The Monte Carlo technique as a tool to predict LOAEL

Jovana B. Veselinović[a], Aleksandar M. Veselinović[a], Alla P. Toropova[b*], Andrey A. Toropov[b]

[a]*University of Niš, Faculty of Medicine, Department of Chemistry, Niš, Serbia*
[b]*IRCCS- Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy*

Published version of this paper could be find here 10.1016/j.ejmech.2016.03.075

European Journal of Medicinal Chemistry Volume 116, 30 June 2016, Pages 71-75

Abstract

Quantitative structure – activity relationships (QSARs) for the Lowest Observed Adverse Effect Level (LOAEL) for a large set of organic compounds (n=565) are suggested. The molecular structures of these compounds are represented by Simplified Molecular Input-Line Entry Systems (SMILES). A criteria for the estimation quality of split into the "visible" training set (used for developing a model) and "invisible" external validation set is suggested. The correlation between the above criterion and the predictive potential of developed QSAR model (root-mean-square error for "invisible" validation set) has been detected. One-variable models are built up for several different splits into the “visible” training set and “invisible” validation set. The statistical quality of these models is quite good. Mechanistic interpretation and the domain of applicability for these models are defined according to probabilistic point of view. The methodology for defining applicability domain in QSAR modeling with SMILES notation based optimal descriptors is presented.

Key words: QSAR; LOAEL; SMILES; Ecology; Drug toxicity; Optimal descriptor

*Corresponding author:
Alla P. Toropova
Laboratory of Environmental Chemistry and Toxicology,

IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
Tel: +39 02 3901 4595
Fax: +3902 3901 4735
E-mail: alla.toropova@marionegri.it

1. Introduction

In recent years a considerable efforts have been made to assess ~~geno~~toxic impurities in pharmaceutical products as well as in products in general, because nowadays, people in the majority of ~~industrial~~ countries are under the influence by many various substances [1,2]. Toxic effects that substance can show are different and they include an adverse alteration of morphology, function, capacity, growth, development, or lifespan of a target organism distinguished from normal organisms of the same species under defined conditions of exposure. It is inconvenient to use human for biochemical and/or medicinal observations and therefore databases on potential risk of different substances are gradually increasing with experiments on animals [3,4]. However, all experiments with animals have serious ethical issues. On the other hand, the definition of endpoints which are the reliable measure of the harmfulness of substances is a task which requires long time and expensive equipment [5]. Further, chronic studies are designed to obtain a dose-response covering overt toxic effects, mild effects (the Lowest Observed Adverse Effect Level, LOAEL), and no effects (the No Observed Effect Level, NOAEL). The numerical data on these endpoints (LOAEL and NOAEL) are not available for hundreds of thousands or millions of substances which can enter the food chain and result in human exposure. For all stated reasons, the risk assessment in the absence of sufficient experimental data is a challenge for scientists, so the search for mathematical approaches which are capable to estimate the harmfulness of various substances (without direct experiment) is an attractive alternative of the experimental definition of risk assessment [6,7].

The quantitative structure – property/activity relationships (QSPRs/QSARs) based on the molecular descriptors are a computational tool used to predict various endpoints and they can be used for risk assessment [8-11]. Therefore, QSAR models for these endpoints can be useful from points of view of medicinal chemistry and ecology [25]. Optimal descriptors give possibility to establish specific one-variable QSPR/QSAR model using the Monte Carlo

method [12-15]. Recently, the optimal descriptors calculations become available with CORAL software [16], where Simplified Molecular Input-Line Entry System (SMILES) [17-19] were used for representation of the molecular structure [20-24].

The aim of the present study is the estimation of SMILES-based optimal descriptors calculated with the CORAL software as a tool to predict of the LOAEL of various organic compounds. Also, in this research the methodology for defining applicability domain in QSAR modeling with SMILES notation based optimal descriptors is presented.

2. Method

2.1. Data

Experimental data on LOAEL (logarithmic scale, mg/kg body weight per day) were taken from literature [25]. These values were converted into negative decimal logarithm, i.e. the pLOAEL is the endpoint examined in this work. It has to be notated that the database from literature contains large number of duplicates. After the extracting of the duplicates, the total number of compounds available for the QSAR analysis was 341. The supplementary materials contains the lists of compounds involved into building up models (n=341) together with the list of duplicates and wrong structures (n=226) from the above mentioned source [25]. Five random splits into the training set, invisible training set, calibration set, and the validation set were prepared according to the following principles: (i) these splits are random; (ii) these splits are not identical (Table 1); and (iii) the number of compounds in the external validation set is about 50 or more.

2.2. Optimal descriptors

The Monte Carlo method simulations, based on iterative algorithms, are run for obtaining the distribution of an unknown probabilistic entity. Therefore, Monte Carlo method develops QSAR model by generating suitable random numbers and observing how that fraction of numbers obeys a property or some properties. Further, a numerical correlation weight value (CW) is randomly assigned to SMILES-based descriptors in each independent Monte Carlo run and for a defined endpoint. Correlation Weights (DCW) for SMILES notation based optimal descriptors used in this study are calculated as the following [26]:

$$DCW(T,N) = \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (1)$$

where S_k is SMILES atoms, i.e. one symbol (e.g. 'C', 'N', '=', etc.) or two symbols which cannot be examined separately (e.g. 'Cl', 'Br', etc.); SS_k and SSS_k are compositions of two and three SMILES atoms, respectively; $CW(S_k)$, $CW(SS_k)$, and $CW(SSS_k)$ are the correlation weights for the S_k, SS_k, and SSS_k, respectively; the numerical data on the correlation weights for above-mentioned SMILES attributes (i.e. S_k, SS_k, and SSS_k) are calculated by the Monte Carlo method where their values should provide the maximum of the target function (TF):

$$TF = R + R' - abs(R - R') \times dR_w \quad (2)$$

where R and R' are correlation coefficients between pLOAEL and DCW(T,N) for the sub-training and calibration sets, respectively; dR_w (= 0.01) is an empirical constant. The parameter T is a threshold that is used to define rare and active SMILES attributes, e.g. T=3 means that if an attribute 'x' is represented only in two (or less) SMILES of the sub-training set, the 'x' is rare and CW(x) is fixed equal to zero (i.e. the 'x' is not involved in the model).

The parameter N is the number of epochs of the Monte Carlo optimization for the TF which gives maximum of correlation coefficient between LOAEL and DCW(N,T) for test set. The values of T=T* and N=N*, which gives maximum of correlation coefficient between LOAEL and DCW(T,N) for the test set are to be preferable in order to build up a model:

$$pLOAEL = C_0 + C_1 \times DCW(T^*,N^*) \quad (3)$$

The "invisible" validation set (no information on these substances is used in the modeling process) is involved in the final checking up of the predictive potential of the model calculated with Eq. 3.

There are two ways to build up a model using the optimal descriptor: (i) classical method, which is based on three sets, namely, training set, calibration set, and validation set; and (ii) balance of correlations which is based on four sets, namely, training set, invisible training set, calibration set, and validation set. In fact, in the traditional classical method, the training set is builder of a model, the calibration set is blocker of the overtraining (situation, where the excellent statistical quality of a model for the training set is accompanied by poor statistical

quality for an external set). In the case of the balance of correlation the training set distributed into active training set and invisible passive training set. The invisible training set is not take part in the optimization of the correlation weights, but permanently during of the optimization the correlation weights are checking up with the compounds of the invisible training set.

2.3. Applicability domain

The applicability domain (AD) is a characteristic of developed QSAR model that can be applied for further validation. AD is defined as biological, structural or physico-chemical space, knowledge or information on which the model of the training set is developed, and for which it is applicable to make predictions for new compounds. QSPR models are more reliable if predicted compounds are within the applicability. However, when a compound is much dissimilar to all compounds of the modeling set, a reliable prediction of its property is uncertain. For reasons stated above defining AD is one of the main aims of all developed QSAR models.

The distribution into the "visible" training set (for the described approach the "visible" training set contains also invisible training, calibration, and validation sets) and "invisible" validation set has apparent influence upon the predictability of a model. A possible measure of the quality of the split can be as the following:

$$SA_{Defect} = \sum_{active} \Sigma \left| P(SA) - P'(SA) \right| \qquad (4)$$

where the probability of an attribute SA in the sub-training set P(SA) and the probability of SA in the test set or in the calibration set P'(SA) are calculated by

$$P(SA) = \frac{N_{set}(SA)}{N_{set}} \qquad (5)$$

where $N_{set}(SA)$ is the number of SMILES which contains SA and N_{set} is the total number of SMILES in the set. The defect is calculated with active (not blocked) SA only (Table 1). If the defect = 0, the split should be estimated as "ideal" one. But in fact, this situation is not

possible. However, the value of the defect calculated with Eq. 4 gives possibility to compare various splits.

Summation of the SA_{defect} of all active SMILES attributes can be a measure of quality (defect) of each SMILES:

$$SMILES_{defect} = \sum_{SA_{defect} \in SMILES} SA_{Defect} \quad (6)$$

Summation of all $SPLIT_{defect}$ can be a measure of quality (defect) of the split into the visible training (calibration) sets and invisible validation set:

$$Split_{defect} = \sum SMILES_{Defect} \quad (7)$$

The probabilistic domain of applicability can be defined via inequality

$$SMILES_{defect} < 2 \times \overline{SMILES_{defect}} \quad (8)$$

In other words, a SMILES characterized by the $SMILES_{defect}$ which is lower than the doubled average value of the characteristics over compounds of the training set, the SMILES falls into the domain of applicability, otherwise the SMILES is out of the domain of applicability.

In addition, one can compare quality (defect) of different splits into the training, calibration, and validation sets: preferable split should be characterized by lower defect calculated with Eq. 7.

3. Results and Discussion

The threshold values from 1 to 5 and the number of epochs of the Monte Carlo optimization from 1 to 35 were examined for five random splits. Models calculated by the classical scheme are:

Split 1: pLOAEL = -2.2833 (± 0.0030) + 0.047777 (± 0.00013)×DCW(1,28) (9)

Split 2: pLOAEL = -2.3329 (± 0.0030) + 0.053500 (± 0.00014)×DCW(1,27) (10)

Split 3: pLOAEL = -2.3364 (± 0.0025) + 0.068323 (± 0.00013)×DCW(1,33) (11)

Split 4: pLOAEL = -2.2836 (± 0.0038) + 0.037157 (± 0.00013)×DCW(1,11) (12)

Split 5: pLOAEL = -2.1870 (± 0.0026) + 0.056755 (± 0.00013)×DCW(1,30) (13)

Models calculated by the balance of correlations are:

Split1: pLOAEL = -2.1723 (± 0.0059) + 0.04342(± 0.0002)×DCW(1,28) (14)

Split2: pLOAEL = -2.1450 (± 0.0072) + 0.04850(± 0.0003)×DCW(1,27) (15)

Split3: pLOAEL = -2.0593 (± 0.0056) + 0.06366(± 0.0003)×DCW(1,33) (16)

Split4: pLOAEL = -2.2813 (± 0.0072) + 0.03791(± 0.0003)×DCW(1,11) (17)

Split5: pLOAEL = -1.8400 (± 0.0057) + 0.04618(± 0.0003)×DCW(1,30) (18)

Applied two approaches gave different models and their statistical characteristics of models are represented in Table 2. One can see below, that the balance of correlations gives more reliable models for the pLOAEL examined in this work, because these models have better statistical characteristics for the validation sets (Table 2). Thus, the training and invisible training sets are united in common set in the case of the classical scheme, but these sets are acting separately, in the case of the balance of correlations. It has to be noted that statistical quality of developed models is quite good (Table 2). The statistical quality of LOAEL model suggested in the literature [25] is the following: n=567, r^2=0.54, s=0.700. One can see, that statistical characteristics of models calculated with the optimal descriptors (Table 2) are comparable with the stated model.

There are several ways to classify SMILES attributes: (i) according to transparency of their physical meaning (e.g. the transparent interpretations are ‘C.....’ is carbon atoms; ‘N.....’ is nitrogen atom; ‘#.....’ is triple covalent bond; but unclear interpretations take place for SMILES attributes such as ‘=...3.....’; ‘(...(...(...’; ‘2...(...(...’, etc.); (ii) according to their roles as the promoter of an endpoint increase (if correlation weights are stable positive in several probes of the Monte Carlo optimization) or *vise versa* promoter of an endpoint decrease (if correlation weights are stable negative in several probes of the Monte Carlo optimization; and (iii) according to their prevalence in training, invisible training, calibration, and validation sets. Consequently, one can estimate possible interpretations for SMILES attributes which have clear physicochemical (structural) meaning. In the case of the LOAEL there are the following stable promoters of the endpoint increase (Table S3): (i) presence of oxygen atoms (‘O.....’, ‘O...=.....’, ‘=...O...(...’, etc.); (ii) double bond and branching (‘C...(...=...’, ‘=...C...(...’); whereas stable promoters of the endpoint decrease are (i) presence of branching (‘(.....’); (ii) presence of cycles (‘1.....‘). Since used endpoint in this

study is pLOAEL SMILES notation based attributes (molecular fragments) defined as promoters of an endpoint increase reduces substance's toxic effect and *vice versa* SMILES notation based attributes defined as promoter of an endpoint decrease will increase substance's toxic effect.

It was expected that a split characterized by smaller value of the defect should give a better prediction than a split with the larger defect. Figure 1 shows that there is a correlation between the defect of split for the training set and standard error for validation set. The correlation should be checked up with a group of various endpoints, but the first experiment (with the LOAEL) one can estimate as successful.

4. Conclusions

The robust QSAR model for LOAEL is suggested. The balance of correlations gives better model than the classic scheme. The model has the mechanistic interpretation and a defined measurement of the quality of distribution into the training, "invisible" training, calibration, and validation set. This measurement gives possibility to check up whether chemicals of the external set (which are not involved in building up model) fall into the domain of applicability of this model.

Acknowledgments

The authors are grateful for the contribution of the EU project PROSIL funded under the LIFE program (project LIFE12 ENV/IT/000154) for financial support. J.B.V. and A.M.V. acknowledge support from Ministry of Education and Science, Republic of Serbia, under Project Number 31060. Authors also express their gratitude to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for valuable consultations on the computer science.

References

[1] A. Giordani, W. Kobel, H. Ulrich Gally, Overall impact of the regulatory requirements for genotoxic impurities on the drug development process, European Journal of Pharmaceutical Sciences 43 (2011) 1–15.

[2] G. M. Williams, K. D. Brunnemann, D. J. Smart, D. Molina, A. M. Jeffrey, J.-D. Duan, N. Krebsfaenger, A. Kampkoetter, G. Schmuck, Relationship of cellular topoisomerase

IIα inhibition to cytotoxicity and published genotoxicity of fluoroquinolone antibiotics in V79 cells, Chemico-Biological Interactions 203 (2013)386–390.

[3] C. Di Paolo, M. Cabré, J.L. Domingo, M. Gómez, Melatonin does not modify the concentration of different metals in AbPP transgenic mice, Food and Chemical Toxicology 70 (2014) 252–259.

[4] Q. Meng, D. M. Walker, J.D. McDonald, R. F. Henderson, M. M. Carter, D. L. Cook Jr., C. L. McCash, S. M. Torres, M. J. Bauer, S. K. Seilkop, P.B. Upton, N. I. Georgieva, G. Boysen, J.A. Swenberg, V.E. Walker, Age-, gender-, and species-dependent mutagenicity in T cells of mice and rats exposed by inhalation to 1,3-butadiene, Chemico-Biological Interactions 166 (2007) 121–131.

[5] K. Rajalakshmi, H. Devaraj, S. Niranjali Devaraj, Assessment of the no-observed-adverse-effect level (NOAEL) of gallic acid in mice, Food and Chemical Toxicology 39 (2001) 919-922.

[6] L. Edler, K. Poirier, M. Dourson, J. Kleiner, B. Mileson, H. Nordmann, A. Renwick, W. Slob, K. Walton, G. Würtzen, Mathematical modelling and quantitative methods, Food and Chemical Toxicology 40 (2002) 283-326.

[7] I.M. Kapetanovic, Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach, Chemico-Biological Interactions 171 (2008) 165–176.

[8] A. Afantitis, G. Melagraki, P.A. Koutentis, H. Sarimveis, G. Kollias, Ligand – based virtual screening procedure for the prediction and the identification of novel b-amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks, European Journal of Medicinal Chemistry 46 (2011) 497–508.

[9] B. Furtula, I. Gutman, Relation between second and third geometric–arithmetic indices of trees, Journal of Chemometrics 25 (2011) 87–91.

[10] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemometrics and Intelligent Laboratory Systems 109 (2011) 146–161.

[11] K. Roy, I. Mitra, Electrotopological state atom (E-State) index in drug design, QSAR, property prediction and toxicity assessment, Current Computer-Aided Drug Design 8 (2012) 135-158.

[12] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases, Journal of Molecular Graphics and Modelling 31 (2011) 10–19.

[13] J.C. Garro Martinez, P.R. Duchowicz, M.R. Estrada, G.N. Zamarbide, E.A. Castro, QSAR study and molecular design of open-chain enaminones as anticonvulsant agents, International Journal of Molecular Sciences 12 (2011) 9354–9368.

[14] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, QSAR on arylpiperazine derivatives with activity on malaria, Chemometrics and Intelligent Laboratory Systems 110 (2012) 81–88.

[15] L.M.A. Mullen, P.R. Duchowicz , E.A. Castro, QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents, Chemometrics and Intelligent Laboratory Systems 107 (2011) 269–275.

[16] CORAL software (2015) http://www.insilico.eu/coral

[17] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, Journal of Chemical Information and Computer Sciences 28 (1988) 31-36.

[18] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, Journal of Chemical Information and Computer Sciences 29 (1989) 97-101.

[19] D. Weininger, Smiles. 3. Depict. Graphical depiction of chemical structures, Journal of Chemical Information and Computer Sciences 30 (1990) 237-243.

[20] A.M. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikolić, SMILES-based QSAR model for arylpiperazines as high-affinity 5-HT1A receptor ligands using CORAL, European Journal of Pharmaceutical Sciences 48 (2013) 532-541.

[21] A.M. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikolić, SMILES-Based QSAR models for the calcium channel-antagonistic effect of 1,4-dihydropyridines, Archiv der Pharmazie 346 (2013) 134-139.

[22] A.P. Toropova, A.A. Toropov, Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO2 nanoparticles, Chemosphere 93 (2013) 2650-2655.

[23] A.P. Toropova, A.A. Toropov, CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method, European Journal of Pharmaceutical Sciences 52 (2014) 21-25.

[24] K. Nesmerak, A.A. Toropov, A.P. Toropova, P. Kohoutova, K. Waisser, SMILES-based quantitative structure-property relationships for half-wave potential of N-benzylsalicylthioamides, European Journal of Medicinal Chemistry 67 (2013) 111-114.

[25] P. Mazzatorta, M. Dominguez Estevez , M. Coulet, B. Benoit Schilter, Modeling Oral Rat Chronic Toxicity, Journal of Chemical Information and Modeling 48 (2008) 1949–1954.

[26] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: Quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats, Journal of Computational Chemistry 32 (2011) 2727–2733.

Table 1. The percentage of identity for splits 1-5 and defects of distributions “Sub-training / Test” together with defects of distributions “Sub-training / Validation” (indicated by bold)

split	Set	Defect	n	Split 1	Split 2	Split 3	Split 4	Split 5
1	Training	216.9	111	100*	80.9	30.9	39.5	29.9
	Invisible training		126	100	81.1	38.7	35.8	36.0
	Calibration		52	100	15.1	24.8	27.3	21.0
	Validation		52	100	16.8	15.4	27.5	18.9
2	Training	202.0	114		100	35.0	38.1	33.0
	Invisible training		118		100	35.9	36.2	30.6
	Calibration		54		100	16.8	21.4	22.4
	Validation		55		100	22.4	12.5	18.3
3	Training	188.9	109			100	33.5	38.4
	Invisible training		127			100	35.7	36.7
	Calibration		53			100	19.8	20.8
	Validation		52			100	16.5	18.9
4	Training	217.9	112				100	32.4
	Invisible training		114				100	29.4
	Calibration		58				100	23.4
	Validation		57				100	12.6
5	Training	218.9	110					100
	Invisible training		124					100
	Calibration		53					100
	Validation		54					100

*) $Identity\,(\%) = \frac{N_{i,j}}{0.5*(N_i + N_j)} \times 100$

where

$N_{i,j}$ is the number of substances which are distributed into the same set for both i-th split and j-th split (set =sub-training, calibration, test, validation) ;

N_i is the number of substances which are distributed into the set for i-th split;

N_j is the number of substances which are distributed into the set for j-th split.

Table 2. The statistical characteristics of QSAR models for pLOAEL

			Training				**Invisible training set**		**Calibration set**		**Validation set**	
Split	**T***	**N***	**r^2**	**q^2**	**s**	**F**	**r^2**	**s**	**r^2**	**s**	**r^2**	**s**
Classical scheme "(Training – calibration) – validation"												
1	1	28	0.71	0.70	0.540	583			0.69	0.420	0.64	0.674
2	1	27	0.72	0.71	0.540	594			0.64	0.540	0.61	0.628
3	1	33	0.80	0.79	0.475	921			0.43	0.742	0.62	0.545
4	1	11	0.62	0.62	0.628	384			0.76	0.524	0.46	0.620
5	1	30	0.75	0.75	0.517	707			0.70	0.532	0.58	0.704
Balance of correlations "(Training – invisible training – calibration) – validation"												
1	1	28	0.70	0.69	0.577	255	0.65	0.634	0.78	0.484	0.69	0.699
2	1	27	0.65	0.64	0.684	210	0.62	0.577	0.67	0.497	0.76	0.501
3	1	33	0.74	0.73	0.557	308	0.74	0.570	0.69	0.525	0.66	0.493
4	1	11	0.63	0.61	0.628	209	0.53	0.713	0.79	0.478	0.46	0.630
5	1	30	0.68	0.67	0.579	231	0.72	0.602	0.77	0.527	0.69	0.632

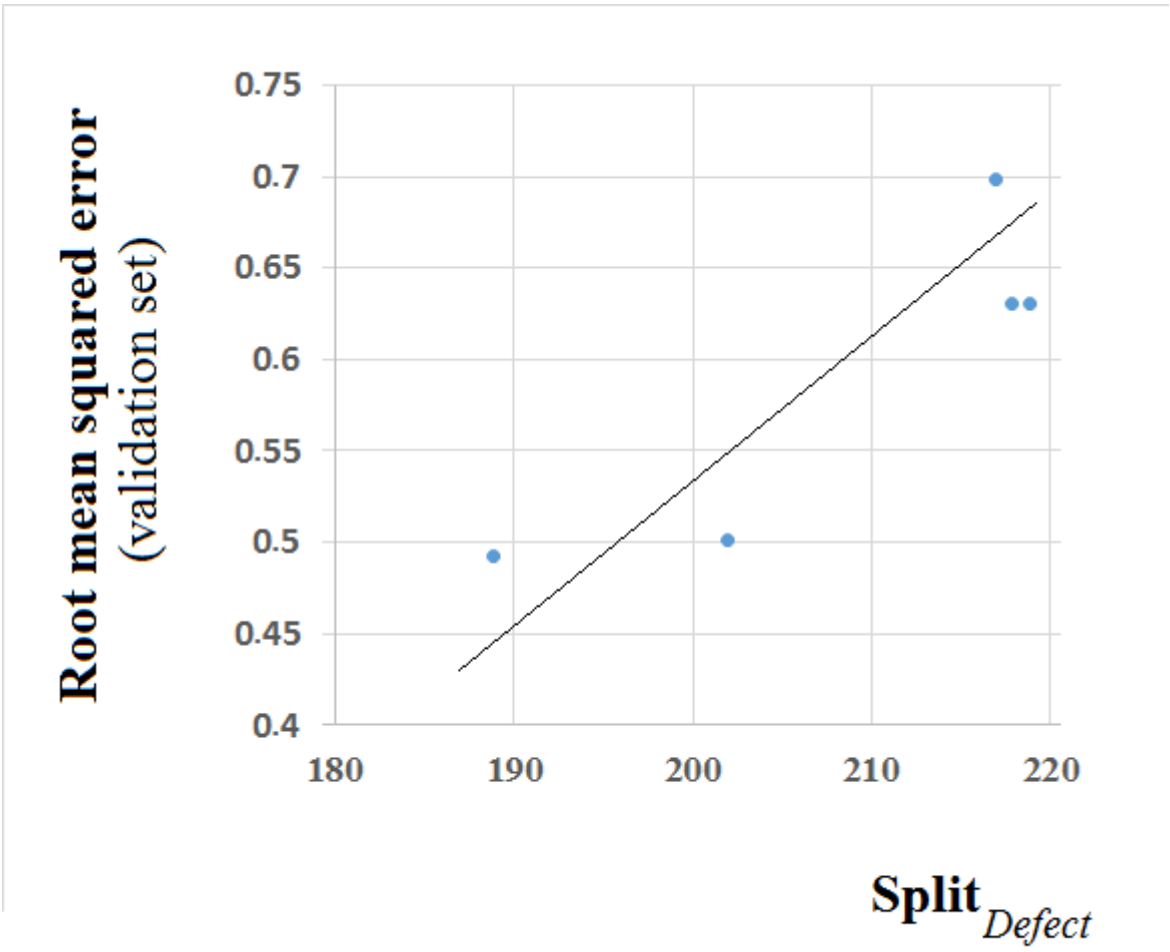

Figure 1. The correlation between the root-mean squared error for the validation set and the $Split_{defect}$ values calculated with Eq. 7 for five splits examined in this work (r^2=0.78).