Transparency in modeling through careful application of OECD's QSAR/QSPR principles via a curated water solubility data set

Charles N. Lowe[1], Nathaniel Charest[2], Christian Ramsland[2], Daniel T. Chang[1], Todd M. Martin[1], and Antony J. Williams[1]

1. Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA

2. ORAU Student Services Contractor to Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA

Table of Contents:

QMRF

1) QSAR Identifier
    1) Title
       EPA’s Computational Chemistry and Cheminformatics Branch Water Solubility Model
    2) Related Models
       No related models.
    3) Software Coding Model
       R Studio Version 1.4.1717 A language and environment for statistical computing.
       https://www.R-project.org/
       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
2) General Information
    1) Date of QMRF
       10/18/2022
    2) QMRF Authors & Contact Information
       Nathaniel Charest, ORAU research fellow at the Center for Computational Toxicology and Exposure, US Environmental Protection Agency, charest.nathaniel@epa.gov
       Charles Lowe, Chemist at Center for Computational Toxicology and Exposure, US Environmental Protection Agency, lowe.charles@epa.gov
    3) Date of QMRF Updates
    4) Summary of QMRF Updates
    5) Model Developers & Contact Information
       Charles Lowe, Chemist at Center for Computational Toxicology and Exposure, US Environmental Protection Agency, lowe.charles@epa.gov
       Nathaniel Charest, ORAU research fellow at the Center for Computational Toxicology and Exposure, US Environmental Protection Agency, charest.nathaniel@epa.gov
    6) Date of Model Development & Publication
       2022
    7) Primary Related Publications
       TBD
3) Dataset Information
    1) Data Curation Strategy
    2) List of Datasets & Availability
        i. eChemPortal (ECHA 2022) is a database of physicochemical properties and toxicity measurements provided by the OECD.
        ii. Advanced Digital Design of Pharmaceutical Therapeutics (ADDoPT 2022) is a collaboration between pharmaceutical companies and academia to establish digital design approaches more usable for drug discovery.
        iii. AqSolDB (Sorkun 2019) is a data collection resulting from work described in a publication by Sorkun et al.
        iv. The Bradley dataset (Bradley 2015) is a collection of measured solubilities from the Open Notebook Science Challenge.

v. The Online chemical modeling environment (Sushko 2011) is a physical property database and modeling platform with data sets provided by the users.

vi. LookChem (LookChem 2022) is a global chemical trading platform which includes physical property values for advertised chemicals. A caveat with this source is that each entry lacks a citation, thus it was difficult to rectify if physicochemical properties were really measured or predicted.

vii. QSAR DataBank (Ruusmann 2015) is a repository of QSA/PR models and associated data following the FAIR (Findable, Accessible, Interoperable, Re-usable) principles.

viii. PubChem (Kim 2021) is an open chemistry database developed by the National Institutes of Health.

ix. The OPEn structure–activity/property Relationship App (OPERA) (Mansouri 2018) is a collection of models and associated data developed by Mansouri et al.

4) OECD Principle 1 – Defining The Endpoint
   1) Species
      i. Not applicable
   2) Endpoint
      i. OECD Physical Chemical Properties 1.3 Water Solubility
   3) Comments on Endpoint
      i. The solubility is the maximum amount of a solute than can be dissolved in a substance at a given temperature. This model predicts solubility for chemicals measured between 20 – 30 degrees Celsius. Predicted solubility is expected to be approximate within this temperature range.
   4) Endpoint Units
      i. Base-10 Logarithm (Moles/Liter) [Log(M)]
   5) Dependent Variable
      i. Water solubility
   6) Theoretical Description of Endpoint
      i. Water solubility has a well-understood mechanism of emergence from straight-forward statistical physics. It is an equilibrium quantity derived from comparing the energetics of a solute molecule being surrounded by solvent molecules (solvated phase) versus the energetics of the solute molecule being surrounded by other solute molecules (solute phase). The transition is a single-step process governed by passage of a molecule through the interface between the bulk and the solvent. There are not multiple mechanisms relating structural features to the endpoint.
5) OECD Principle 2 – Defining The Algorithm
   1) Structural Representation
      i. PaDEL descriptors [ref] were selected based on expert judgement of their relationship to capturing the solvation energetics
   2) Descriptors In Model
      i. XLogP

1. Additive SAR model predicting the water-octanol partition coefficient. Thermodynamically relatable to water-bulk partition coefficient, thus relating to water solubility.
2. Ref

ii. SIC0
    1. Structural information content index. Briefly, captures atomic diversity in molecular graph normalized by number of atoms. Relatable to structural complexity and potential interactions with solvent.
    2. Roy, Basak, Harriss, Magnuson. “Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications”. Mathematical Modeling In Science And Technology. Fourth International Conference. 1983.

iii. ZMIC3
    1. Z-Modified information content index, order 3. Briefly, captures diversity of 3rd order atomic connectivity within the molecular graph. This can encode common functional groups with relevance to water solubility, such as amine or alcohols.
    2. King, J. A Z-Weighted Information Content Index. *Int. J. of Quan. Chemistry.* 1989.

iv. piPC7
    1. Path count of 7th-order pi-conjugation in the molecular graph. High degrees of pi-conjugation allow for greater induction effects than can affect polar moments or polarizability, thus influencing solute-solute and solvent-solute interactions.

v. piPC5
    1. Path count of 5th-order pi-conjugation in molecular graph. Captures benzene rings. High degrees of pi-conjugation will affect the polarizability of the molecule and influence the energetics of aqueous solvation.
    2.

vi. nAcid
    1. Number of acidic protons. The ability to form acid-base pairs directly influences the charge state of the aqueous species and therefore affects the energetics of solvation.

vii. nHBAcc
    1. Number of hydrogen bond acceptors. The ability to interact with the hydrogen bond network will influence the energetics of solvation.

viii. nHBDon
    1. Number of hydrogen bond donors. The ability to interact with the hydrogen bond network will influence the energetics of solvation.

ix. GATS1s
    1. Geary autocorrelation coefficient, lag one, weighted by Gasteiger charge. Spatial autocorrelation of atomic charge separated by one bond. Captures local motifs that affect the energetics of solvent interaction.

x. GATS1m
    1. Geary autocorrelation coefficient, lag one, weighted by mass. Spatial autocorrelation of weighted by atomic mass separated by one bond. Captures local motifs that affect the energetics of solvent interaction.

xi. GATS1e
    1. Geary autocorrelation coefficient, lag one, weighted by electronegativity. Spatial autocorrelation of weighted by electronegativity separated by one bond. Captures local motifs that affect the energetics of solvent interaction.

xii. GATS1p
    1. Geary autocorrelation coefficient, lag one, weighted by polarizability. Spatial autocorrelation of weighted by polarizability separated by one bond. Captures local motifs that affect the energetics of solvent interaction.

xiii. GATS1i
    1. Geary autocorrelation coefficient, lag one, weighted by ionization potential. Spatial autocorrelation of weighted by ionization potential separated by one bond. Captures local motifs that affect the energetics of solvent interaction.

xiv. GATS2e
    1. Geary autocorrelation coefficient, lag one, weighted by electronegativity. Spatial autocorrelation of weighted by electronegativity separated by two bonds. Captures local motifs that affect the energetics of solvent interaction. Longer range separation of electronegativities could embed polar moments that can interact with the aqueous solvent.

xv. GATS1v
    1. Geary autocorrelation coefficient, lag one, weighted by van der Waals volume. Spatial autocorrelation of van der Waals volume separated by one bond. Captures local motifs of atomic arrangements.

xvi. GATS1c
    1. Geary autocorrelation coefficient, lag one, weighted by gasteiger charge Spatial autocorrelation of charge separated by one bond. Captures separation of charges across single bonds, which may influence interactions with a polar solvent like water.

xvii. nHBAcc
    1. Number of hydrogen bond acceptors. Interaction with the hydrogen bond network of the aqueous solvent is expected to impact the energetics of solvation.

xviii. nHBDon
    1. Number of hydrogen bond donors. Interaction with the hydrogen bond network of the aqueous solvent is expected to impact the energetics of solvation.

xix. nAcid

                1. Number of acidic groups. Acidic groups that can deprotonate and form formal charges

    3) Descriptor Selection

        i. Descriptors were first decorrelated by identifying descriptors with Spearman correlations of 0.95 or above. Pairs were resolved by taking the descriptor with the higher correlation with the endpoint. This process was implemented by Caret, the R package. Descriptors were then selected first by training a random forest algorithm and determining the importance assigned by the algorithm to each descriptor. The top 16 most important descriptors were reviewed and reconciled with chemical intuition, with expert judgements and replacements made where another descriptor was considered to be more holistically representative of the energetics of solvation.

    4) Descriptor Calculation Software

        i. Descriptors were calculated by OPERA v2.7 software by Kamel Mansouri & Antony Williams (Mansouri & Williams 2018). Inputs to the software are SMILES strings that were standardized to be QSAR ready by the Hazard Comparison Dashboard's standardizer (ref). These standardized SMILES were then passed to OPERA for descriptor calculation.

    5) Regression Algorithm

        i. Random Forest

            1. Ensemble method based on Breiman, 2004. 1000 decision trees are trained. Each tree is exposed to a data set bootstrapped with replacement from the total training set. At each split in a decision tree, 4 descriptors are considered. Splits are determined based on minimization of the variance of the dependent variable in each leaf. This process is repeated, growing trees to unrestricted depth with no pruning. The model's final estimate is the arithmetic average of all decision trees' predictions.

    6) Software Details

    7) Chemicals/Descriptors Ratio

        i. 8037 chemicals / 16 descriptors = 502.31

6) Defining The Applicability Domain

    1) Qualitative Description

        i. The data used to train the models overwhelmingly represents the space of small organic molecules with primarily light atoms as composition. Experimental results reported in the publication text show some transferability of learning between small organic functional groups when they are excluded, however significant failure is shown when heavier atomic functional groups like metallics or metalloids are considered.

    2) Quantitative Description

        i. The OPERA Local Index was applied to quantitatively compare a similarity index to the performance of the model. There is a roughly linear decline in root mean squared error as compounds with increasingly high similarity indices are excluded. The publication text illustrates the full detail.

3) Limits of Applicability
    i. This model is not recommended for molecules with metallic or metalloid atoms. Based on the usage, the OPERA local index can be used to approximate the expected performance of the model based on a compound's similarity to nearest neighbors in the training set.

7) Internal Validation
    1) Training Set Availability
        i. Yes
    2) Training Set Identifiers
    3) Descriptor Available
        i. Yes
    4) Endpoints Available
        i. Yes
    5) Training Set Construction
        i. Training set was constructed via a stratified splitting to capture 8037 of 10207 chemicals from the complete pool (~79%).
    6) Internal Statistics
        i. Performance on Training Set
            1. 0.97 Coefficient of Determination
            2. 0.41 Root-mean-squared Error
        ii. Leave-many-out on Training Set
            1. 5-fold Cross Validation
                a. 0.82 Average of 5 folds
                b. 0.96 Root-mean-squared Error
                c. Plot of predictions is available in Section 10
        iii. Out-of-bag estimation
            1. 0.81 Coefficient of Determination
            2. 0.98 Root-mean-squared Error

8) External Validation
    1) Testing Set Availability
        i. Yes
    2) Testing Set Identifiers
    3) Descriptors Available
        i. Yes
    4) Endpoints Available
        i. Yes
    5) Testing Set Construction
        i. Testing set was constructed as the remaining set after the training set was selected using the procedure outlined in section 7.5
    6) External Statistics
        i. 0.82 Coefficient of Determination
        ii. 0.97 Root-mean-squared Error
    7) Comments on Predictivity

i. The agreement between internal validation statistics and external statistics suggests the model is relatively stable within the chemical space covered by the training set. This is reinforced by the behavior of the performance statistics versus the OPERA Local Index thresholds used to characterize the applicability domain.

ii. The model is not suitable for compounds outside chemical space that is covered by the training set.

9) Mechanistic Interpretation

1) Each of the descriptors passed to the model were selected for their ability to parameterize structural elements that can contribute to the energetics of solvation. Detailed relating of these descriptors to the energetics of solvation is discussed in section 5.2. Random forest regressors derive internal forms of similarity by reducing variance of the endpoint in clusters after performing splits on descriptors. Because these descriptors are related to the energetics of solvation, it is expected that the decision tree clustering process will result in clusters with similar structural contributions to the energetics determining the ratio of molecules in the bulk versus the solvated phase. Thus, the clusters are expected to capture structural relationships to the water solubility based on first principle arguments.

   Individual decision trees are exposed to different samplings of the training set based on bootstrapping with replacement. This results in differing predictions of most similar training compound by differing trees, mitigating overfitting and better abstracting higher-level patterns in the descriptors that result in similar solvation energetics.

2) This mechanism is a combined a prior and a posteriori mechanism. Descriptors were initially found by a machine learning approach, requiring a posteriori interpretation; however, based on expert judgement certain descriptors were exchange based on a priori recommendations about the relevance of descriptors to the energetics of solvation.

Bibliography

ADDoPT, Advanced Digital Design of Pharmaceutical Therapeutics. 2022. *Advanced Digital Design of Pharmaceutical Therapeutics.* https://www.addopt.org/.

Bradley, Jean-Claude and Abraham, Michael H and Acree, William E and Lang, Andrew SID and Beck, Samantha N and Bulger, David A and Clark, Elizabeth A and Condron, Lacey N and Costa, Stephanie T and Curtin, Evan M and others. 2015. "Determination of Abraham model solute descriptors for the monomeric and dimeric forms of trans-cinnamic acid using measured solubilities from the Open Notebook Science Challenge." *Chemistry Central Journal* 9: 1-6. doi:10.1186/s13065-015-0080-9.

ECHA, European Chemicals Agency. 2022. *Registration Dossier.* Accessed 2022. https://echa.europa.eu/registration-dossier/.

Kim, Sunghwan and Chen, Jie and Cheng, Tiejun and Gindulyte, Asta and He, Jia and He, Siqian and Li, Qingliang and Shoemaker, Benjamin A and Thiessen, Paul A and Yu, Bo and others. 2021.

"PubChem in 2021: new data content and improved web interfaces." *Nucleic acids research* 49: 1388-1395. doi:10.1093/nar/gkaa971.

LookChem. 2022. *https://www.lookchem.com/about/about.htm.* Accessed 2022. https://www.lookchem.com/.

Mansouri, Kamel and Grulke, Chris M and Judson, Richard S and Williams, Antony J. 2018. "OPERA models for predicting physicochemical properties and environmental fate endpoints." *Journal of cheminformatics* 10: 1-19. doi:10.1186/s13321-018-0263-1.

Ruusmann, Villu and Sild, Sulev and Maran, Uko. 2015. "QSAR DataBank repository: open and linked qualitative and quantitative structure--activity relationship models." *Journal of Cheminformatics* 7 (1): 1-11. doi:10.1186/s13321-015-0082-6.

Sorkun, Murat Cihan, Abhishek Khetan, and Süleyman Er. 2019. "AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds." *Scientific data* 6 (1): 1-8. doi:10.1038/s41597-019-0151-1.

Sushko, Iurii and Novotarskyi, Sergii and Korner, Robert and Pandey, Anil Kumar and Rupp, Matthias and Teetz, Wolfram and Brandmaier, Stefan and Abdelaziz, Ahmed and Prokopenko, Volodymyr V and Tanchuk, Vsevolod Y and others. 2011. "Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information." *Journal of computer-aided molecular design* 25 (6): 533-554. doi:10.1007/s10822-011-9440-2.