	QMRF identifier (JRC Inventory): **Q17-14-0013**
	QMRF Title: **OPERA-model for** Vapor pressure
	Printing Date: **Apr 16, 2018**

1.QSAR identifier

1.1.QSAR identifier (title):

OPERA-model for Vapor pressure

1.2.Other related models:

No related models

1.3.Software coding the model:

OPERA V1.5

OPERA (OPEn (quantitative) structure-activity Relationship Application) is a standalone free and open source command line application. It provides a suite of QSAR models to predict physicochemical properties and environmental fate of organic chemicals based on PaDEL descriptors. It is available for download in Matlab, C and C++ languages from github under MIT license.

Kamel Mansouri (mansourikamel@gmail.com)

https://github.com/kmansouri/OPERA.git

PaDEL descriptors V2.21

Open source software to calculate molecular descriptors and fingerprints.

Chun Wei Yap (phayapc@nus.edu.sg)

http://padel.nus.edu.sg/software/padeldescriptor

MATLAB V8.2

MATrix LABoratory is a multi-paradigm numerical computing environment and fourth-generation programming language

http://www.mathworks.com/company/aboutus/contact_us/?s_tid=gn_cntus

http://www.mathworks.com/products/matlab/

2.General information

2.1.Date of QMRF:

1 November 2016

2.2.QMRF author(s) and contact details:

[1]Kamel Mansouri, ORISE research fellow at National Center for Computational Toxicology (NCCT), U.S. Environmental Protection Agency, mansourikamel@gmail.com

[2]Antony Williams, National Center for Computational Toxicology (NCCT), U.S. Environmental Protection Agency, Williams.Antony@epa.gov

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Kamel Mansouri, ORISE research fellow at National Center for Computational Toxicology (NCCT), U.S. Environmental Protection Agency, mansourikamel@gmail.com

2.6. Date of model development and/or publication:

2016

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

[1]An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling. 2016. Kamel Mansouri, Chris M. Grulke, Ann M. Richard, Richard S. Judson and Antony J. Williams. SAR & QSAR in Environ. Res; Vol. 27 , Iss. 11,2016. doi: 10.1080/1062936X.2016.1253611.

[2]OPERA: A free and open source QSAR tool for physicochemical properties and environmental fate predictions. Kamel Mansouri, Chris Grulke, Richard Judson, Antony Williams, Journal of Cheminformatics (2017)

[3]PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. Chun Wei Yap. (2011). J. Comput. Chem., 32: 1466–1474. doi:10.1002/jcc.21707 http://onlinelibrary.wiley.com/doi/10.1002/jcc.21707/abstract

[4]A KNIME workflow for chemical structures curation and standardization in QSAR modeling. Kamel Mansouri, Sherif Farag, Jayaram Kancherla, Regina Politi, Eugene Muratov, Denis Fourches, Nikolai Nikolov, Eva Bay Wedebay, Christopher Grulke, Ann Richard, Richard Judson, Alexander Tropsha. (in preparation)

[5]The influence of data curation on QSAR Modeling – examining issues of quality versus quantity of data (SOT). Williams, A., K. Mansouri, A. Richard, AND C. Grulke. Presented at Society of Toxicology, New Orleans, LA, March 13 - 17, 2016. https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=311418

[6]An Online Prediction Platform to Support the Environmental Sciences (American Chemical Society). Richard, A., C. Grulke, K. Mansouri, R. Judson, AND A. Williams. Presented at ACS Spring Meeting, San Diego, CA, March 13 - 17, 2016. https://cfpub.epa.gov/si/si_public_record_Report.cfm?dirEntryId=311655

[7]The importance of data curation on QSAR Modeling: PHYSPROP open data as a case study. Kamel Mansouri, Christopher Grulke Ann Richard Richard Judson Antony Williams. Presented at QSAR2016 14 June 2016, Miami, FL http://www.qsar2016.com/program

[8]Mansouri K. (2017) OPERA: A QSAR tool for physicochemical properties and environmental fate predictions. doi: 10.6084/m9.figshare.4836428 https://figshare.com/articles/OPERA_A_QSAR_tool_for_physicochemical_properties_and_environmental_fate_predictions/4836428

2.8. Availability of information about the model:

Non-proprietary suite of QSAR models freely available as a command line standalone application (OPERA: OPEn saR App) from github under MIT license: https://github.com/kmansouri/OPERA.git. Its predictions for the full DSSTox 720k chemicals are published on the EPA CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard). Training and validation sets are available for visualization on the dashboard and as SDF files provided in supporting information Section 9.3 and from the paper [ref 1-2, Section 2.7]. (ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/PHYSPROP_Analysis)

2.9. Availability of another QMRF for exactly the same model:

Not to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Not applicable

3.2.Endpoint:

1.Physical Chemical Properties 1.4.Vapour pressure

3.3.Comment on endpoint:

The vapor pressure of a liquid is the pressure exerted by its vapor
when the liquid and vapor are in dynamic equilibrium. This property is the
vapor pressure (in mm/Hg) predicted at 25°C

3.4.Endpoint units:

Logmm/Hg

3.5.Dependent variable:

VP

3.6.Experimental protocol:

The experimental data were downloaded from the EPI Suite data
webpage (http://esc.syrres.com/interkow/EpiSuiteData.htm).
These data are from PHYSPROP (The Physical Properties
Database) which is a collection of a wide variety of sources built by
Syracuse Research Corporation (SRC). Experimental protocols of the
different parts of data can be traced back to the original referenced
literature from the database.

3.7.Endpoint data quality and variability:

The original data collected from the PHYSPROP
database (3037 chemicals) have undergone a series of processes to curate
the chemical structures and remove duplicates, obvious outliers and
erroneous entries. This procedure also included a consistency check to
ensure only good quality data is used for the development of the QSAR
model (2840 chemicals).

Then, QSAR-ready structures were generated by
standardizing all chemical structures and removing duplicates, inorganic
and metallo-organic chemicals (2716 chemicals). The descriptions of
KNIME workflows that were developed for the purpose of the cleaning and
standardization of the data are available in the papers [ref 1 and ref 4
Section 2.7].

The curated outlier-free experimental data (2716
chemicals) was divided into training and validation sets before the
machine learning and modeling steps.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR/QSPR model derived derived from weighted nearest neighbors
algorithm (kNN) based on PaDEL descriptors [ref2 Sect 1.3].

4.2.Explicit algorithm:

Distance weighted k-nearest neighbors (kNN). k=5
This is a refinement of the classical k-NN classification algorithm where the contribution of each of

the k neighbors is weighted according to their distance to the query point, giving greater weight to closer neighbors.The used distance is the Euclidean distance. kNN is an unambiguous algorithm that fulfills the transparency requirements of OECD principle 2 with an optimal compromise between model complexity and performance.

4.3.Descriptors in the model:

[1]MLFER_S, Unitless, Molecular linear free energy relation: Combined dipolarity/polarizability. Platts JA, Butina D, Abraham MH, Hersey A. Estimation of molecular free energy relation descriptors using a group contribution approach. J Chem Inf Comput Sci. 1999;39(5):835-45.

[2]SHBd, Unitless, Atom type electrotopological state: Sum of E-States for (strong) hydrogen bond donors. Hall, L. H., and Kier, L. B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 35, 1039-1045; Liu, R., Sun, H., and So, S. S. (2001). Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. J Chem Inf Comput Sci 41, 1623-1632.; Gramatica, P., Corradi, M., and Consonni, V. (2000). Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. Chemosphere 41, 763-777.

[3]TopoPSA, Unitless, Topological polar surface area: Topological polar surface area. Ertl, P. and Rohde, B. and Selzer, P., Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties, J. Med. Chem., 2000, 43:3714-3717

[4]nHBDon, Unitless, Hbond donor count: Number of hydrogen bond donors (using CDK HBondDonorCountDescriptor algorithm).

[5]MLFER_L, Unitless, Molecular linear free energy relation: Solute gas-hexadecane partition coefficient. Platts JA, Butina D, Abraham MH, Hersey A. Estimation of molecular free energy relation descriptors using a group contribution approach. J Chem Inf Comput Sci. 1999;39(5):835-45.

[6]MLFER_E, Unitless, Molecular linear free energy relation: Excessive molar refraction. Platts JA, Butina D, Abraham MH, Hersey A. Estimation of molecular free energy relation descriptors using a group contribution approach. J Chem Inf Comput Sci. 1999;39(5):835-45.

[7]AATSC0v, Unitless, Average centered Broto-Moreau autocorrelation - lag 0 / weighted by van der Waals volumes. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 27-37

[8]nssO, Unitless, Atom type electrotopological state: Count of atom-type E-State: -O-.Hall, L. H., and Kier, L. B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 35, 1039-1045; Liu, R., Sun, H., and So, S. S. (2001). Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. J Chem Inf Comput Sci 41, 1623-1632.; Gramatica, P., Corradi, M., and Consonni, V. (2000). Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. Chemosphere 41, 763-777.

[9]MDEC-23, Unitless, Molecular distance edge: Molecular distance edge between all secondary and tertiary carbons. Liu, S. and Cao, C. and Li, Z. , Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance Edge (MDE) Vector, lambda, Journal of Chemical Information and Computer Sciences, 1998, 38:387-394

[10]nHBAcc_Lipinski, Unitless, Hbond acceptor count: Number of hydrogen bond acceptors (using Lipinski's definition: any nitrogen; any oxygen).

[11]piPC7, Unitless, Path counts: Conventional bond order ID number of order 7. Todeschini, R. and

Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 574-579, pg 395-402

[12]EE_Dt, Unitless, Estrada-like index from detour matrix. Todeschini, R. and Consonni, V. (2009). Molecular descriptors for chemoinformatics, (Weinheim: Wiley VCH) pg 714-726

4.4. Descriptor selection:

PaDEL software was used to calculate 1440 molecular descriptors. A first filter was applied in order to remove descriptors with missing values, constant and near constant (standard deviation of 0.25 as a threshold) and highly correlated descriptors (96% as a threshold). The remaining 770 descriptors were used in a feature selection procedure to select a minimum number of variables encoding the most relevant structural information to the modeled endpoint. This step consisted of coupling Genetic Algorithms (GA) with the weighted kNN algorithm and was applied in 5 fold cross validation on the training set (2034 chemicals). This procedure was run for 200 consecutive independent runs maximizing Q^2 in cross-validation and minimizing the number of descriptors. The number of k neighbors is optimized within the range of 3 to 7. The descriptors were then ranked based on their frequency of selection during the GA runs. The best model showed an optimal compromise between the simplicity (minimum number of descriptors) and performance (Q^2 in cross-validation) to ensure transparency and facilitate the mechanistic interpretation as required by OECD principles 2 and 5. More details in paper [ref2 Section 2.7].

4.5. Algorithm and descriptor generation:

PaDEL descriptors were calculated based on two-dimensional (2D) chemical structures generated by the Indigo cheminformatics suite of tools implemented in KNIME. 2D descriptors were selected over 3D to avoid complicated and usually irreproducible geometrical optimizations. The calculated descriptors fall into different groups such as constitutional indices, ring descriptors, topological indices, 2D matrix based descriptors, functional group counts and atom counts. Details and references provided in Section 4.3.

4.6. Software name and version for descriptor generation:

PaDEL-Descriptors V2.21

An open source software to calculate molecular descriptors and fingerprints.

Chun Wei Yap (phayapc@nus.edu.sg)

http://padel.nus.edu.sg/software/padeldescriptor

4.7. Chemicals/Descriptors ratio:

2034 chemicals (trainingset)/12 descriptors= 169.5

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

The model is applicable to heterogeneous organic chemicals. In the implementation of the model several pieces of information are given to help the user in evaluating the reliability of a prediction. The chemical structure is first assessed to see if it is falling within the

Applicability Domain of the model or not. Then the accuracy of the predicted value is reported based on the similarity of the query chemical to its neighboring chemicals in the training set of the model.

This fullfills the requirements of the 3rd OECD principle by defining the limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the model can generate reliable predictions.

5.2. Method used to assess the applicability domain:

The applicability domain of the model is assessed in two independent levels using two different distance-based methods. First, a global applicability domain is determined by means of the leverage approach that checks whether the query structure falls within the multidimensional chemical space of the whole training set.

The leverage of a query chemical is proportional to its Mahalanobis distance measure from the centroid of the training set. The leverages of a given dataset are obtained from the diagonal values of the hat matrix. This approach is associated with a threshold leverage that corresponds to $3*p/n$ where p is the number of model variables while n is the number of training compounds. A query chemical with leverage higher than the threshold is considered outside the AD and can be associated with unreliable prediction.

The leverage approach has specific limitations, in particular with respects to gaps within the descriptor space of the model or at the boundaries of the training set. To obviate such limitations, a second tier of applicability domain assessement was added. This comprised a local approach which only investigated the vicinity of the query chemical. This local approach provides a continuous index ranging from 0 to 1 which is different from the first approach which only provides Boolean answers (yes/no). This local AD-index is relative to the similarity of the query chemical to its 5 nearest neighbors in the p dimensional space of the model. The higher this index, the more the prediction is likely to be reliable.

5.3. Software name and version for applicability domain assessment:

Implemented in OPERA V1.5

An implementation of a local similarity index and the leverage approach based on the work of Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 2012, 17, 4791-4810.

Kamel Mansouri (mansourikamel@gmail.com);
https://github.com/kmansouri/OPERA.git

5.4. Limits of applicability:

These two AD methods described in Section 5.2 are complementary and can be interpreted in the following way:

- If a chemical is considered outside the global AD with a low local AD-index, the prediction can be unreliable
- If a chemical is considered outside the global AD but the local

AD-index is average or relatively high, this means the query chemical is
on the boundaries of the training set but has quite similar neighbors.
The prediction can be trusted.
- If a chemical is considered inside the global AD but the local
AD-index is average or relatively low, this means the query chemical
fell in a "gap" of the chemical space of the model but still within the
boudaries of the training set and surrounded with training chemicals.
The prediction should be considered with caution.
- If a chemical is considered inside the global AD with a high
local AD-index, the prediction should be considered reliable.
Even though the applicability domain is necessary to set the
limits of the interpolation space of the model, it doesn't necessarily
inform about the quality of the prediction especially in the empty
spaces and around the edges of the descriptor space. In order to
overcome this limitation and help the user decide about the reliability
of a prediction, we added a confidence level index ranging from 0 to 1
relative to the accuracy of prediction of the 5 nearest neighbors to the
query chemical. The higher this index, the more the prediction is likely
to be reliable.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:
Yes

6.2.Available information for the training set:
Internal ID; CAS checksum; name validity; preferred name; IUPAC name; Original SMILES; QSAR-
ready canonical smiles; InChI; Salt information; DSSTox GSID; Experimental reference; Consistency
flag
CAS RN: Yes
Chemical Name: Yes
Smiles: Yes
Formula: No
INChI: Yes
MOL file: Yes

6.3.Data for each descriptor variable for the training set:
All

6.4.Data for the dependent variable for the training set:
All

6.5.Other information about the training set:
The training set consists of 2034 chemicals. The
structures are randomly selected to represent 75% of the available data
keeping a similar normal distrubution of VP vlaues in both training and
test sets using the Venetian blinds method. The values are ranging from
~-13 to ~-5. A plot of the distribution of VP values is provided in the
supporting information Section 9.3.

6.6.Pre-processing of data before modelling:

No preprocessing of the values.

6.7.Statistics for goodness-of-fit:

Performance in training:

R^2=0.91

RMSE=1.08

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

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

Performance in 5-fold cross-validation:

Q^2=0.91

RMSE=1.08

A plot of the experimental versus predicted values

for the training set is provided in supporting information Section 9.3.

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

Internal ID; CAS checksum; name validity; preferred name; IUPAC name; Original SMILES; QSAR-ready canonical smiles; InChI; Salt information; DSSTox GSID; Experimental reference; Consistency flag

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: Yes

MOL file: Yes

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

All

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

All

7.5.Other information about the external validation set:

The validation set consists of 679 chemicals.

The values are ranging from ~-11 to ~5.

7.6.Experimental design of test set:

The structures are randomly selected to represent

25% of the available data keeping a similar normal distrubution of VP

vlaues in both training and test sets using the Venetian blinds method.

A plot of the distribution of VP values is provided in the supporting

information Section 9.3.

7.7. Predictivity - Statistics obtained by external validation:

Performance in test:

R^2=0.92

RMSE=1

7.8. Predictivity - Assessment of the external validation set:

The validation set consisting of 679 chemicals
which is equivalent to a third (1/3) of the training set is sufficient
for the evaluation of the predictivity of the model and a good
representation of the chemical space as shown in the multi-dimensional
scaling plot provided in supporting information Section 9.3. A plot of
the experimental versus predicted values for the validation set is
provided in supporting information Section 9.3.

7.9. Comments on the external validation of the model:

The choice of proportions between the training set and the validation
set as well as the splitting method helped in accurately evaluating the
model and covering most of the training set chemical space. This goal
was accomplished without the need to do a structural sampling that
usually shows over-optimistic evaluation of the predictivity or a
complete random selection that risks biasing the evaluation towards a
certain region of the chemical space.

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

The model descriptors were selected statistically but they can also be
mechanistically interpreted.
Vapor pressure is one of the key parameters in the assessment of fate
and distribution of environmental pollutants in various phases (air,
water, and soil) and is as important as water solubility, air-water
partition coefficient and octanol-water partition coefficient.This physico-chemical property provides
an indication of the attractive
forces between the molecules that are directly related to the structure of
the compound, and thus can be correlated to the structure [Liang 1998,
Basak and Mills 2005, and Oberg and Liu 2007].The stronger the
intermolecular forces are, the more tightly the molecules are held
together in a condensed phase and, hence, the lower the vapor pressure
will be [Schwarzenbach et al. 1993]. Molecular interactions in most
nonionic liquids can be divided into three types: dipole-dipole,
dipole-induced dipole, and induced dipole-induced dipole or dispersion
forces (as function of the molecule's polarizability) [Atkins 1978].Therefore, topological descriptors,
including molecular connectivity
indices, electrotopological state descriptors, general polarity
descriptors, and hydrogen bonding descriptors can be used to model
intramolecular forces and predict VP. Such descriptors included in our
model for VP prediction are:
SHBd: Atom type electrotopological state: Sum of E-States for
(strong) hydrogen bond donors.

nHBDon: Hbond donor count: Number of hydrogen bond donors (using
CDK HBondDonorCountDescriptor algorithm).

nHBAcc_Lipinski: Hbond acceptor count: Number of hydrogen bond
acceptors (using Lipinski's definition: any nitrogen; any oxygen).

MLFER_S: Molecular linear free energy relation: Combined
dipolarity/polarizability.

MLFER_E: Molecular linear free energy relation: Excessive molar
refraction.

TopoPSA: Topological polar surface area: Topological polar surface
area.

AATSC0v: Average centered Broto-Moreau autocorrelation - lag 0 /
weighted by van der Waals volumes.MLFER_L: Molecular linear free energy relation: Solute
gas-hexadecane partition coefficient.

nssO: Atom type electrotopological state: Count of atom-type
E-State: -O-.HallMDEC-23: Molecular distance edge: Molecular distance edge between
all secondary and tertiary carbons.piPC7: Path counts: Conventional bond order ID number of
order 7.

EE_Dt: Estrada-like index from detour matrix.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation.

8.3.Other information about the mechanistic interpretation:

For more details and full reference, see
references in Section 4.3 and Section 9.2.

9.Miscellaneous information

9.1.Comments:

This QSAR model for VP prediction is part of the
NCCT_Models Suite that is a free and open-source standalone application
for the prediction of physicochemical properties and environmental fate
of chemicals. This application is available in the Supporting
information Section 9.3 of this report and in the paper ref 2 Section 2.7.
The detailed results of this suite of models applied on more than 700k
DSSTox chemicals are available on the iCSS chemistry dashboard
(https://comptox.epa.gov/dashboard).

This current version of the model is mainly based
on curated and standardized data collected from the Physprop database.
All NCCT_Models are designed to fulfil the requirement of the 5 OECD
principles to ensure transparency and reproducibility of the results. In
order to predict new chemicals, the models only require 2D chemical
structures that are used to calculate molecular descriptors by PaDEL
2.21 software. Then a simple weighted kNN algorithm is used to make the
prediction based on the observed values of the k closest molecules. All
models showed high robustness and statistics stability between training,
5-fold cross-validation and the external validation set.

Considering the full applicability domain of the 2716
chemicals with available data and the same models parameters described
earlier, the calibration statistics would be an R ^{2} of 0.9 and
an RMSE of 1.11.

9.2.Bibliography:

[1]Liang C and Gallagher DA (1998) QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors. Journal of Chemical Information and Computer Sciences. 38 (2) 321-324 DOI: 10.1021/ci970289c

[2]Basak SC, Mills D (2005) Development of quantitative structure-activity relationship models for vapor pressure estimation using computed molecular descriptors Volume 2005, Issue 10, Commemorative Issue in Honor of Prof. Alexandru T. Balaban on the occasion of his 75th anniversary, 308-320 http://dx.doi.org/10.3998/ark.5550190.0006.a23

[3]Öberg T and Liu T (2008) Global and Local PLS Regression Models to Predict Vapor Pressure. QSAR Comb. Sci. 27, 273–279 doi:10.1002/qsar.200730038

[4]Schwarzenbach RP, Gschwead PM, Imboden DM (1993) Environmental Organic Chemistry. John Wiley & Son, Inc.: New York

[5]Atkins RW (1978) Physical Chemistry; W. H. Freeman Company: San Francisco

9.3.Supporting information:

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-14-0013

10.2.Publication date:

2017-09-21

10.3.Keywords:

Vapor pressure;PaDEL;OPERA;

10.4.Comments:

To be entered by JRC