess the applicability domain:

For the applicability domain estimation three methods are assessed, the Tanimoto value to check the structural similarity, the KDE (Kernel Density Estimation), as well as the Euclidean distance and the Leverage to check the descriptor values distribution.

5.2. Description of the applicability domain of the model:

The Tanimoto coefficient allows to compare the structural similarity of two chemical structures by computing a set of fingerprints for each chemical compound. A value from 0 to 1 is obtained, where 1 corresponds to identical structures and is closer to zero if they are very different.

The kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. It is a technique employed to estimate the unknown probability distribution of a random variable, based on a sample of points taken from that distribution.

The Euclidean distance is a measure of the separation between two points in Euclidean space. We can compute the distance of a descriptor value to the median of that descriptor in the training set and determine if it is inside the applicability domain or not.

The leverage of a compound measures the distance of this compound from the model experimental space (the structural centroid of the training set) and is a measure of its influence on the model.

5.3. Software name and version for applicability domain assessment:

ProtoQSAR proprietary software v1.0.

6 Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

The training set is comprised of 6899 compounds from a curated dataset of 9199 compounds.

CAS RN: No

Chemical Name: No

SMILES: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for the dependent variable for the training set:

Yes

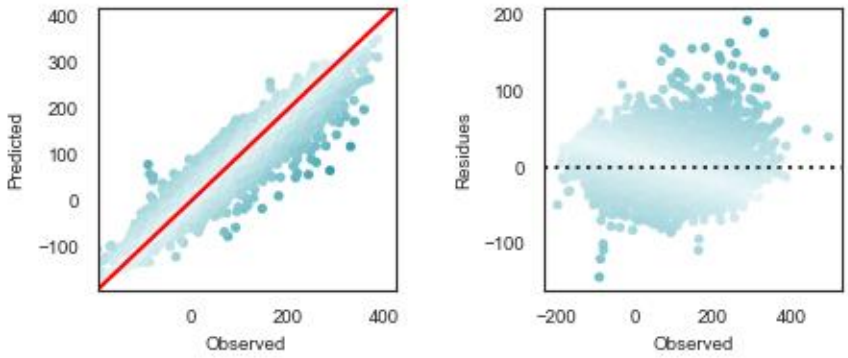
6.4. Other information about the training set:

6899 compounds (75% of the curated dataset) were randomly selected for the training set.

6.5. Pre-processing of data before modelling:

The experimental data of this dataset was curated following a standard procedure in order to guarantee its quality. Compounds with unclearly defined chemical structures were deleted, as well as inorganics compounds, metal complexes, salts containing organic polyatomic counterions, mixtures and substances of unknown or variable composition (UVCB). Also, duplicates and tautomers were checked. A further check of duplicates was done using the best tautomer to ensure that only one compound was present in the final dataset.

6.6. Statistics for goodness-of-fit:



6.7. Model performance

Parameters	Training	Validation
Explained variance	0.89	0.76
Mean absolute error (MAE)	23.74	36.76
Mean squared error (MSE)	1013.38	2395.75
Median absolute error	18.58	28.00
R2 score	0.89	0.76

7 External validation - OECD Principle 4

7.1. Availability of the external validation set:

The validation set is comprised of 2300 compounds from a curated dataset of 9199 compounds.

7.2. Available information for the external validation set:

CAS RN: No

Chemical Name: No

SMILES: Yes

Formula: No

INChI: No

MOL file: No

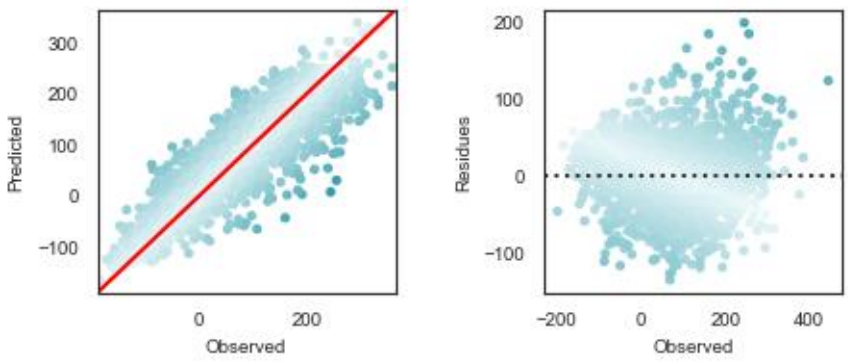
7.3. Data for the dependent variable for the external validation set:

Yes

7.4. Other information about the external validation set:

2300 compounds (25%) were randomly selected for the test set.

7.5. Predictivity - Statistics obtained by external validation:



7.6. Comments on the external validation of the model:

N/A

8 Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

The presented model identifies chemical structural features and physicochemical properties, which during the construction of the model were found be of relevance to melting point.

8.2. A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation. The identified chemical structural features and physicochemical properties may serve as starting point for a posteriori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation:

N/A

9 Miscellaneous information

9.1. Comments:

The model can be applied to estimate melting point.

ProtoPRED provides prediction for more than 25 endpoints, including physicochemical, toxicological and ecotoxicological, by using proprietary QSAR models. All ProtoPRED models meet OECD criteria and are valid for regulatory purposes.

9.2. Bibliography:

[1] OECD guideline: Test No. 102: Melting Point/ Melting Range. https://www.oecd-ilibrary.org/environment/test-no-102-melting-point-melting-range_9789264069527-en

[2] PHYSPROP database from EPI suite (https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface)

[3] R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley- VCH, 2009.

9.3. Supporting information:

N/A