ORIGINAL SCIENTIFIC PAPER

CROATICA CHEMICA ACTA

Croat. Chem. Acta 2020, *93*(4), 311–319
Published online: September 2, 2021
DOI: 10.5562/cca3776

HKD 1926

QSPR Models for Prediction of Aqueous Solubility: Exploring the Potency of Randić-type Indices

Janja Sluga,[1,2] Katja Venko,[1] Viktor Drgan,[1] Marjana Novič[1,*]

[1] National Institute of Chemistry, Theory Department, Laboratory for Cheminformatics, Hajdrihova ulica 19, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Pharmacy, Chair of Pharmaceutical Chemistry, Aškerčeva cesta 7, Ljubljana, Slovenia
[*] Corresponding author's e-mail address: marjana.novic@ki.si

RECEIVED: April 26, 2021 ✶ REVISED: June 28, 2021 ✶ ACCEPTED: June 29, 2021

— THIS PAPER IS DEDICATED TO PROF. MILAN RANDIĆ ON THE OCCASION OF HIS 90TH BIRTHDAY, AND TO THE MEMORY OF PROF. MIRCEA DIUDEA —

Abstract: The development of QSPR models to predict aqueous solubility (logS) is presented. A structurally diverse set of over 1600 compounds with experimentally determined solubility values (AqSolDB database) is used for building the data-driven models based on multiple linear regression (MLR) and artificial neural network (ANN) methods to predict aqueous solubility. Molecular structures are encoded by numerous structural descriptors, including the connectivity index developed by Randić in 1975, and many later derived variations. To evaluate the potency of Randić-like descriptors in the structure-property relationship, we developed models based on two sets of descriptors, first using only Randić-like descriptors calculated with Dragon, and second using 17 commonly applied descriptors available in the AqSolDB database. All models were validated with external prediction sets, with the RMSE ranging from 0.8 to 1.1. Interestingly, the RMSE of predicted LogS values of models based only on the Randić-like descriptors were in average just 0.1 larger than the models with 17 descriptors preselected as suitable for modelling logS.

Keywords: aqueous solubility, QSPR model, MLR, ANN, connectivity index, Randić-like indices.

INTRODUCTION

AQUEOUS solubility is an important physical property describing a complex process of the interaction of a solute with water. It is of a special importance in the pharmaceutical industry, as the drug discovery relays upon solubility data to help improve drug delivery systems. In our previous work we found a potential autolysin E (AtlE) inhibitor,[1] which has to be subjected to a structure optimization procedure due to its low solubility. There are several models available to predict aqueous solubility (logS) based generally on two main approaches, either calculation of solution free energy using physics-based theory alone, or using machine learning/quantitative structure–property relationship (QSPR) models.

Theoretical calculations by using molecular simulation of aqueous solubility are extremely demanding. Physics-based solubility computations equilibrate the free energy of a molecule in the crystal lattice to the solvation energy of a molecule in saturated aqueous solution. Crystal lattice free energy is experimentally observed as free energy of sublimation, while free energy for transfer of the molecule from the gas phase to the saturated solution is solvation free energy. Both values can be calculated by molecular dynamics simulation in conjunction with one of the methods for free energy calculation, such as thermodynamic integration, thermodynamic perturbation, or metadynamics. Solvation free energy calculation[2,3] is specially demanding since in the case of solubility simulation it depends on the solute concentration. Therefore, solvation free energy value should be determined for several values of solute concentration. Obviously, a huge computational effort for one species and first principle approach is definitively not practical for purposes like drug design where hundreds of drug candidates should be screened. For much simpler tasks like octanol/water partition calculations where solvent reaction field methods should work, only few solvation models reproduce experimental values e.g. SMD (solvation model density) of Truhlar and Cramer works, while PCM - polarizable continuum model of Tomasi does not work.[2]

Often the pragmatic approach of the machine or deep learning outperforms the theoretical calculations, particularly after having more and more data available. It has been reported by J. L. McDonagh *et al.*[3] that direct theoretical calculation did not give accurate results, while machine learning was able to predict the logS with a root mean squared error (RMSE) of ~1.1 log units in a 10-fold cross-validation for 100 drug-like molecules. The group of Schneider performed a review of the drug discovery with explainable artificial intelligence.[4] They concluded that deep learning bears promise for drug discovery, but there is a demand for 'explainable' deep learning methods to address the need for a new narrative of the machine language of the molecular sciences. QSPRs are widely used *in silico* methods to predict chemical properties for untested chemicals, since they present time and cost-effective approach, which is in most cases sufficient alternative method to experimental testing. For building and validation of models, the appropriate statistical algorithms and the data matrix that includes numerical values of chemical structures and empirical values of property are needed. In the literature and on web platforms, QSPR models for aqueous solubility based on various sets of molecules are published.[3,5–16] Models are developed using various number of compounds and linear or non-linear methodologies (e.g.: multiple linear regression, partial least squares, ordinary least squares, multivariate adaptive regression splines, hierarchical clustering, group contribution, nearest neighbour, support vectors, random forest, or artificial neural networks). By utilizing information derived only from SMILES strings, the available models make predictions of aqueous solubilities (logS) in simple and straightforward procedures, which do not require molecular geometry optimizations. Models generate predictions including various molecular properties from molecular descriptors (Chemistry Development Kit (CDK), PaDEL, RDKit, Dragon, alvaDesc, SASA) to molecular structures signatures (distance graph based signatures (GBS), MACCS keys, *etc*.). Most of the models include logP parameter for prediction[3,5,7,9,12,17–19], which is also a core variable in General Solubility Equation (GSE)[20] While GSE is method of choice only when melting point is estimated, the other models, in which variables are calculated from molecular structure, can be used without limitations.

On SwissADME platform,[13] three aqueous solubility models are available: ESOL,[5] Ali[12] and SILICOS-IT. ESOL model by Delaney[5] was developed from a set of 2874 compounds using multiple linear regression and nine molecular properties like logP, molecular weight, proportion of heavy atoms in aromatic systems, and number of rotatable bonds. The model has good performance ($R^2_{\text{TR}} = 0.69$, $R^2_{\text{V}} = 0.85$) and is competitive with General Solubility Equation for medicinal/agrochemical molecules. Model Ali is based on set of 1256 compounds by using partial lest squares, MACCS keys, TPSA and logP and having performance $R^2_{\text{TR}} = 0.81$, $R^2_{\text{V}} = 0.83$. [12] While model SILICOS-IT[21] is based on fragmental method corrected by molecular weight and having $R^2_{\text{TR}} = 0.75$. Models of McElroy and Jurs[14] were generated with 399 heteroatom-containing organic compounds by using multiple linear regression (MLR) and computational neural network. The best results were obtained with non-linear CNN models ($\text{RMSE}_{\text{TR}} = 0.6$, $\text{RMSE}_{\text{V}} = 1.5$; subscripts $_{\text{TR}}$ and $_{\text{V}}$ referring to training and validation sets, respectively). The models available on VEGA and EPA platforms are based on 5020 compounds from EPI Suite database. VEGA water solubility model v.1.0.0[7] is based on artificial neural network algorithm and 15 DRAGON descriptors with performance $\text{RMSE}_{\text{TR}} = 0.84$ and $\text{RMSE}_{\text{V}} = 0.93$. EPA model is available in Toxicity Estimation Software Tool (T.E.S.T. v5.1) and makes consensus prediction from various modeling algorithms (hierarchical clustering, group contribution, nearest neighbour) with estimated $\text{RMSE}_{\text{V}} = 0.87$. [6] Another available prediction model is on pkCSM platform[11] generated with 1708 compounds and graph based signatures ($R^2_{\text{TR}} = 0.82$, $R^2_{\text{V}} = 0.73$). The model from admetSAR 2.0 is also based on the same set of compounds ($R^2 = 0.81$). [9,10] On Alvascience platform model ($R^2_{\text{TR}} = 0.76$, $R^2_{\text{V}} = 0.76$) based on ordinary least squares, 8825 compounds and five alvaDesc descriptors is presented and its predictions are in high correlation (> 0.9) with ESOL model.

The molecular connectivity index developed by Milan Randić[22] was shown to correlate almost perfectly with the boiling points of alkane isomers having two to seven carbon atoms. Hall *et al.*[23] demonstrated its relationship to water solubility and boiling point. After 25 years, in 2001, Randić published a comprehensive review of the developments of the connectivity indices as molecular descriptors in multiple linear regression analysis structure–property–activity studies.[24] The review is focused on the elaboration of higher order connectivity indices and the valence connectivity indices. The discussion has shed light on further development in chemical graph theory, novel directions in mathematical characterization of chemical, biochemical, and biological systems, all stimulated by the connectivity index. Connectivity-based molecular descriptors were later applied in many QSPR models, including those for prediction of aqueous solubility[25–29] A few years ago, Gutman *et al.*[30] depicted an interesting connection between the degree-based information content of a (molecular) graph and Randić index. However, detailed inspection of the correlation studies revealed that I(G) (degree-based information content), converse to R(G) (Randić index), carries information only on degree distribution in graphs and not on their mutual relationship, which results in the insensitivity of vertex-degree-based information of a graph on subtle structural differences among graphs. This illustrates additionally the advantage of the Randić index and its application potential in chemistry. An interesting use of connectivity indices is presented for the estimation of stability constants of metal-complexes.[31,32]

The aim of this study was to evaluate the potency of Randić-like descriptors in the structure-property relationship regarding aqueous solubility. In particular, our goal is to develop the QSPR (Quantitative Structure-Property Relationship) models to predict aqueous solubility (logS) of the potential (AtlE) inhibitors in order to optimize the initial chemical structure of poorly soluble hit compounds that were obtained in previous work.[1] Therefore, we developed and compared models based on two sets of descriptors, first using only Randić-like descriptors calculated with Dragon, and second using 17 commonly applied descriptors, as described in the literature, and available in the AqSolDB database.[33]

MATERIALS AND METHODS

Dataset

The dataset for modeling included 1674 compounds. The solubility data were obtained from AqSolDB database.[33] In this database logS values were collected from different sources.[5,20,34–41] AqSolDB consists of over 9900 unique compounds, which are coded with SMILES strings. For our modelling, we have chosen only compounds from the most reliable groups G3 and G5, *i.e*. groups composed of compounds with logS values found more than once in merged dataset and having reliability label of standard deviation < 0.5. In this way, 1818 compounds were obtained that were further reduced because of limitations of calculation of molecular descriptors (MD) in Dragon software.[42] This led to 1674 compounds in our dataset. Compounds were classified in four classes according to the thresholds reported in AqSolDB: insoluble compounds (logS > −4), moderately soluble compounds (log S −4 to −2), soluble compounds (logS −2 to 0) and highly soluble compounds (logS > 0)[33] The classes are labelled as I, L, S and H (see Supplementary file Figure S1 for distribution of the classes on a Kohonen top-map). Among the 1674 compounds included in this study, 458 compounds are insoluble (I), 563 compounds have moderate low solubility (L), while 519 compounds are soluble (S) and 134 compounds are highly soluble (H). In general, chemical structures of all compounds were numerically coded with molecular descriptors (MDs). SMILES representations of structures and experimental values of all compounds are listed in Table S1 (Supplementary Material). The experimental values available as standardized logS units in AqSolDB[33] were obtained from aqueous solubility assays that followed the OECD guidelines for testing of chemicals. Further on, 1674 compounds were divided into three datasets: 1004 compounds in the training set TR (60 %), 335 in the test set TE (20 %) and 335 in the validation set V (20 %). The split of compounds into datasets was based on mapping and visualization of compounds on top-map with CPANNatNIC software[43]. Several Kohonen maps of different network sizes (19 × 19, 20 × 20, 21 × 21) were tested for mapping the compounds according to their structural similarity and logS values. After the statistical analysis and inspection of occupied/empty neurons, the optimal neural network for the splitting purpose was of size 20 × 20 (Table S2). The most similar compounds that were located on the same neuron were then split in training, test and validation datasets. See the methodology of splitting of data using Kohonen ANN in the paper of Minovski et al.[46] and Refs. [36,37] ibid. The Kohonen map of the optimal network is visualized in Figure S1.

Molecular Descriptors

Two set of molecular descriptors were used in development of our QSPR predictive models. The first set, so called AqSol set, was composed of 17 MDs topological and physico-chemical 2D descriptors that are published in AqSolDB.[33] Originally, they were calculated with RDKit software[44] and are listed in Table S8 (Supplementary Material). The second set, so called Dragon set, was generated by Dragon 7.0 software.[42] Molecular structures in the format of SMILE strings were an input and the program calculated over 3000 molecular descriptors for each molecule. Therefore, the reduction of the number of initially calculated descriptors was crucial for QSPR modelling. We focused on Randić connectivity index and other Randić-like descriptors. In this way we ended up with 94 Randić descriptors (Table S3). Prior to be used in building of QSPR models, both set of descriptors were normalised to zero mean and unit standard deviation for each descriptor.

Generation of Models

The inputs for model building were various *m*-dimensional vectors representing the chemical structure; *m* being the number of selected MDs (independent variables), and the target (property, *i.e.* dependent variable) corresponding to logS of compounds. For model fitting, the linear and nonlinear regression approaches were used. Firstly, we used the supervised learning algorithm of counter-propagation artificial neural network (CP-ANN) and in-house software (CP-ANNatNIC).[43] During the model optimisation, the size of CP-ANN (number of neurons), number of epochs, minimal and maximal coefficients used for correction of CPANN weights, and selection of descriptors were simultaneously optimised. Secondly, the multiple linear regression models were developed with Qsarins software (DiSTA, Varese, Italy, www.qsar.it)[45] In optimization the genetic algorithm (GA) was used for selection of the influential descriptors and improvement of predictive ability and robustness of the models, which is available in Qsarins and combined with the CPANN in-house program. All models were externally validated and had defined applicability domain (AD). The AD was applied to evaluate the reliability of model predictions within the established chemical space limits. Two approaches were used for AD evaluation, the cumulative distributions of Euclidean distances to central neurons (MEDS) for CP-ANN models[46] and the leverage values for MLR models (hat values).[45] At the end, the models with the best performance parameters were selected and are presented in this paper as models NN-AqSol (NN-A), Q-AqSol (Q-A), NN-Dragon (NN-D) and Qsarins-Dragon (Q-D). These optimized regression models can be further used for prediction of logS value of any chemical of interest, considering the limitations specified with the models, such as electro-neutrality or structural applicability domain.

Statistical Evaluation of Models

First it has to be stressed that the problem of overfitting can always be an issue using ANN method. The methodology applied in this work relies on division of the data into training, test and validation set of compounds. Here, we have to take care that the compounds in the validation set, which serves for the final validation of the models, is separated from the rest of compounds at very beginning, prior to any data curation. Then the test set is selected as described, and it serves for internal intermediate testing during the model development and optimization. This standard procedure has proven to be the most robust against potential overfitting.

The criteria used for selection of the best regression models were several validation parameters, which are RMSE, R^2, Q^2_{F3}, CCC.[47–49] The models with the highest quality indicators of the validation set were selected. The consensus approach was also applied for the NN-A and NN-Q models (consNN) and for all four regression models NN-A, Q-A, NN-D and Q-D (cons4) by using the weighted average response. The calculation of the final predicted value of logS was performed following the Equation (1); M—number of models, y_k—response estimated by the k-th model, h_k—leverage[50]

$$\overline{y_w} = \frac{\sum_{k=1}^{M} \frac{y_k}{h_k}}{\sum_{k=1}^{M} \frac{1}{h_k}} \tag{1}$$

RESULTS AND DISCUSSION

Splitting of the Data

Initially, we followed recommended methodology for building QSAR models[51] and performed precise splitting of the initial data to avoid inconsistent results. The rate of the division of compounds into the training set (TR), test set (TS), and validation set (V) sets was done according to the optimal distribution of 1674 compounds described with 17 AqSol molecular descriptors on the Kohonen top map. In Kohonen neural network, molecular descriptors were mapped according to similarity and consequently few of them have place on the same neurons. A network with optimal distribution of compounds had the following parameters: 20 × 20 neuron grid, 100 learning epochs, RMSE = 0.397 and R = 0.918, 0.5 maximal learning rate, 0.01 minimal learning rate, non-toroidal NN boundary conditions, and triangular correction function of the neighborhood (Table S2). Splitting of 1674 compounds was performed following the optimal rate (60 % training set / 20 % test set/ 20 % validation set) to cover as much information as possible. In models using Dragon descriptors, the number of compounds was reduce to 1665, since for nine compounds some descriptors were not calculated. The validation set (n = 335, 334 for Dragon models) was the same for all developed models. The training set for MLR models was composed of 1339 compounds (1331 for Dragon models), merged TR and TS sets, while in CP-ANN models the training and test sets were composed of 1004 (996 for Dragon models) and 335 compounds, respectively (Table 1).

Distribution of logS

Analysis of the logS distribution reveals that the compounds have solubility values in the range between −12.1 and 1.5. Figure 1 shows the distribution of solubility values in solubility classes according to AqSolDB classification.[33] Our rates of compounds distribution according to aqueous solubility classes are comparable with rates of compounds as available in AqSolDB.

Development of Predictive Models

The logS values of 1674 compounds selected from AqSolDB and algorithms of multiple linear regression (MLR) or counter-propagation artificial neural networks (CP-ANN) with genetic algorithm (GA) were applied for development of regression models for predicting solubility in water. During optimization process hundreds of models were generated, but only the best four models were selected and represented in this study (models NN-A, Q-A, NN-D, Q-D) for two sets of descriptors (A: AqSol 17 descriptors and D:

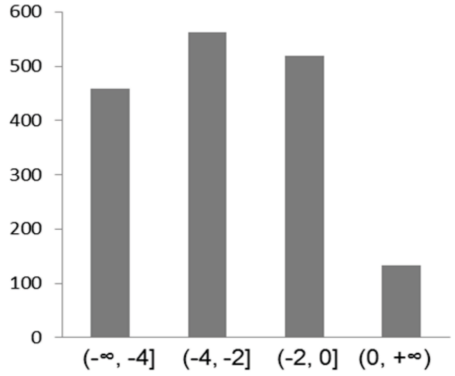

Figure 1. Distribution of compounds according to aqueous solubility ranges (logS).

Dragon Randić-like 94 descriptors) and two modelling methods (NN for neural networks and Q for MLR). The QSARINS and CP-ANNatNIC software are well known and frequently used tools for linear and non-linear QSAR models.[43,45] The best models were chosen by using Root Mean Square Error for training set (RMSE_{TR}) and validation set (RMSE_{V}) as the optimization value criteria. In building and optimization process of CP-ANN models numerous GA runs by changing different parameters like number of neurons and learning epochs, learning rate, minimal and maximal coefficients used for correction of CP-ANN weights and number of descriptors were performed. The final selection of the best models was based on the optimal values of performance indexes (Table 1, Table S13). The best four regression models are presented in Table 1.

The performance statistics for eight logS predictive QSPR models obtained by Qsarins or CP-ANN tools are given in Table 2. The best results were obtained with NN-A model (RMSE_{TR} = 0.69, RMSE_{V} = 0.76), which has a good ability to predict aqueous solubility. Next reliable model NN-D has also good performance (RMSE_{TR} = 0.81, RMSE_{V} = 0.96). The linear Q-A model has values for RMSE_{TR} = 1.38, RMSE_{V} = 1.12, while another non-linear prediction model Q-D have parameters RMSE_{TR} = 1.22 and RMSE_{V} = 1.24. If several models are developed the consensus approach could be applied according to the OECD guidance. Consensus models cons4 and aver4 are including predictions of four chosen models, but do not show better results than single models NN-A and NN-D. On the other hand, the consensus models consNN and averNN, which includes predictions from both CP-ANN models (NN-A and NN-D), are the best according to validation parameters. The consensus approach in consNN has decreased the RMSE_{TR} to 0.59, if compared with single models NN-A (RMSE_{TR} = 0.69) and NN-D (RMSE_{TR} = 0.81). The newly developed models have broad applicability domain and cover wide chemical space. Almost all compounds are in AD of our models. Therefore, models are robust and predictions are reliable, based on compounds from similar structural domain.

The best CP-ANNs models, NN-A (7 AqSol MDs) has $R^2_{\text{all}} = 0.90$ and NN-D (22 Dragon MDs) has $R^2_{\text{all}} = 0.84$, while models from literature has lowest R^2.[6,7,9–11,13] We were able to compare our results with models available on VEGA ($R^2_{\text{TR}} = 0.86$, $R^2_{\text{V}} = 0.83$) and EPA ($R^2_{\text{V}} = 0.84$) platforms, Ali *et al.* study ($R^2_{\text{TR}} = 0.81$, $R^2_{\text{V}} = 0.83$), [12] pkCSM model ($R^2_{\text{TR}} = 0.82$, $R^2_{\text{V}} = 0.73$) or admetSAR model ($R^2 = 0.81$). Anyhow, analysis of the R^2 parameters shows that also our models Q-A (9 AqSol MDs) with $R^2_{\text{all}} = 0.65$ and Q-D (12 Dragon MDs) with $R^2_{\text{all}} = 0.72$ are comparable with ESOL model[5] ($R^2_{\text{TR}} = 0.69$, $R^2_{\text{V}} = 0.85$), SILICOS-IT[21] ($R^2_{\text{TR}} = 0.75$) and AlvaScience model[17] ($R^2_{\text{TR}} = 0.76$, $R^2_{\text{V}} = 0.76$). In general, our CP-ANN models have better validation parameters in comparison with other publically available models. Furthermore, X1 (connectivity index of order 1 - Randić connectivity index) with other connectivity indices demonstrates very good choice as molecular descriptors for predicting solubility in water. Results RMSE_{TR} = 0.81 for NN-D model with Randić-like descriptors were just 0.1 log units larger than the RMSE_{TR} = 0.69 of NN-A model with selected 7 out of 17 AqSol descriptors preselected as suitable for predicting solubility in water.

Table 1. Summary of results for regression models.

Model ID	No. of compounds (TR/TS*/V)	No. of MDs	Algorithm / optimization criterion	Network size / No. of epochs	RMSE_{TR}	$\text{RMSE}_{\text{TS/CV*}}$	RMSE_{V}
NN-A	1674 (1004/335/335)	7 AqSol	CP-ANN/R_{TR}+R_{TS}	20×20/328	0.69	0.72	0.76
Q-A	1674 (1339*/335)	9 AqSol	MLR	–	1.38	1.44*	1.12
NN-D	1665 (996/335/334)	22 Dragon	CP-ANN/rmse_{TR}	20×20/329	0.81	1.07	0.96
Q-D	1665 (1331*/334)	12 Dragon	MLR	–	1.22	1.24*	1.08

* rmse for cross-validation of training set

Table 2. Statistical parameters of the best four single models and consensus models.

Model ID	RMSE$_{\text{TR}}$	R^2_{TR}	Q_{F3TR}	CCC$_{\text{TR}}$	RMSE$_{\text{V}}$	R^2_{V}	Q_{F3V}	CCC$_{\text{V}}$	RMSE$_{\text{all}}$	R^2_{all}	Q_{F3all}
NN-A	**0.69**	**0.90**	**0.90**	**0.95**	**0.76**	**0.88**	**0.88**	**0.94**	**0.71**	**0.90**	**0.90**
Q-A	1.38	0.62	0.62	0.77	1.12	0.53	0.75	0.84	1.33	0.65	0.65
NN-D	0.81	0.87	0.87	0.93	0.96	0.80	0.81	0.91	0.90	0.84	0.83
Q-D	1.22	0.58	0.70	0.83	1.08	0.69	0.77	0.87	1.19	0.72	0.72
cons4	1.10	0.82	0.76	0.85	1.04	0.63	0.78	0.87	1.09	0.77	0.76
aver4	0.83	0.86	0.86	0.92	0.78	0.83	0.88	0.93	0.82	0.87	0.86
consNN	**0.59**	**0.93**	**0.93**	**0.96**	**0.70**	**0.90**	**0.90**	**0.95**	**0.63**	**0.92**	**0.92**
averNN	0.63	0.92	0.92	0.96	0.73	0.88	0.90	0.95	0.68	0.91	0.91

Graphs on Figure 2 show the correlation among predicted and experimental logS values of six models. The best correlation is observed in consensus models like consNN ($R^2 = 0.922$) and averNN ($R^2 = 0.910$). Among single models the non-linear CP-ANN models (NN-D, $R^2 = 0.815$, and NN-A, $R^2 = 0.901$) have better performance than linear MLR models (Q-D, $R^2 = 0.705$ and Q-A, $R^2 = 0.654$).

Selection of Influential Descriptors

The descriptors selected in the NN-A and Q-A models are shown in Table S10 (Supplementary Material). Model NN-A is developed on base of 7 AqSol MDs: MolLogP, MolMR, HeavyAtomCount, NumHeteroatoms, NumAliphaticRings, RingCount, and BertzCT. In model Q-A 9 AqSol MDs were selected: MolLogP, HeavyAtomCount, NumHeteroatoms, NumRotatableBonds, NumValenceElectrons, NumAromaticRings, RingCount, LabuteASA, and BertzCT. The MDs selected in the NN-D and Q-D models are listed in Table S5 (Supplementary Material). For model NN-D, 22 Dragon MDs we selected: PW2, X1v, X4v, X0Av, X1Av, X0sol, X3sol, X5sol, RDCHI, X1Kup, X1Per, X1MulPer, Chi_H2, Chi_Dt, ChiA_Dt, Chi_Dz(Z), Chi_Dz(p), Chi_Dz(i), Chi_B(p), ChiA_B(p), VR3_B(p), and VR2_B(i). In model Q-D, 12 Dragon MDs were selected: CID, X0Av, X1sol, Chi_D, ChiA_X, ChiA_Dt, Chi_B(m), Chi_B(e), Chi_B(p), ChiA_B(p), VR2_B(i), and ChiA_B(s). Several descriptors presented in Table S5 are correlated with original Randić connectivity index (X1) and some other connectivity indices X1v, X0sol, X3sol, X5sol, RDCHI, X1Kup, X1Per, X1MulPer, and 2D matrix-based descriptor Chi_H2 in model NN-D. We also observed high correlation with walk and path counts descriptor CID,

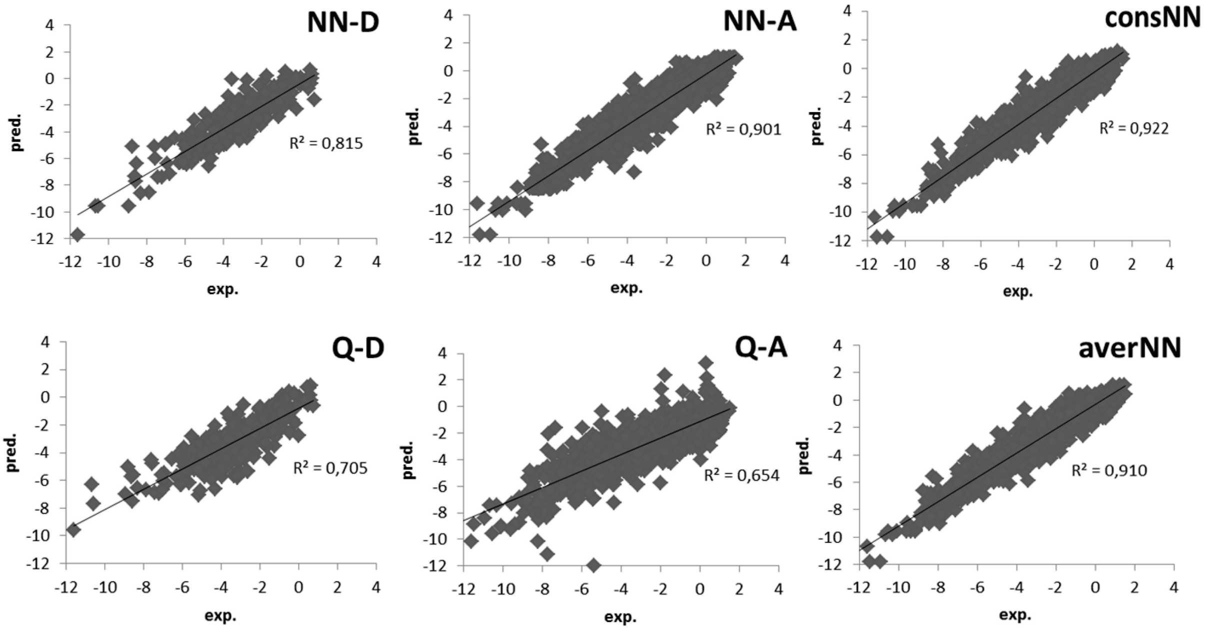

Figure 2. Correlation between experimental and predicted logS values of all compounds in the joint dataset.

connectivity index X1sol, and 2D matrix-based descriptors Chi_B(e), Chi_B(p) in model Q-D. In Table S5, we can also see, that five Dragon descriptors (X0Av, ChiA_Dt, Chi_B(p), ChiA_B(p), VR2_B(i)) are represented in both models, NN-D and Q-D.

The top ten most frequently selected Dragon MDs in model optimization are summarized in Table S4, which are X0Av, ChiA_B(p), X1Per, X1Av, X1Kup, X1MulPer, X3sol, RDCHI, X1v, and Chi_H2 for non-linear approach. X0Av, CID, ChiA_B(p), Chi_D, ChiA_X, VR2_B(s), Chi_B(m), Chi_B(s), Chi_B(p), and Chi_B(e) were most frequently selected when using linear methodology. The top ten most frequently selected AqSol MDs are listed in Table S9. MolLogP, MolMR, BertzCT, NumHAcceptors, NumHeteroatoms, MolWt, TPSA, NumHDonors, RingCount, and NumAromaticRings are represented in CP-ANN models, while NumAromaticRings, MolLogP, NumRotatableBonds, NumDonors, NumHeteroatoms, NumAliphaticRings, MolWt, BalabanJ, NumHAcceptors, and RingCount are in Qsarins models. Correlation analysis for Dragon descriptors (correlation coefficient (CC) > 0.8) are represented in Table S6, where we can see that descriptors X1, CID, X1A, PW4, X4v, X0Av, Chi_D, ChiA_Dt, ChiA_B(p), and VR3_B(p) have many correlated molecular decsriptors. MolMR, BertzCT, NumAliphaticRings, TPSA, and NumHAcceptors from AqSol descriptor dataset have also high correlation with few descriptors (Table S11). Correlation matrix for all 94 Dragon and 17 AqSol MDs are represented in Table S7 and S12, respectively.

The correlation coefficients (CC) of aqueous solubility predictions among our prediction models, experimental logS and predictions generated with six publically available models are represented in correlation matrix (Table S14). We can observe high correlation (CC > 0.9) for models NN-A, NN-D, aver4, consNN, averNN and VEGA with experimental logS values. High correlation (CC > 0.8) was also observed among predictions from NN-A, NN-D, cons4, aver4, consNN and averNN with predictions in VEGA and pkCSM models, while predictions of other three public models (ESOL, Ali and SILICOS-IT) are less correlated with our models and experimental logS, but still in reasonable high rate (CC > 0.7).

To summarize, our trial was to select a really large set of reliable solubility data of structurally diverse compounds, from different data sets, all collected in AqSol database. So compiled data (chemical diversity) combined with CP-ANN method, which is able to automatically organize smaller clusters (subsets) of compounds from which the prediction are performed, give better results than any of available models compared in the discussion, with large applicability domain.

Software together with the model of Aqueous solubility, is available from the authors upon request. Software written in Java is also freely available, the user can download it from: https://www.ki.si/en/departments/d01-theory-department/laboratory-for-cheminformatics/software/ (SOM tool developed within LIFE+ project LIFE PROSIL).

CONCLUSIONS

In this work, linear and non-linear QSPR models were constructed to predict solubility in water. A dataset of 1674 chemicals (splitting in 60 % training set, 20 % test set and 20 % validation set) and their experimentally measured aqueous solubility values (logS) was used for model development. Dragon and AqSol molecular descriptors, and MLR and CP-ANN algorithms were implemented in modelling process. Multiple calculations were performed in order to obtain the optimal model (cons NN with $R^2 = 0.92$, $\text{RMSE}_{\text{TR}} = 0.59$). Non-linear models were shown to give better results and predict the water solubility of chemicals more accurately than the linear ones. An interesting conclusion can be drawn from the comparison of models based on descriptors derived from the connectivity index (Randić-like indices) on one side, and on AqSol descriptors (recognized as most suitable descriptors for logS modelling) on the other side. The RMSE of the models based on the Randić-like descriptors only, were in average just 0.1 log units larger than the models with AqSol descriptors, which demonstrates a huge potential of connectivity index in capturing molecular structural properties that may correlate with physico-chemical as well as biochemical properties of compounds. In drug design, the solubility is one of the key properties that have to be considered. For drug design purposes it would be worth mentioning protonation states that depend on solution pH and solute pKa values. The latter is also concentration dependent. Besides, protonation states may be different in crystal than in the solution. The methodology applied in this work, however, cannot explicitly consider such effects. Nevertheless, some underlying information about the effects mentioned above is present in the solubility data used for training and developing data-driven models. Once the database of tested chemicals is large enough and covers an extensive area of chemical structure space, the models would become more reliable, also regarding the protonation state, but the interpretation of it definitively remains as a challenge for future.

Based on the results obtained in this work we are confident that these newly developed models could be a valuable guidance for design and optimization of more soluble compounds in the future. Since models are robust and reliable, we hope they will be very useful in our further drug development of autolysin E inhibitors.

Data Availability: The data used to support the findings of this study are included within the article and Support Information.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding: This work was financially supported by the grants from the Slovenian Research Agency (national program P1-0017 and MR 39008).

Acknowledgments: We thank prof. Paola Gramatica for availability of QSARINS license.

Supplementary Information. Supporting information to the paper is attached to the electronic version of the article at: https://doi.org/10.5562/cca3776.

REFERENCES

[1] J. Borišek, S. Pintar, M. Ogrizek, S. G. Grdadolnik, V. Hodnik, D. Turk, A. Perdih, M. Novič, *J. Enzyme. Inhib. Med. Chem*. **2018,** *33*(1):1239–1247. https://doi.org/10.1080/14756366.2018.1493474

[2] A.V. Marenich, C.J. Cramer, D.G. Truhlar, *J. Phys. Chem. B*, **2009**, *113*(18), 6378–6396. https://doi.org/10.1021/jp810292n

[3] J. L. McDonagh, N. Nath, L. D. Ferrari, T. V. Mourik, J. B. O. Mitchell, *J. Chem. Inf. Model*. **2014**, *54*, 844–856. https://doi.org/10.1021/ci4005805

[4] J. J. Luna, F. Grisoni, G. Schneider, *Nat. Mach. Intell.* **2020**, *2*, 573–584. https://doi.org/10.1038/s42256-020-00236-4

[5] J. S. Delaney, *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1000–1005. https://doi.org/10.1021/ci034243x

[6] T. M Martin; Toxicity Estimation Software Tool (TEST); U.S. Environmental Protection Agency, 2020, Washington, DC, https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

[7] E. Benfenati, A. Roncaglioni, A. Lombardo, A. Manganaro, *Integrating QSAR, Read-Across, and Screening Tools: The VEGAHUB Platform as an Example* in: (ed.: H. Hong) Advances in Computational Toxicology. Challenges and Advances in Computational Chemistry and Physics, vol 30., **2019**, Springer, Cham. https://www.vegahub.eu/ https://doi.org/10.1007/978-3-030-16443-0_18

[8] P. J. Nathan, J. Clarke, J. Lloyd, C. W. Hutchison, L. Downey, C. Stough, *Hum. Psychopharmacol*. **2001** *16*(4), 345–351. https://doi.org/10.1002/hup.306

[9] J. Wang, G. Krudy, T. Hou, W. Zhang, G. Holland, X. Xu, *J. Chem. Inf. Model*. **2007**, *47*, 1395–1404. https://doi.org/10.1021/ci700096r

[10] H. Yang, C. Lou, L. Sun, J. Li, Y. Cai, Z. Wang, W. Li, G. Liu, Y. Tang, B*ioinform*. **2019**, *35*(6), 1067–1069. https://doi.org/10.1093/bioinformatics/bty707

[11] D. E. V. Pires, T. L. Blundell, D. B. Ascher, *J. Med. Chem*. **2015**, *58*(9), 4066. http://biosig.unimelb.edu.au/pkcsm/prediction https://doi.org/10.1021/acs.jmedchem.5b00104

[12] J. Ali, P. Camilleri, M. B. Brown, A. J. Hutt, S. B. Kirton, *J. Chem. Inf. Model*. **2012**, *52*, 420–428. https://doi.org/10.1021/ci200387c

[13] A. Daina, O. Michielin, V. Zoete, *Sci. Rep*. **2017**, *7.*, 42717. https://doi.org/10.1038/srep42717

[14] N. R. McElroy, P. C. Jurs, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1237–1247. https://doi.org/10.1021/ci010035y

[15] M. Przybyłek, T. Jeliński, P. Cysewski, *J. Chem*. **2019**, 9858371. https://doi.org/10.1155/2019/9858371

[16] A. L. Perryman, D. Inoyama, J. S. Patel, S. Ekins, J. S. Freundlich, *ACS Omega*. **2020**, *5*(27), 16562–16567. https://doi.org/10.1021/acsomega.0c01251

[17] Alvascience models for aqueous solubility (LogS). https://www.alvascience.com/tutorial-build-models-for-aqueous-solubility-logs/

[18] D. Butina, J. M. R Gola, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 837–841. https://doi.org/10.1021/ci020279y

[19] A. Cheng, K. M. Merz, *J. Med. Chem*. **2003**, *46*, 3572–3580. https://doi.org/10.1021/jm020266b

[20] N. Jain, S. H. Yalkowsky, *J. Pharm. Sci*. **2001**, *90*(2), 234–252. https://doi.org/10.1002/1520-6017(200102)90:2%3C234::AID-JPS14%3E3.0.CO;2-V

[21] SILICOS-IT – aqueous solubility predictor http://silicos-it.be.s3-website-eu-west-1.amazonaws.com/software/filter-it/1.0.2/filter-it.html

[22] M. Randić, *J. Am. Chem. Soc*. **1975**, *97*(23), 6609–6615. https://doi.org/10.1021/ja00856a001

[23] L. H. Hall, L. B. Kier, W. J. Murray, *J. Pharm. Sci.* **1975**, *64*(12), 1974–1977. https://doi.org/10.1002/jps.2600641215

[24] M. Randić, *J. Mol. Graph. Model.* **2001**, *20*, 19–35. https://doi.org/10.1016/S1093-3263(01)00098-5

[25] M. Hewitt, M. T. Cronin, S. J. Enoch, J. C. Madden, D. W. Roberts, J. C. Dearden, *J. Chem. Inf. Model*. **2009**, *49*(11): 2572–2587. https://doi.org/10.1021/ci900286s

[26] S. Nikolić, N. Trinajstić, D. Amic, D. Beslo, S. Basak in *QSAR/QSPR Studies by Molecular Descriptors, Vol. 1* (Ed.: M. V. Diudea), Nova Science Publishers, New York, **2001**, pp. 63–81.

[27] Y. D. Hu, Y. L. Wang, *Asian J. Chem*. **2007**, *19*, 407–416.

[28] C. Zhong, Q. Hu, *J. Pharm. Sci*. **2003**, *92*, 2284–2294. https://doi.org/10.1002/jps.10499

[29] E. Estrada, E. J. Delgado, J. B. Alderete, G. A. Jaña, *J. Comput. Chem*. **2004**, *25*(14), 1787–1796. https://doi.org/10.1002/jcc.20099

[30] I. Gutman, B. Furtula, V. Katanić, *AKCE Int. J. of Graphs Comb.* **2018**, *15*(3), 307–312. https://doi.org/10.1016/j.akcej.2017.09.006

[31] N. Raos, G. Branica, A. Miličević, *Croatica Chemica Acta*, **2008**, *81*(3), 511–517.

[32] A. Miličević, N. Raos, *J. Phys. Chem. A*, **2008**, *112*(33), 7745–7749. https://doi.org/10.1021/jp802018m

[33] M. C. Sorkun, A. Khetan, S. Er, *Sci Data*. **2019**, *6*, 143. https://doi.org/10.1038/s41597-019-0151-1

[34] OECD. eChemPortal - The Global Portal to Information on Chemical Substances, 2019. https://www.echemportal.org/echemportal/propertysearch/addblock_input.action

[35] US EPA. EPI Suite Data. WATERNT (Water Solubility Fragment) Program Methodology & Validation Documents, 1995. http://esc.syrres.com/interkow/Download/WaterFragmentDataFiles.zip

[36] US EPA. EPI Suite Data. WSKOWWIN Program Methodology and Validation Documents, 1994. http://esc.syrres.com/interkow/Download/WSKOWWIN_Datasets.zip

[37] O. A. Raevsky, V. Y. Grigor'ev, D. E. Polianczyk, O. E. Raevskaja, J. C. Dearden, *J. Chem. Inf. Comput. Sci*. **2014**, *54*, 683–691. https://doi.org/10.1021/ci400692n

[38] J. Huuskonen, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777. https://doi.org/10.1021/ci9901338

[39] J. Wang, T. Hou, X. Xu, *J. Chem. Inf. Model*. **2009**, *49*, 571–581. https://doi.org/10.1021/ci800406y

[40] Goodman Group website, http://www-jmg.ch.cam.ac.uk/data/solubility/

[41] A. Llinas, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model*. **2008**, *48*, 1289–1303. https://doi.org/10.1021/ci800058v

[42] Dragon 7.0 – Software for molecular descriptors calculation, **2016**, Kode srl., Pisa, Italy, https://chm.kode-solutions.net/products_dragon.php

[43] V. Drgan, Š. Župerl, M. Vračko, C. I. Cappelli, M. Novič, *J. Cheminform*. **2017**, *9*, 30. https://doi.org/10.1186/s13321-017-0218-y

[44] RDKit software, **2019**, http://wwww.rdkit.org

[45] P. Gramatica, N. Chirico, E. Papa, S. Kovarich, S. Cassani, *J. Comput. Chem*. **2013**, *34*, 2121–2132. https://doi.org/10.1002/jcc.23361

[46] N. Minovski, Š. Župerl, V. Drgan, M. Novič, *Anal. Chim. Acta*. **2013**, *759*, 28–42. https://doi.org/10.1016/j.aca.2012.11.002

[47] V. Consonni, D. Ballabio, R. Todeschini, *J. Chem. Inf. Model*. **2009**, *49*(7), 1669–1678. https://doi.org/10.1021/ci900115y

[48] N. Chirico, P. Gramatica, *J. Chem. Inf. Model*. **2012**, *52*, 2044–2058. https://doi.org/10.1021/ci300084j

[49] X. Wang, Q. Wang, M. E.Morris, *APPS J*. **2008**, 10, 47–55. https://doi.org/10.1208/s12248-007-9001-8

[50] R. Todeschini, V. Consonni, P. Gramatica in *Comprehensive Chemometrics, Vol. 4* (Eds.: S.D. Brown, R. Tauler, B. Walczak), Elsevier, Oxford, **2009**, pp. 129–172. https://doi.org/10.1016/B978-044452701-1.00007-7

[51] P. Gramatica, *QSAR Comb. Sci*. **2007**, *26*, 694–701. https://doi.org/10.1002/qsar.200610151