Med Chem Res (2015) 24:1241–1264
DOI 10.1007/s00044-014-1193-8

MEDICINAL CHEMISTRY RESEARCH

ORIGINAL RESEARCH

Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model

Vijay H. Masand · Devidas T. Mahajan · Gulam M. Nazeruddin · Taibi Ben Hadda · Vesna Rastija · Ahmed M. Alfeefy

Received: 22 May 2014 / Accepted: 15 July 2014 / Published online: 12 August 2014

Abstract Quantitative Structure–Activity Relationship not only provides guidelines regarding structural features responsible for biological activity but it can be used also for prediction of desired activity prior to synthesis of untested chemicals. Therefore, an appropriate validation of any QSAR is of utmost importance to judge its external predictive ability. Generally, internal and external validations (preferred by many) are used in the absence of a true external dataset. The model developed using external method may not be reliable as it may not capture all essential features required for the particular SAR due to omission of some compounds, especially for small datasets. In external validation, the splitting is done either rationally or in random manner before descriptor selection. In the present study, rational splitting of dataset was performed using a novel method and its effect on statistical parameters was analyzed. The analysis reveals that the predictive ability of a QSAR model is sensitive toward (1) the method of splitting and (2) distribution of the training and the prediction sets. In addition, purposeful selection can be used to influence the statistical parameters; therefore, external validation based on single split is insufficient to guarantee the true predictive ability of a QSAR model. Besides, it appears that the selection of descriptors prior to splitting (information leakage) has little role to play in deciding external predictivity of the model. The present study reveals that as many as possible statistical parameters should be examined along with boot-strapping instead of single external validation.

Keywords QSAR · External validation · Statistical parameters · Splitting methods · Predictivity

Electronic supplementary material The online version of this article (doi:10.1007/s00044-014-1193-8) contains supplementary material, which is available to authorized users.

V. H. Masand (✉) · D. T. Mahajan
Department of Chemistry, Vidya Bharati College, Camp, Amravati, Maharashtra, India
e-mail: vijaymasand@gmail.com; vijaymasand@rediffmail.com

G. M. Nazeruddin
Department of Chemistry, Poona College, Pune, Maharashtra, India

T. B. Hadda
Laboratoire Chimie des Matériaux, Université Mohammed Premier, 60000 Oujda, Morocco

V. Rastija
Department of Chemistry, Faculty of Agriculture, Josip Juraj Strossmayer University of P. Svacica 1d, Osijek, Croatia

A. M. Alfeefy
Department of Pharmaceutical Chemistry, College of Pharmacy, Salman Bin Abdulaziz University, P.O. Box 173, Alkharj 11942, Saudi Arabia

Introduction

Under the umbrella of modern drug designing, Computer Assisted Drug Designing (CADD) is the method of choice due to faster, cheaper, and result-oriented analysis (Kubinyi, 2002; Van Drie, 2007; Yuriev *et al.*, 2011). Over the years, CADD has matured with the advent of new techniques, algorithms, and software programs. Quantitative Structure–Activity Relationship (QSAR), molecular docking, pharmacophore modeling, etc. are thriving techniques from the tenant of CADD. Of these, QSAR has gained much attention because of its applicability in risk assessment, toxicity prediction, and regulatory decisions apart from drug discovery and lead optimization. Further

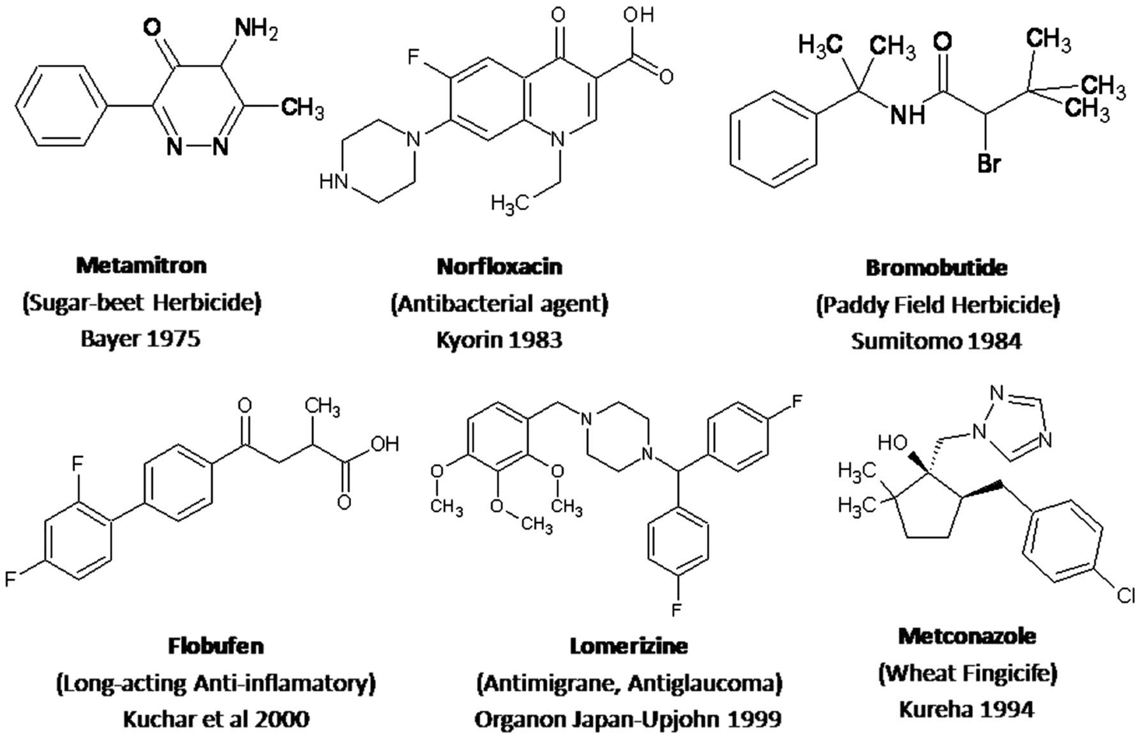


Fig. 1 Some of the commercial drugs developed with the aid of QSAR

application of QSAR models includes prediction of desired activity/property for a molecule before its synthesis and testing (Mahajan *et al*., 2012, 2013; Masand *et al*., 2012, 2010, 2013). In last decades, QSAR has contributed significantly in bringing many successful drugs in the market (see Fig. 1) (Selassie, 2003).

Therefore, QSAR models are routinely built to establish the statistical correlation between structural features (independent or predictor variables) that govern the biological activity or a physico-chemical property (dependent variable) (Scior *et al*., 2009). The four main steps involved in QSAR model building are (1) Structure drawing and geometry optimization, (2) calculation of myriad number of descriptors, (3) generation of model using least (optimal) number of descriptors, and (4) appropriate validation of mathematical model (Tropsha, 2010). The success of any QSAR model depends on various factors like accuracy of the experimental (input) data, selection of appropriate number and type of descriptors, statistical method (or algorithms), and most significantly on apposite validation of the developed model (see Fig. 2) (Huang and Fan, 2011). The utility of a QSAR model depends on its ability to predict accurately for unknown chemicals with some known degree of certainty (Roy *et al*., 2008). The prediction ability is a crucial aspect related to appropriate validation of the QSAR models. A QSAR model is considered appropriately statistically validated if it possesses good internal and external predictive ability, such models are successful in predicting the activity/property of unknown chemical (Scior *et al*., 2009; Tropsha, 2010).

Recently, appropriate validation of QSAR model is under hot debate. For thriving QSAR models, validation must be primarily for statistical robustness, prediction abilities, and applicability domain of the models (Sahigara *et al*., 2012, 2010). There are two standard ways of doing this (1) internal validation (2) external validation (Hawkins *et al*., 2003). These are performed in five different ways: leave-one-out cross-validation, leave-many-out cross-validation, Y-randomization, bootstrapping (least known among the five), and external validation (Hawkins *et al*., 2003; Kiralj and Ferreira, 2009).

The widely accepted parameter Q^2 (also symbolized as R^2_{cv}, r^2_{cv}, q^2, and Q^2_{LOO}) for internal validation is calculated by the formula (Consonni *et al*., 2010; Todeschini *et al*., 2004):

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}i - yi)^2}{\sum_{i=1}^{n} (yi - \bar{y})^2}.$$

Internal validation, a statistical method regularly performed using leave-one-out or by leave-many-out cross-validation, leads to an overestimation of predictive

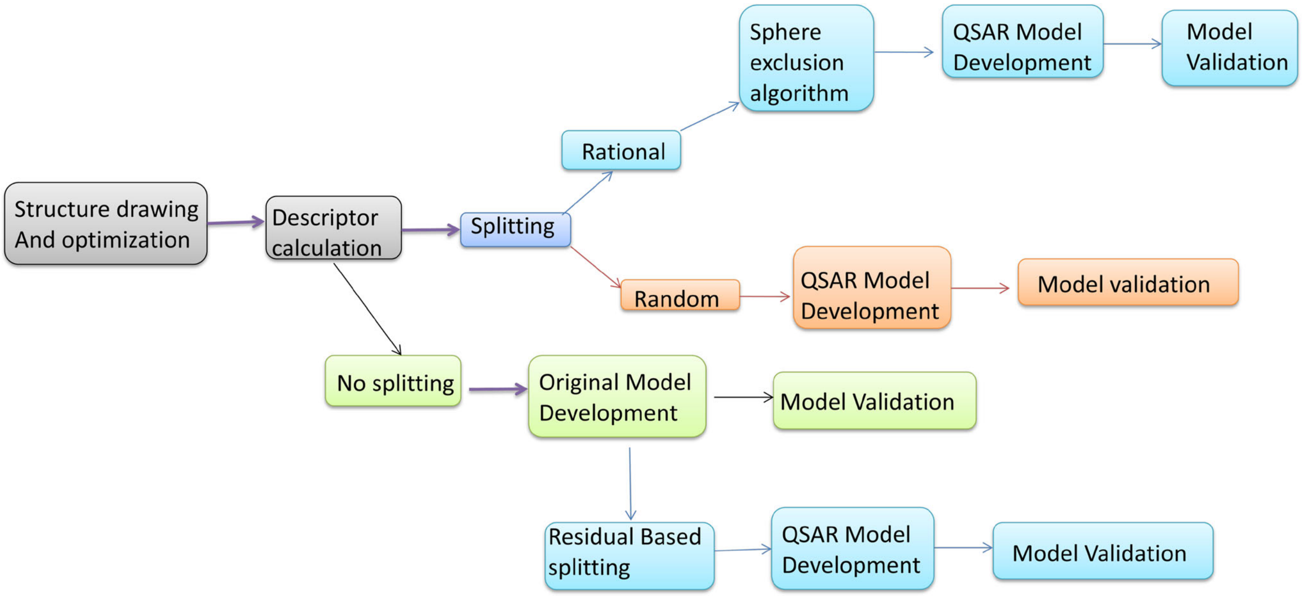


Fig. 2 Flowchart diagram for the methodology used in present study

capacity in many instances. But, it is useful for verification of robustness of the model. Therefore, internal validation may not be sufficient for validation, but it is essential (Consonni *et al.*, 2010; Golbraikh and Tropsha, 2002; Gramatica, 2013; Tropsha, 2010). It is still useful, especially, when the dataset is small or of modest size (Hawkins *et al.*, 2003).

On the other hand, external validation involves splitting the available data into training (or learning) and test (or prediction) sets. For external validation, selection of proper size of training and prediction sets is very crucial (Kiralj and Ferreira, 2009; Roy *et al.*, 2008). Generally, this splitting is performed using random division, but purposeful or rational splitting for selection of compounds whose chemistry covers the whole (or maximum) population, but does not introduce any bias is a good idea (Hawkins *et al.*, 2003). Rational or purposeful splitting methods can divide datasets into training and prediction sets in an intelligent fashion (Martin *et al.*, 2012). Different algorithms like Kennard-Stone, minimal prediction set dissimilarity, and sphere exclusion algorithms have been developed for smarter way of dividing the datasets into training and prediction sets with the aim of producing more predictive models (Chirico and Gramatica, 2012; Consonni *et al.*, 2010; Gramatica 2013; Huang and Fan 2011; Kiralj and Ferreira 2009; Martin *et al.*, 2012; Scior *et al.*, 2009). Even though, earlier studies have pointed out the superiority of rational division algorithms over the simple random splitting and activity sorting methods. Yet, appropriate selection of rational division method is still unclear because of the conflicting results (Huang and Fan 2011). Recent literature survey indicates that the method/algorithm of choice for splitting has little influence on the statistical performance of a QSAR model. Recently, Martin and co-workers reported the influence of rational selection of training and prediction sets on the model's predictivity (Martin *et al.*, 2012). However, if the prediction set is small, unknowingly, the researcher may get a prediction set for which the developed model might show a high predictive ability (Baumann and Stiefl, 2004; Chirico and Gramatica, 2012; Consonni *et al.*, 2009; Consonni *et al.*, 2010; Hawkins, 2004; Huang and Fan, 2011; Martin *et al.*, 2012; Scior *et al.*, 2009; Todeschini *et al.*, 2004; Tropsha, 2010).

The aim of the present study is to compare the statistical performance of different algorithms of rational selection, and to study the effect of descriptors selection prior to splitting (information leakage) on the external predictive ability of the model. In addition, the aim of the present study is to devise, evaluate, and compare a novel non-algorithmic method for rational splitting that influences the statistical parameters of QSAR model.

Experimental section

Datasets

For the present study, three datasets of varying size are used. The first dataset consists of forty-four N-Phenyl Ureidobenzenesulfonate Derivatives (N-PUSs) with wide variety of substituents present at different positions, as shown in Table 1, was selected from the literature (Turcotte *et al.*, 2012). The activities of these compounds

Table 1 Substituted N-Phenyl Ureidobenzenesulfonate derivatives along with $-\log IC_{50}$ (pIC_{50}) and descriptor values

S. no.	X	R_1	R_2	pIC_{50} (M) (HT-29)	F07 [C–N]	F05 [C–C]	Mor29e	Mor03m	RDF095v
1	O	4-OH	4-CEU	5.824	1	12	0.728	−3.81	1.503
2	O	2-Me	3-CEU	4.481	2	13	0.04	−5.025	5.266
3	O	2-CH_2-CH_3	3-CEU	5.367	2	14	0.283	−5.342	3.314
4	O	2-$(CH_2)_2$-CH_3	3-CEU	4.824	2	16	0.04	−4.293	5.093
5	O	4-OH	3-CEU	3.921	2	12	0.408	−4.627	4.138
6	O	2-CH_2-CH_3	4-CEU	4.770	1	14	0.246	−2.51	1.679
7	O	2-$(CH_2)_2$-CH_3	4-CEU	5.602	1	16	0.249	−4.542	3.791
8	NH	2-Me	3-CEU	4.149	2	13	0.044	−5.776	5.265
9	NH	2-CH_2-CH_3	3-CEU	4.319	2	14	0.003	−4.2	3.792
10	NH	2-$(CH_2)_2$-CH_3	3-CEU	4.824	2	16	−0.015	−5.66	6.562
11	NH	2-Me	4-CEU	4.260	2	13	−0.056	−4.842	4.379
12	NH	2-CH_2-CH_3	4-CEU	4.398	2	14	−0.199	−4.036	3.316
13	NH	2-$(CH_2)_2$-CH_3	4-CEU	4.678	2	16	−0.109	−5.379	4.276
14	O	2-Me	3-CPU	4.678	2	13	−0.099	−4.91	4.18
15	O	2-CH_2-CH_3	3-CPU	4.638	2	14	0.057	−5.709	3.351
16	O	2-$(CH_2)_2$-CH_3	3-CPU	4.854	2	16	0.09	−4.777	5.737
17	O	4-OH	3-CPU	4.292	2	12	0.371	−3.81	3.557
18	O	2-Me	4-CPU	4.585	1	13	0.026	−3.933	2.952
19	O	2-CH_2-CH_3	4-CPU	4.824	1	14	0.1	−4.001	2.835
20	O	2-$(CH_2)_2$-CH_3	4-CPU	4.886	1	16	−0.033	−4.132	4.372
21	O	4-OH	4-CPU	4.301	1	12	0.137	−3.322	2.426
22	NH	2-Me	3-CPU	4.377	2	13	0.144	−4.557	4.698
23	NH	2-CH_2-CH_3	3-CPU	4.018	2	14	−0.233	−5.079	4.639
24	NH	2-$(CH_2)_2$-CH_3	3-CPU	4.824	2	16	−0.265	−5.077	5.298
25	NH	2-Me	4-CPU	4.194	2	13	−0.247	−2.874	2.342
26	NH	2-$(CH_2)_2$-CH_3	4-CPU	4.585	2	16	−0.199	−5.308	4.241
27	O	2-Me	4-CEU	5.328	1	13	0.182	−4.774	3.14
28	O	2-Me	3-EU	4.357	2	13	0.048	−3.826	3.679
29	O	2-CH_2-CH_3	3-EU	4.481	2	14	0.123	−3.92	3.658
30	O	2-$(CH_2)_2$-CH_3	3-EU	4.602	2	16	0.064	−4.152	5.884
31	O	4-OH	3-EU	4.125	2	12	0.245	−3.361	3.747
32	O	2-Me	4-EU	4.921	1	13	0.226	−3.517	2.458
33	O	2-CH_2-CH_3	4-EU	4.921	1	14	0.289	−3.603	2.334
34	O	2-$(CH_2)_2$-CH_3	4-EU	5.620	1	16	0.179	−3.78	3.85
35	O	4-OH	4-EU	4.921	1	12	0.275	−2.822	2.042
36	O	3-Me	4-CEU	5.143	1	12	0.179	−5.158	2.729
37	NH	2-Me	3-EU	3.991	2	13	0.114	−4.003	5.481
38	NH	2-CH_2-CH_3	3-EU	4.824	2	14	−0.077	−3.948	3.99
39	NH	2-$(CH_2)_2$-CH_3	3-EU	4.387	2	16	−0.121	−3.772	4.862
40	NH	2-CH_2-CH_3	4-EU	4.066	2	14	−0.106	−3.638	2.681
41	NH	2-$(CH_2)_2$-CH_3	4-EU	4.495	2	16	−0.179	−3.383	3.517
42	O	4-Me	4-CEU	4.523	1	12	0.411	−2.191	3.995
43	O	4-OMe	4-CEU	4.745	1	13	0.579	−3.646	2.832
44	O	4-N$(Me)_2$	4-CEU	4.409	2	14	0.274	−3.852	5.562

CEU 2-chloroethylurea, *CPU* 3-chloropropylurea, *EU* ethylurea reported as IC_{50} (μM) against HT-29 colon carcinoma cells were converted to pIC_{50} (M). These derivatives of N-PUS, their corresponding $-\log IC_{50}$ (pIC_{50}) values along with the values of descriptor are presented in Table 1.

The second data consists of one hundred and twelve 4-aminoquinoline derivatives (Hwang *et al*., 2011) with a variety of substituents at different positions (see Table 2). The anti-malarial activity tested against chloroquine (CQ) sensitive (3D7) strain of *P. falciparum* reported as EC_{50} (μM) values were converted to pEC_{50} (M) for smoother statistical calculations. These derivatives of 4-aminoquinolines, their corresponding $-\log EC_{50}$ (pEC_{50}) values and values of descriptor are presented in Table 2.

The third dataset, which is a subset of the dataset 2, comprises cytotoxicity data of one hundred 4-aminoquinolines (Hwang *et al*., 2011) tested against HepG2 cell lines (see Table 3). For convenience, EC_{50} (μM) values were converted to pEC_{50} (M).

Calculation and selection of descriptors

The structures were drawn using Chemsketch 12 freeware, optimized using MMFF94 force field in TINKER, and then subjected to calculation of a large number of descriptors using e-Dragon, and PowerMV. Objective feature selection was performed to eliminate highly correlated and constant variables using QSARINS v1.2 and RapidMiner 5.0. Redundant descriptors were identified and eliminated using objective feature selection (Chirico and Gramatica, 2012; Gramatica, 2013; Mahajan *et al*., 2013; Masand *et al*., 2012, 2010, 2013). The procedure reported in the literature was employed for objective feature selection (Chirico and Gramatica 2012; Gramatica, 2013; Mahajan *et al*., 2013; Masand *et al*., 2012, 2010, 2013). As a general rule, constant for >80 % molecules, low-variance and correlated ($|R| \geq 0.6$) descriptors were excluded prior to modeling.

Methodology

The general procedure of external validation involves selection of descriptor on the basis of training set after splitting. It is well established that a QSAR model well predicts for a prediction molecule that is structurally very similar to the training set molecules because the descriptor (hence, the model) has captured common features of the training set molecules and is proficient to detect them in the new molecule (Consonni *et al*., 2009, 2010; Huang and Fan, 2011; Schuurmann *et al*., 2008; Todeschini *et al*., 2004), reverse is true for a new molecule which has very little in common with the training set data. That is, the confidence in its prediction should be low. Recently, Roy et al. proposed a new approach to overcome this critical issue, in which they used undivided dataset for selection of variables and performed internal validation (LOO cross-validation) in two different ways to ensure external predictivity of the developed model (Mitra *et al*., 2010). In the present work, descriptor selection was performed for the whole data set prior to splitting (information leakage) to determine the effect of selection of descriptors on external predictivity and behavior of different statistical parameters of the model. Genetic Algorithm (GA) available in QSARINS v1.2 was employed for the selection of optimum number and the set of descriptors applying the default settings (Chirico and Gramatica, 2011, 2012). Though, this step contravenes the basic rule that prediction set compounds should be excluded from the model development procedure, that is, they should be unknown to the developed model. But, this ensures that the selected descriptors capture the essential features that control the biological activity. In addition, it allows determining the effect of early descriptor selection (that is, prior to splitting or information leakage) on external predictivity of the models. Ferreira and Kiralj have termed such models as 'Auxiliary models' (Kiralj and Ferreira, 2009).

In QSARINS (Gramatica *et al*., 2014, 2013), CV (cross-validation) techniques are used as the optimization parameter (fitness function) for GA-based variable selection and also to verify model robustness and to avoid naïve Q^2 (Chirico and Gramatica, 2011, (2012). The novel methodology for splitting (first time reported in this work, termed as residual-based method (RBM)) begins with the creation of an original model on the basis of undivided dataset followed by splitting of dataset into training and prediction sets on the basis of sign of residuals (difference between the actual and predicted value by original model) for each sample. In short, for the whole undivided dataset, a statistically robust GA-MLR model (Original Model) was built. For some molecules, this original model resulted in positive residuals and negative for the rest. Now, for the novel methodology of splitting i.e., RBM, the whole dataset was divided rationally into training and prediction sets on the basis of sign of the residuals (obtained in the original model with the condition that the bigger set as training set). A GA-MLR QSAR model was built for the training and the prediction sets created by RBM method. For comparison purpose, the whole dataset was again divided randomly (random splitting model, termed as RSM) and rationally (using sphere exclusion model, termed a SEM method) into the training and the prediction sets with number of compounds similar to training and prediction sets as in RBM, that is, during these various splitting, the number of molecules in training and prediction set is identical in RBM, RSM, and SEM. A molecule in the training set of one method (RBM or RSM or SEM) may or may not be in the training set of other method (RBM or RSM or SEM). The identical data split with

Table 2 4-aminoquinolines used in present study along with pEC_{50} and descriptors

Sr. no.	R_1	R_2	pEC_{50} (M)	Mor13e	RDF040v	F06 [N–O]
1	PhO	Furfuryl	5.620	−0.218	7.086	2
2	PhO	2-HO-3-MeO-Bn	5.854	−0.81	7.784	1
3	PhO	Piperonyl	5.921	−1.085	8.62	1
4	PhO	3-F-6-MeO-Bn	5.959	−0.506	7.366	1
5	2-MeO-PhO	Furfuryl	6.143	−0.729	8.331	1
6	2-MeO-PhO	2-HO-3-MeO-Bn	6.152	−1.446	9.537	1
7	2-MeO-PhO	Piperonyl	6.223	−1.455	9.351	1
8	2-MeO-PhO	3-F-6-MeO-Bn	6.236	−0.523	9.875	1
9	3-MeO-PhO	Furfuryl	6.503	−0.319	7.272	1
10	3-MeO-PhO	2-HO-3-MeO-Bn	6.527	−0.748	9.432	1
11	3-MeO-PhO	Piperonyl	6.545	−1.079	9.32	1
12	3-MeO-PhO	3-F-6-MeO-Bn	6.547	−0.64	9.523	1
13	4-MeO-PhO	Furfuryl	6.600	−0.245	7.489	1
14	4-MeO-PhO	2-HO-3-MeO-Bn	6.652	−0.838	9.878	1
15	4-MeO-PhO	Piperonyl	6.682	−1.017	9.645	1
16	4-MeO-PhO	3-F-6-MeO-Bn	6.754	−0.392	9.149	1
17	4-F-PhO	Furfuryl	6.790	−0.428	6.34	1
18	4-F-PhO	2-HO-3-MeO-Bn	6.790	−0.651	8.072	1
19	4-F-PhO	Piperonyl	6.842	−0.842	8.321	1
20	4-F-PhO	3-F-6-MeO-Bn	6.860	−0.691	7.491	1
21	4-Cl-PhO	Furfuryl	6.863	−0.252	8.598	1
22	4-Cl-PhO	Furfuryl	6.879	−0.778	9.636	1
23	4-Cl-PhO	Piperonyl	6.893	−0.749	10.379	1
24	4-Cl-PhO	3-F-6-MeO-Bn	6.896	−0.636	10.35	1
25	3-Me2 N-PhO	Furfuryl	6.928	−0.024	7.162	1
26	3-Me2 N-PhO	2-HO-3-MeO-Bn	6.936	−0.727	8.607	1
27	3-Me2 N-PhO	Piperonyl	6.975	−0.858	8.745	1
28	3-Me2 N-PhO	3-F-6-MeO-Bn	7.018	−0.212	7.919	1
29	4-tertBu-PhO	Furfuryl	7.036	0.559	8.276	1
30	4-tertBu-PhO	2-HO-3-MeO-Bn	7.046	1.005	9.795	1
31	4-tertBu-PhO	Piperonyl	7.051	0.206	9.918	1
32	4-tertBu-PhO	3-F-6-MeO-Bn	7.066	−0.375	10.119	1
33	4-F-Ph	Furfuryl	7.125	−0.585	5.107	0
34	4-F-Ph	2-HO-3-MeO-Bn	7.180	−0.77	6.327	0
35	4-F-Ph	Piperonyl	7.244	−0.931	6.653	0
36	4-F-Ph	3-F-6-MeO-Bn	7.252	−1.058	6.982	0
37	3,5-CF3-Ph	Furfuryl	7.260	−0.722	5.599	0
38	3,5-CF3-Ph	2-HO-3-MeO-Bn	7.268	−0.92	7.079	0
39	3,5-CF3-Ph	Piperonyl	7.268	−1.35	7.336	0
40	3,5-CF3-Ph	3-F-6-MeO-Bn	7.268	−0.935	7.85	0
41	1-Naphtyl	Furfuryl	7.301	−0.703	6.82	0
42	1-Naphtyl	2-HO-3-MeO-Bn	7.337	−0.761	7.719	0
43	1-Naphtyl	Piperonyl	7.337	−1.118	8.451	0
44	1-Naphtyl	3-F-6-MeO-Bn	7.387	−0.556	8.566	0
45	4-CF3-Ph	Furfuryl	7.387	−0.594	5.258	0
46	4-CF3-Ph	2-HO-3-MeO-Bn	7.398	−0.331	6.622	0
47	4-CF3-Ph	Piperonyl	7.398	−0.877	7.122	0
48	4-CF3-Ph	3-F-6-MeO-Bn	7.398	−0.562	7.114	0

Table 2 continued

Sr. no.	R_1	R_2	$p\mathrm{EC}_{50}$ (M)	Mor13e	RDF040v	F06 [N–O]
49	Ph	Furfuryl	7.398	−0.449	5.193	0
50	Ph	2-HO-3-MeO-Bn	7.409	−0.4	6.467	0
51	Ph	Piperonyl	7.409	−0.754	6.74	0
52	Ph	3-F-6-MeO-Bn	7.420	−0.588	6.815	0
53	4-tertBu-Ph	Furfuryl	7.469	−0.104	7.086	0
54	4-tertBu-Ph	2-HO-3-MeO-Bn	7.481	0.348	8.467	0
55	4-tertBu-Ph	Piperonyl	7.495	−0.658	8.7	0
56	4-tertBu-Ph	3-F-6-MeO-Bn	7.509	0.472	9.679	0
57	Piperonyl	Furfuryl	7.509	0.326	4.927	0
58	Piperonyl	2-HO-3-MeO-Bn	7.509	−0.592	7.546	0
59	Piperonyl	Piperonyl	7.538	−0.501	7.801	0
60	Piperonyl	3-F-6-MeO-Bn	7.538	−0.268	8.098	0
61	4-MeO-Ph	Furfuryl	7.553	0.539	6.398	0
62	4-MeO-Ph	2-HO-3-MeO-Bn	7.569	−0.221	8.104	0
63	4-MeO-Ph	Piperonyl	7.569	−0.536	8.312	0
64	4-MeO-Ph	3-F-6-MeO-Bn	7.569	−0.071	7.819	0
65	4-F-Bn	Furfuryl	7.569	0.959	5.326	0
66	4-F-Bn	2-HO-3-MeO-Bn	7.585	−0.233	8.929	0
67	4-F-Bn	Piperonyl	7.585	−0.067	10.705	0
68	4-F-Bn	3-F-6-MeO-Bn	7.602	0.219	8.134	0
69	iso-butyl	Furfuryl	7.602	1.507	6.712	0
70	iso-butyl	2-HO-3-MeO-Bn	7.638	0.808	9.125	0
71	iso-butyl	Piperonyl	7.658	0.497	6.044	0
72	iso-butyl	3-F-6-MeO-Bn	7.658	0.112	10.014	0
73	cHex	Furfuryl	7.699	1.544	7.704	0
74	cHex	2-HO-3-MeO-Bn	7.699	0.438	11.277	0
75	cHex	Piperonyl	7.699	0.752	10.997	0
76	cHex	3-F-6-MeO-Bn	7.699	0.756	9.319	0
77	1-Et-Pr	Furfuryl	7.721	1.399	6.966	0
78	1-Et-Pr	2-HO-3-MeO-Bn	7.721	0.074	9.59	0
79	1-Et-Pr	Piperonyl	7.745	0.193	8.59	0
80	1-Et-Pr	3-F-6-MeO-Bn	7.745	0.864	7.615	0
81	3-CF3-Bn	Furfuryl	7.745	0.117	7.599	0
82	3-CF3-Bn	2-HO-3-MeO-Bn	7.745	−0.196	9.004	0
83	3-CF3-Bn	Piperonyl	7.770	0.261	8.227	0
84	3-CF3-Bn	3-F-6-MeO-Bn	7.770	0.413	8.458	0
85	4-CN-Bn	Furfuryl	7.770	1.409	8.512	0
86	4-CN-Bn	2-HO-3-MeO-Bn	7.824	−0.322	9.482	0
87	4-CN-Bn	Piperonyl	7.824	0.399	9.507	0
88	4-CN-Bn	3-F-6-MeO-Bn	7.824	0.776	9.032	0
89	Bn	Furfuryl	7.854	1.122	8.522	0
90	Bn	2-HO-3-MeO-Bn	7.886	0.25	10.837	0
91	Bn	Piperonyl	7.886	−0.019	6.437	0
92	Bn	3-F-6-MeO-Bn	7.886	0.494	10.293	0
93	3,5-Me-Bn	Furfuryl	7.886	1.444	6.402	0
94	3,5-Me-Bn	2-HO-3-MeO-Bn	7.959	1.194	10.845	0
95	3,5-Me-Bn	Piperonyl	7.959	1.419	11.419	0
96	3,5-Me-Bn	3-F-6-MeO-Bn	7.959	0.887	8.162	0

Table 2 continued

Sr. no.	R_1	R_2	pEC_{50} (M)	Mor13e	RDF040v	F06 [N–O]
97	2-Cl-4-F-Bn	Furfuryl	7.959	0.839	11.876	0
98	2-Cl-4-F-Bn	2-HO-3-MeO-Bn	7.959	−0.479	9.369	0
99	2-Cl-4-F-Bn	Piperonyl	7.959	−0.34	10.094	0
100	2-Cl-4-F-Bn	3-F-6-MeO-Bn	8.000	0.067	9.652	0
101	iso-pentyl	Furfuryl	8.046	1.491	7.984	0
102	iso-pentyl	2-HO-3-MeO-Bn	8.046	1.243	8.878	0
103	iso-pentyl	Piperonyl	8.046	0.237	9.109	0
104	iso-pentyl	3-F-6-MeO-Bn	8.046	0.566	7.314	0
105	cHexmethyl	Furfuryl	8.046	1.37	8.807	0
106	cHexmethyl	2-HO-3-MeO-Bn	8.046	0.976	10.656	0
107	cHexmethyl	Piperonyl	8.097	0.756	11.296	0
108	cHexmethyl	3-F-6-MeO-Bn	8.155	1.236	8.876	0
109	PhEt	Furfuryl	8.398	0.843	7.842	0
110	PhEt	2-HO-3-MeO-Bn	8.398	0.615	9.586	0
111	PhEt	Piperonyl	8.398	0.336	7.842	0
112	PhEt	3-F-6-MeO-Bn	9.000	0.636	14.152	0

respect to number of compounds in the training and the prediction sets was used in external validation for all models of a dataset to allow better comparison between the respective statistics (Kiralj and Ferreira, 2009). GA-MLR models were built also for the RSM and SEM. Briefly, four models were generated for each dataset.

Results and discussion

For small- and moderate-sized datasets, which is the realistic situation for a QSAR modeler, a very serious problem in developing QSAR models with reduced sets of data (splitting the sets) is the loss of considerable amount of information due to holding out of some compounds for validation purpose (Chirico and Gramatica, 2011, 2012; Consonni *et al.*, 2009, 2010; Hawkins, 2004; Hawkins *et al.*, 2008; Huang and Fan, 2011; Mitra *et al.*, 2010; Roy *et al.*, 2008; Schuurmann *et al.*, 2008; Scior *et al.*, 2009). Other confines associated in using small datasets include fortuitous correlation, poor regression statistics, failure of carrying out various statistical tests, and abnormal behavior in performed tests (Kiralj and Ferreira, 2009). This may lead to spurious conclusions in model interpretation and incorrect proposals for the mechanism of action of the compounds.

In the present study, the main emphasis is on various methods for splitting the dataset. For external validation, random as well as rational splitting methods were adopted to create training and prediction sets. For the rational splitting, a special method RBM was also evaluated. Interestingly, for all the datasets, the residual-based method resulted in radical splitting with ∼55–60 % and ∼40–45 % compounds in the training and the prediction sets, respectively. GA-MLR models were rebuilt for training and prediction sets using the same descriptors that were used for building the original model. In addition to RBM, sphere exclusion algorithm (SEM) and random (RSM) methods were also used for creating training and prediction sets, keeping the number of compounds the same as in residual-based method in the training and prediction sets. This ensures better comparison of various statistical parameters.

The analysis of Tables S1–S3 and 4, 5, 6 indicates that (i) the training and prediction sets used in RBM, RSM, and SEM models cover the diversity of the datasets and (ii) many compounds in the training and the prediction sets are close to each other (see supplementary figure S1, S2, and S3).

The statistical results for the original model for the three datasets are presented in Table 7. The minimum acceptable statistics (or recommended threshold values of statistical parameters) (Chirico and Gramatica, 2011, 2012; Huang and Fan, 2011; Kiralj and Ferreira, 2009; Martin *et al.*, 2012) for regression models in QSAR include following conditions: $R^2 > Q^2$, $Q^2 \geq 0.5$, $R^2_{\mathrm{tr}} \geq 0.6$, $R^2_{\mathrm{ex}} \geq 0.6$, $RMSE_{tr} < RMSE_{cv}$, $\Delta\mathrm{K} \geq 0.05$, $CCC \geq 0.85$, Q^2-$F^{\mathrm{n}} \geq 0.70$, and $r^2_{\mathrm{m}} \geq 0.6$ with *RMSE,* and *MAE* should be close to zero. In addition, the chance correlation of a QSAR model is validated on following criteria: $R^2_{Yrand} > Q^2_{Yrand}$,

$Q^2_{Yrand} < 0.2$ and $R^2_{Yrand} < 0.2$ → no chance correlation;

Table 3 4-aminoquinolines used in present study along with *p*EC$_{50}$ and descriptors

Sr. no.	R$_1$	R$_2$	*p*EC$_{50}$	GATS1p	E3u	E1 m	H6u	R2e
1	PhO	2-HO-3-MeO-Bn	5.046	0.959	0.41	0.552	1.306	1.88
2	PhO	Piperonyl	5.886	1.052	0.383	0.578	1.446	2.004
3	PhO	3-F-6-MeO-Bn	6.097	0.907	0.396	0.576	1.404	1.952
4	2-MeO-PhO	Furfuryl	5.538	1.089	0.309	0.483	1.1	1.988
5	2-MeO-PhO	2-HO-3-MeO-Bn	4.812	1.011	0.279	0.572	1.212	2
6	2-MeO-PhO	Piperonyl	5.215	1.085	0.288	0.595	1.355	2.133
7	2-MeO-PhO	3-F-6-MeO-Bn	5.076	0.964	0.353	0.586	1.485	2.008
8	3-MeO-PhO	Furfuryl	5.086	1.089	0.447	0.507	1.073	1.934
9	3-MeO-PhO	3-F-6-MeO-Bn	5.161	0.964	0.316	0.602	1.42	1.983
10	4-MeO-PhO	Furfuryl	5.553	1.089	0.45	0.507	1.07	1.935
11	4-MeO-PhO	2-HO-3-MeO-Bn	5.367	1.011	0.364	0.523	1.445	1.925
12	4-MeO-PhO	Piperonyl	5.066	1.085	0.429	0.607	1.096	2.03
13	4-MeO-PhO	3-F-6-MeO-Bn	5.041	0.964	0.375	0.602	1.402	1.959
14	4-F-PhO	2-HO-3-MeO-Bn	4.857	0.875	0.435	0.657	1.331	1.906
15	4-F-PhO	3-F-6-MeO-Bn	5.066	0.845	0.395	0.683	1.402	1.974
16	4-Cl-PhO	Furfuryl	4.996	0.974	0.471	0.686	1.21	1.89
17	4-Cl-PhO	Furfuryl	5.167	0.901	0.395	0.771	1.293	1.88
18	4-Cl-PhO	Piperonyl	5.056	0.986	0.389	0.786	1.446	1.997
19	4-Cl-PhO	3-F-6-MeO-Bn	5.215	0.858	0.389	0.798	1.264	1.899
20	3-Me2 N-PhO	2-HO-3-MeO-Bn	4.963	1.049	0.361	0.55	1.153	1.948
21	3-Me2 N-PhO	3-F-6-MeO-Bn	5.699	0.993	0.382	0.581	1.17	1.965
22	4-tertBu-PhO	Furfuryl	5.409	1.043	0.446	0.427	1.59	1.972
23	4-tertBu-PhO	2-HO-3-MeO-Bn	5.921	0.949	0.413	0.482	1.646	1.957
24	4-tertBu-PhO	Piperonyl	5.921	1.043	0.404	0.501	1.858	2.096
25	4-tertBu-PhO	3-F-6-MeO-Bn	5.056	0.898	0.389	0.505	1.617	1.977
26	4-F-Ph	Furfuryl	5.013	0.84	0.332	0.592	0.974	1.888
27	4-F-Ph	2-HO-3-MeO-Bn	5.027	0.799	0.393	0.665	1.252	1.91
28	4-F-Ph	Piperonyl	4.921	0.882	0.281	0.706	1.214	2.013
29	4-F-Ph	3-F-6-MeO-Bn	6	0.77	0.424	0.69	1.269	1.955
30	3,5-CF3-Ph	Furfuryl	5.523	0.739	0.262	0.663	0.878	2.48
31	3,5-CF3-Ph	2-HO-3-MeO-Bn	5.921	0.735	0.353	0.759	1.538	2.267
32	3,5-CF3-Ph	Piperonyl	5.854	0.765	0.307	0.762	1.322	2.519
33	3,5-CF3-Ph	3-F-6-MeO-Bn	6	0.73	0.347	0.783	1.441	2.283
34	1-Naphtyl	Furfuryl	5.092	0.98	0.452	0.48	1.156	1.903
35	1-Naphtyl	2-HO-3-MeO-Bn	5.432	0.863	0.367	0.575	1.259	1.918
36	1-Naphtyl	Piperonyl	5.161	0.986	0.4	0.582	1.469	2.008
37	4-CF3-Ph	Furfuryl	5.167	0.749	0.34	0.688	0.919	2.109
38	4-CF3-Ph	2-HO-3-MeO-Bn	5.377	0.74	0.414	0.756	1.279	2.107
39	4-CF3-Ph	Piperonyl	5.481	0.79	0.406	0.77	1.306	2.245
40	4-CF3-Ph	3-F-6-MeO-Bn	5.092	0.728	0.381	0.775	1.169	2.136
41	Ph	Furfuryl	5.092	1.012	0.33	0.51	0.975	1.854
42	Ph	2-HO-3-MeO-Bn	5.409	0.892	0.354	0.566	1.181	1.883
43	Ph	Piperonyl	5.328	1.015	0.279	0.637	1.21	1.979
44	Ph	3-F-6-MeO-Bn	5.456	0.835	0.374	0.603	1.165	1.919
45	4-tertBu-Ph	Furfuryl	5.409	1.009	0.353	0.447	1.453	2.019
46	4-tertBu-Ph	2-HO-3-MeO-Bn	6.155	0.886	0.377	0.497	1.569	2.037
47	4-tertBu-Ph	Piperonyl	5.796	1.011	0.415	0.528	1.756	2.163
48	4-tertBu-Ph	3-F-6-MeO-Bn	6	0.83	0.357	0.528	1.577	2.061

Table 3 continued

Sr. no.	R_1	R_2	$p\text{EC}_{50}$	GATS1p	E3u	E1 m	H6u	R2e
49	Piperonyl	Furfuryl	5.076	1.057	0.255	0.582	1.033	2.073
50	Piperonyl	2-HO-3-MeO-Bn	5.721	0.985	0.213	0.589	1.608	2.026
51	Piperonyl	Piperonyl	5.119	1.057	0.218	0.604	1.474	2.155
52	Piperonyl	3-F-6-MeO-Bn	5.046	0.939	0.208	0.688	1.718	2.084
53	4-MeO-Ph	Furfuryl	5.119	1.05	0.197	0.532	1.306	1.908
54	4-MeO-Ph	2-HO-3-MeO-Bn	5.602	0.956	0.354	0.504	1.68	1.96
55	4-MeO-Ph	Piperonyl	5.187	1.049	0.225	0.502	1.641	2.051
56	4-MeO-Ph	3-F-6-MeO-Bn	5.041	0.905	0.32	0.497	1.551	1.963
57	4-F-Bn	Furfuryl	5.215	0.838	0.35	0.597	0.765	1.941
58	4-F-Bn	2-HO-3-MeO-Bn	5.027	0.797	0.28	0.605	1.016	1.945
59	4-F-Bn	Piperonyl	5.155	0.88	0.433	0.663	0.961	1.977
60	4-F-Bn	3-F-6-MeO-Bn	5.538	0.768	0.407	0.645	1.403	1.874
61	iso-butyl	2-HO-3-MeO-Bn	5.409	0.925	0.286	0.395	1.138	2.053
62	iso-butyl	Piperonyl	5.569	1.05	0.423	0.438	1.042	2.107
63	iso-butyl	3-F-6-MeO-Bn	5.444	0.865	0.303	0.373	1.305	2.041
64	cHex	Furfuryl	5.18	1.012	0.312	0.476	1.052	2.075
65	cHex	2-HO-3-MeO-Bn	5.77	0.892	0.353	0.464	1.251	2.082
66	cHex	Piperonyl	5.268	1.015	0.256	0.541	1.231	2.129
67	cHex	3-F-6-MeO-Bn	5.495	0.835	0.228	0.517	1.377	2.031
68	1-Et-Pr	Furfuryl	4.987	1.049	0.299	0.484	1.331	1.969
69	1-Et-Pr	2-HO-3-MeO-Bn	6.959	0.925	0.361	0.444	1.619	2.021
70	1-Et-Pr	Piperonyl	5.131	1.05	0.304	0.527	1.334	2.121
71	1-Et-Pr	3-F-6-MeO-Bn	5.032	0.863	0.271	0.584	1.882	1.973
72	3-CF3-Bn	Furfuryl	5.409	0.746	0.369	0.722	1.296	2.168
73	3-CF3-Bn	2-HO-3-MeO-Bn	5.678	0.737	0.356	0.499	1.701	2.132
74	3-CF3-Bn	Piperonyl	5.602	0.787	0.337	0.732	1.565	2.246
75	3-CF3-Bn	3-F-6-MeO-Bn	6.398	0.725	0.311	0.525	1.738	2.205
76	4-CN-Bn	Furfuryl	4.943	0.963	0.23	0.602	0.981	1.931
77	4-CN-Bn	2-HO-3-MeO-Bn	5.066	0.872	0.32	0.566	1.335	1.853
78	4-CN-Bn	Piperonyl	5.066	0.983	0.152	0.578	1.257	1.97
79	4-CN-Bn	3-F-6-MeO-Bn	5.119	0.824	0.157	0.581	1.388	1.802
80	Bn	2-HO-3-MeO-Bn	5.092	0.891	0.14	0.491	1.554	1.792
81	Bn	Piperonyl	5.086	1.014	0.436	0.481	1.119	2.066
82	Bn	3-F-6-MeO-Bn	5.081	0.834	0.211	0.508	1.435	1.787
83	3,5-Me-Bn	Furfuryl	5.143	1.009	0.355	0.494	1.25	1.983
84	3,5-Me-Bn	2-HO-3-MeO-Bn	5.602	0.887	0.176	0.473	1.993	1.875
85	3,5-Me-Bn	Piperonyl	5.523	1.012	0.189	0.49	1.639	1.991
86	3,5-Me-Bn	3-F-6-MeO-Bn	6.046	0.831	0.318	0.499	1.465	2.046
87	2-Cl-4-F-Bn	Furfuryl	5.06	0.789	0.385	0.674	1.245	1.82
88	2-Cl-4-F-Bn	2-HO-3-MeO-Bn	5.538	0.757	0.266	0.597	1.555	1.975
89	2-Cl-4-F-Bn	Piperonyl	5.276	0.834	0.273	0.744	1.297	2.003
90	2-Cl-4-F-Bn	3-F-6-MeO-Bn	5.495	0.733	0.403	0.624	1.312	1.957
91	iso-pentyl	2-HO-3-MeO-Bn	5.194	0.922	0.332	0.625	1.365	2.086
92	iso-pentyl	Piperonyl	5.538	1.047	0.3	0.689	1.134	2.228
93	iso-pentyl	3-F-6-MeO-Bn	6	0.865	0.306	0.451	1.106	2.031
94	cHexmethyl	Furfuryl	5.658	1.011	0.33	0.478	1.266	2.051
95	cHexmethyl	2-HO-3-MeO-Bn	5.921	0.891	0.278	0.473	1.733	2.079
96	cHexmethyl	Piperonyl	5.027	1.014	0.234	0.553	1.335	2.27

Table 3 continued

Sr. no.	R_1	R_2	pEC$_{50}$	GATS1p	E3u	E1 m	H6u	R2e
97	cHexmethyl	3-F-6-MeO-Bn	5.553	0.834	0.293	0.476	1.355	2.082
98	PhEt	2-HO-3-MeO-Bn	5.119	0.799	0.278	0.535	1.744	1.963
99	PhEt	Piperonyl	5.229	0.832	0.204	0.554	1.546	1.854
100	PhEt	3-F-6-MeO-Bn	5.252	1.013	0.317	0.542	1.794	1.96

Any Q^2_{Yrand} and $0.2 < R^2_{Yrand} < 0.3$ → negligible chance correlation;
Any Q^2_{Yrand} and $0.3 < R^2_{Yrand} < 0.4$ → tolerable chance correlation;
Any Q^2_{Yrand} and $R^2_{Yrand} > 0.4$ → recognized chance correlation.
$(1\text{-}r^2/r^2_o) < 0.1$, $0.9 \le k \le 1.1$ or $(1\text{-}r^2/r'^2_o) < 0.1$, $0.9 \le k' \le 1.1$ with $| r^2_o - r'^2_o| < 0.3$

Except for the dataset-3, the statistical parameters point out that the GA-MLR original models for the dataset-1 and 2 are statistically robust with statistically acceptable values of R^2_{tr}, $R^2_{adj.}$, R^2_{cv}, R^2_{LMO}, R^2_{Yrand}, *s*, *Kxx*, ΔK, *RMSE$_{tr}$*, *RMSE$_{cv}$*, *CCC$_{tr}$*, *CCC$_{cv}$*, *MAE$_{tr}$*, *MAE$_{cv}$*, and *F*. Thus, from the internal validation point of view, the original models for the dataset 1 and 2 are satisfying all the essential conditions and criteria. The positive or negative contribution of a descriptor to activity remains the same during the data split and building original model indicating self-consistency of data(Kiralj and Ferreira, 2009), which is useful for model interpretation and mechanism of action.

Since, for a dataset, the same descriptors that cover the diversity of training and prediction sets are used to build models for different types of training and prediction sets, the statistical performance of residual based, random splitting and sphere exclusion should be comparable with each other for all the datasets. But, the statistical performance of each model is different (see Tables 7, 8, 9, 10). This indicates that the method of splitting has significant effect on the behavior of statistical parameters. Additionally, since the descriptors have been selected prior to splitting, the built models have captured common features of training and prediction set molecules, therefore, the models are capable to detect them in the test molecules, also. Consequently, the external predictivity of models should be high and comparable to each other for a dataset. However, the analysis of Tables 8, 9, and 10 indicates that the external predictivity of different models is different. Thus, it appears that the selection of descriptors prior to splitting has little role to play in deciding external predictivity of model. In fact, it is the diversity of training and prediction set that decides the external predictivity of any QSAR model. In other words, if the compounds in prediction set resemble the training set compounds, high predictive ability is observed for the developed model. Therefore, more number of model based on different training and prediction sets for a dataset must be developed, else, boot-strapping is an attractive option.

Results for the dataset-1

A comparison of various statistical parameters viz. R^2_{tr}, $R^2_{adj.}$, R^2_{cv}, R^2_{LMO}, R^2_{Yrand}, *s*, R^2_{ex}, *Kxx*, ΔK, *RMSE$_{tr}$*, *RMSE$_{cv}$*, *CCC$_{tr}$*, *CCC$_{cv}$*, *MAE$_{tr}$*, *MAE$_{cv}$*, r^2_m *av*, and *F* reveals that the performance of RBM model is better than the other models, which suggests that the model is statistically soundful and has good predictive ability. The r^2_m statistic, which penalizes the model profoundly for large difference between predicted and the corresponding experimental response, is higher for residual-based model indicating good external predictivity (Mitra *et al.*, 2010; Roy and Mitra, 2012). A plausible reason for this could be the distribution of the training and the prediction sets in the chemical space because both the sets used in RBM model covers diversity of the dataset. Though RBM model appears statistically robust but apropos of many statistical parameters, everything is not rosy-red for it.

For a good predictive ability *RMSE$_{ex}$* and *MAE$_{ex}$* should be as low as possible (Chirico and Gramatica, 2011), but for RBM model, the values for these parameters are higher than the rest of the models. The large difference between *RMSE$_{tr}$* (=0.118) and *RMSE$_{ex}$* (=0.441) as well as between *MAE$_{tr}$* (=0.089) and *MAE$_{ex}$* (=0.394) raises question on residual-based model's generalizability (Chirico and Gramatica, 2011, 2012). In addition, the lower values of *CCC$_{ex}$*, $Q^2\!-\!F^1$, $Q^2\!-\!F^2$, and $Q^2\!-\!F^3$ for RBM model than RSM and SEM models indicate low external predictivity of this model (Chirico and Gramatica 2011, 2012; Consonni *et al.*, 2009, 2010; Schuurmann *et al.*, 2008). Thus, the RBM model is appearing statistically soundful on the basis of many parameters, but some parameters raise doubts on its external predictivity. A possible reason could be the sensitivity of $Q^2\!-\!F^1$and $Q^2\!-\!F^2$ toward the presence of outliers in the prediction set (Consonni *et al.*, 2010). That is, the presence of more number of outliers in the prediction set of RBM model than the other models is responsible for its low external predictivity. Therefore, it can be stated

Table 4 Experimental and predicted $p\mathrm{IC}_{50}$ by different models for dataset-1

ID	$p\mathrm{IC}_{50}$	Status	Pred. $p\mathrm{IC}_{50}$ (Originalmodel)	Status	Pred. $p\mathrm{IC}_{50}$ RBM	Status	Pred. $p\mathrm{IC}_{50}$ RSM	Status	Pred. $p\mathrm{IC}_{50}$ SEM
1	5.8240	Training	5.2945	Prediction	4.9468	Training	5.5246	Training	5.2338
2	4.4810	Training	4.1434	Prediction	4.0651	Prediction	4.3532	Prediction	4.0669
3	5.3670	Training	5.1114	Prediction	4.6446	Training	5.1312	Prediction	5.0960
4	4.8240	Training	4.6201	Prediction	4.5856	Prediction	4.6241	Training	4.7066
5	3.9210	Training	4.3456	Training	4.2006	Prediction	4.7011	Training	4.2484
6	4.7700	Training	4.9528	Training	4.8453	Training	4.7697	Training	5.0715
7	5.6020	Training	5.4692	Prediction	5.2327	Training	5.4758	Prediction	5.5231
8	4.1490	Training	4.4768	Training	4.1421	Prediction	4.5794	Training	4.3759
9	4.3190	Training	4.4971	Training	4.2888	Training	4.3606	Prediction	4.5313
10	4.8240	Training	4.8351	Prediction	4.5429	Prediction	4.8050	Training	4.8310
11	4.2600	Training	4.2126	Prediction	4.0629	Prediction	4.2534	Prediction	4.1652
12	4.3980	Training	4.3672	Prediction	4.1746	Prediction	4.0790	Prediction	4.4456
13	4.6780	Training	4.8172	Training	4.6643	Prediction	4.8165	Prediction	4.8690
14	4.6780	Training	4.2410	Prediction	4.0579	Training	4.2332	Training	4.2110
15	4.6380	Training	4.8594	Training	4.5173	Training	4.9219	Training	4.8445
16	4.8540	Training	4.7210	Prediction	4.6080	Prediction	4.7724	Prediction	4.7689
17	4.2920	Training	4.3798	Training	4.1487	Training	4.4670	Prediction	4.3226
18	4.5850	Training	4.6987	Training	4.5177	Training	4.6012	Training	4.6975
19	4.8240	Training	4.9861	Training	4.7799	Training	4.8915	Prediction	5.0179
20	4.8860	Training	5.0662	Training	4.9380	Prediction	4.9073	Prediction	5.1382
21	4.3010	Training	4.5949	Training	4.3935	Training	4.4711	Training	4.5774
22	4.3770	Training	4.3227	Prediction	4.1462	Training	4.4158	Prediction	4.2844
23	4.0180	Training	4.2580	Training	4.1286	Training	4.2083	Training	4.2826
24	4.8240	Training	4.5794	Prediction	4.4278	Prediction	4.4112	Training	4.6660
25	4.1940	Training	3.8950	Prediction	3.9259	Prediction	3.6105	Training	3.9953
26	4.5850	Training	4.6599	Training	4.5970	Training	4.6743	Prediction	4.7344
27	5.3280	Training	5.0269	Prediction	4.6933	Prediction	5.0460	Training	4.9675
28	4.3570	Training	4.3042	Prediction	4.1023	Training	4.1680	Training	4.3111
29	4.4810	Training	4.5794	Training	4.3587	Prediction	4.4580	Prediction	4.6178
30	4.6020	Training	4.6885	Training	4.5141	Training	4.5384	Prediction	4.7488
31	4.1250	Training	4.2266	Training	3.9974	Prediction	4.1417	Training	4.1867
32	4.9210	Training	4.9349	Training	4.6648	Training	4.8049	Prediction	4.9387
33	4.9210	Training	5.2090	Training	4.9217	Prediction	5.0860	Training	5.2472
34	5.6200	Training	5.3283	Prediction	5.1025	Prediction	5.1492	Training	5.4025
35	4.9210	Training	4.7116	Prediction	4.4781	Prediction	4.5533	Prediction	4.7109
36	4.9590	Training	4.9578	Prediction	4.5756	Training	5.0387	Prediction	4.8530
37	3.9910	Training	4.1116	Training	3.9965	Training	4.1346	Training	4.0684
38	4.8240	Training	4.3522	Prediction	4.1887	Prediction	4.1562	Training	4.4158
39	4.3870	Training	4.5814	Training	4.4422	Training	4.2705	Prediction	4.7137
40	4.0660	Training	4.3520	Training	4.2611	Training	4.1534	Prediction	4.4426
41	4.4950	Training	4.5532	Training	4.4897	Prediction	4.2078	Training	4.7186
42	4.5230	Training	4.3868	Prediction	4.3279	Training	4.3650	Training	4.3516
43	4.7450	Training	5.1121	Training	4.8920	Training	5.2956	Training	5.0681
44	4.4090	Training	4.3261	Prediction	4.2793	Prediction	4.4605	Prediction	4.3016

RBM Residual-based model, *RBM* Random splitting model, *SEM* Sphere exclusion model

Table 5 Experimental and predicted pEC_{50} by different models for dataset-2

ID	pEC_{50}(M)	Status	Pred. pEC_{50} (Original model)	Status	Pred. pEC_{50} RBM	Status	Pred. pEC_{50} RSM	Status	Pred. pEC_{50} SEM
1	5.620	Training	5.6830	Training	5.3535	Prediction	5.8460	Training	5.6082
2	5.854	Training	6.5075	Training	6.2924	Prediction	6.5813	Training	6.4897
3	5.921	Training	6.5065	Training	6.2847	Training	6.5663	Prediction	6.4877
4	5.959	Training	6.5481	Training	6.3330	Prediction	6.6285	Training	6.5248
5	6.143	Training	6.5698	Training	6.3449	Prediction	6.6337	Training	6.5420
6	6.152	Training	6.4911	Training	6.2631	Training	6.5357	Prediction	6.4734
7	6.223	Training	6.4744	Training	6.2494	Training	6.5223	Training	6.4594
8	6.236	Training	6.7401	Training	6.4880	Prediction	6.7765	Prediction	6.6853
9	6.503	Training	6.5859	Training	6.3683	Training	6.6674	Prediction	6.5571
10	6.527	Training	6.6512	Training	6.4103	Prediction	6.6959	Prediction	6.6101
11	6.545	Training	6.5626	Training	6.3303	Prediction	6.6101	Prediction	6.5347
12	6.547	Training	6.6844	Training	6.4399	Training	6.7272	Prediction	6.6382
13	6.600	Training	6.6207	Training	6.3983	Training	6.6982	Prediction	6.5865
14	6.652	Training	6.6644	Training	6.4187	Training	6.7015	Prediction	6.6207
15	6.682	Training	6.6030	Prediction	6.3645	Training	6.6446	Prediction	6.5687
16	6.754	Training	6.7149	Prediction	6.4709	Prediction	6.7637	Training	6.6648
17	6.790	Training	6.4868	Prediction	6.2853	Training	6.5849	Training	6.4738
18	6.790	Training	6.5683	Prediction	6.3457	Training	6.6367	Prediction	6.5412
19	6.842	Training	6.5417	Prediction	6.3193	Training	6.6062	Training	6.5182
20	6.860	Training	6.5133	Prediction	6.3001	Training	6.5919	Prediction	6.4950
21	6.863	Training	6.7057	Prediction	6.4669	Prediction	6.7638	Prediction	6.6576
22	6.879	Training	6.6599	Prediction	6.4166	Prediction	6.7011	Prediction	6.6172
23	6.893	Training	6.7250	Prediction	6.4700	Prediction	6.7531	Training	6.6718
24	6.896	Training	6.7500	Prediction	6.4931	Training	6.7783	Training	6.6931
25	6.928	Training	6.6484	Prediction	6.4264	Prediction	6.7312	Prediction	6.6106
26	6.936	Training	6.5918	Prediction	6.3628	Training	6.6509	Training	6.5605
27	6.975	Training	6.5710	Prediction	6.3426	Training	6.6280	Prediction	6.5426
28	7.018	Training	6.6622	Prediction	6.4328	Training	6.7322	Prediction	6.6214
29	7.036	Training	6.8761	Prediction	6.6254	Prediction	6.9378	Training	6.8033
30	7.046	Training	7.1023	Training	6.8198	Prediction	7.1363	Training	6.9942
31	7.051	Training	6.9193	Prediction	6.6514	Training	6.9531	Prediction	6.8380
32	7.066	Training	6.7948	Prediction	6.5360	Training	6.8266	Prediction	6.7317
33	7.125	Training	7.2652	Training	7.1980	Training	7.3070	Prediction	7.3185
34	7.180	Training	7.3159	Training	7.2344	Training	7.3368	Prediction	7.3602
35	7.244	Training	7.3026	Training	7.2195	Prediction	7.3181	Prediction	7.3484
36	7.252	Training	7.2977	Training	7.2123	Prediction	7.3078	Training	7.3438
37	7.260	Training	7.2706	Training	7.1989	Training	7.3042	Training	7.3225
38	7.268	Training	7.3385	Training	7.2489	Training	7.3466	Training	7.3785
39	7.268	Training	7.2549	Prediction	7.1703	Training	7.2595	Prediction	7.3069
40	7.268	Training	7.3952	Training	7.2943	Prediction	7.3897	Prediction	7.4258
41	7.301	Training	7.3706	Training	7.2803	Prediction	7.3827	Training	7.4062
42	7.337	Training	7.4269	Training	7.3244	Training	7.4233	Training	7.4530
43	7.337	Training	7.3980	Training	7.2920	Prediction	7.3824	Prediction	7.4275
44	7.387	Training	7.5425	Training	7.4232	Prediction	7.5236	Prediction	7.5505
45	7.387	Training	7.2748	Prediction	7.2056	Prediction	7.3140	Training	7.3265
46	7.398	Training	7.4449	Training	7.3498	Prediction	7.4595	Prediction	7.4697
47	7.398	Training	7.3523	Prediction	7.2611	Prediction	7.3594	Prediction	7.3902

Table 5 continued

ID	pEC_{50}(M)	Status	Pred. pEC_{50} (Original model)	Status	Pred. pEC_{50} RBM	Status	Pred. pEC_{50} RSM	Status	Pred. pEC_{50} SEM
48	7.398	Training	7.4276	Training	7.3300	Prediction	7.4341	Training	7.4544
49	7.398	Training	7.3047	Prediction	7.2334	Training	7.3447	Training	7.3521
50	7.409	Training	7.4161	Training	7.3248	Prediction	7.4336	Prediction	7.4454
51	7.409	Training	7.3521	Prediction	7.2640	Training	7.3657	Training	7.3905
52	7.420	Training	7.3980	Prediction	7.3054	Training	7.4098	Prediction	7.4295
53	7.469	Training	7.5359	Training	7.4292	Prediction	7.5418	Training	7.5467
54	7.481	Training	7.7528	Training	7.6162	Training	7.7333	Prediction	7.7299
55	7.495	Training	7.5284	Training	7.4092	Training	7.5073	Prediction	7.5383
56	7.509	Training	7.8774	Training	7.7202	Prediction	7.8364	Training	7.8346
57	7.509	Training	7.4709	Prediction	7.3874	Training	7.5136	Training	7.4941
58	7.509	Training	7.4541	Prediction	7.3507	Training	7.4532	Training	7.4765
59	7.538	Training	7.4960	Prediction	7.3869	Training	7.4903	Prediction	7.5118
60	7.538	Training	7.5754	Training	7.4571	Training	7.5639	Prediction	7.5791
61	7.553	Training	7.6372	Training	7.5275	Prediction	7.6535	Training	7.6340
62	7.569	Training	7.5872	Training	7.4678	Prediction	7.5755	Training	7.5892
63	7.569	Training	7.5275	Prediction	7.4115	Training	7.5129	Prediction	7.5380
64	7.569	Training	7.6011	Training	7.4829	Prediction	7.5941	Prediction	7.6014
65	7.569	Training	7.6547	Training	7.5522	Training	7.6888	Prediction	7.6503
66	7.585	Training	7.6488	Training	7.5174	Training	7.6226	Training	7.6406
67	7.585	Training	7.8276	Training	7.6663	Prediction	7.7699	Training	7.7908
68	7.602	Training	7.6956	Training	7.5667	Training	7.6824	Training	7.6816
69	7.602	Training	7.8951	Training	7.7607	Prediction	7.9036	Training	7.8535
70	7.638	Training	7.9151	Training	7.7593	Prediction	7.8830	Training	7.8675
71	7.658	Training	7.5994	Prediction	7.4958	Prediction	7.6221	Training	7.6022
72	7.658	Training	7.8168	Training	7.6621	Training	7.7707	Prediction	7.7825
73	7.699	Training	7.9816	Training	7.8316	Training	7.9726	Training	7.9259
74	7.699	Training	7.9941	Training	7.8138	Training	7.9251	Training	7.9320
75	7.699	Training	8.0479	Training	7.8653	Prediction	7.9831	Training	7.9783
76	7.699	Training	7.9178	Training	7.7601	Training	7.8823	Training	7.8694
77	7.721	Training	7.8890	Training	7.7530	Training	7.8932	Training	7.8479
78	7.721	Training	7.7745	Training	7.6269	Prediction	7.7360	Training	7.7469
79	7.745	Training	7.7250	Prediction	7.5898	Prediction	7.7038	Training	7.7061
80	7.745	Training	7.8106	Training	7.6761	Prediction	7.8049	Training	7.7803
81	7.745	Training	7.6292	Prediction	7.5104	Training	7.6256	Prediction	7.6257
82	7.745	Training	7.6636	Prediction	7.5303	Training	7.6360	Training	7.6531
83	7.770	Training	7.7130	Prediction	7.5819	Training	7.6980	Prediction	7.6963
84	7.770	Training	7.7678	Prediction	7.6300	Training	7.7483	Prediction	7.7427
85	7.770	Training	8.0122	Training	7.8530	Training	7.9893	Prediction	7.9510
86	7.824	Training	7.6705	Prediction	7.5327	Prediction	7.6349	Training	7.6585
87	7.824	Training	7.8463	Training	7.6933	Training	7.8085	Training	7.8083
88	7.824	Training	7.9001	Training	7.7463	Training	7.8697	Training	7.8548
89	7.854	Training	7.9437	Training	7.7903	Prediction	7.9214	Training	7.8926
90	7.886	Training	7.9143	Training	7.7445	Training	7.8535	Prediction	7.8646
91	7.886	Training	7.5056	Prediction	7.4069	Prediction	7.5227	Training	7.5218
92	7.886	Training	7.9307	Training	7.7639	Training	7.8788	Prediction	7.8792
93	7.886	Training	7.8557	Prediction	7.7272	Training	7.8698	Training	7.8203
94	7.959	Training	8.1426	Training	7.9532	Training	8.0793	Training	8.0592

Table 5 continued

ID	pEC_{50}(M)	Status	Pred. pEC_{50} (Original model)	Status	Pred. pEC_{50} RBM	Status	Pred. pEC_{50} RSM	Status	Pred. pEC_{50} SEM
95	7.959	Training	8.2417	Training	8.0391	Training	8.1678	Prediction	8.1430
96	7.959	Training	7.8589	Prediction	7.7158	Training	7.8435	Training	7.8208
97	7.959	Training	8.1376	Training	7.9401	Prediction	8.0571	Training	8.0536
98	7.959	Training	7.6238	Prediction	7.4910	Training	7.5906	Training	7.6188
99	7.959	Training	7.7140	Prediction	7.5675	Prediction	7.6677	Prediction	7.6948
100	8.000	Training	7.7776	Prediction	7.6292	Prediction	7.7381	Training	7.7496
101	8.046	Training	7.9907	Prediction	7.8377	Training	7.9769	Training	7.9333
102	8.046	Training	8.0007	Prediction	7.8395	Prediction	7.9719	Training	7.9408
103	8.046	Training	7.7762	Prediction	7.6324	Prediction	7.7457	Training	7.7490
104	8.046	Training	7.7153	Prediction	7.5914	Prediction	7.7155	Training	7.6994
105	8.046	Training	8.0258	Prediction	7.8630	Training	7.9979	Prediction	7.9622
106	8.046	Training	8.0753	Training	7.8931	Training	8.0159	Training	8.0020
107	8.097	Training	8.0722	Prediction	7.8851	Prediction	8.0021	Training	7.9986
108	8.155	Training	7.9989	Prediction	7.8379	Prediction	7.9701	Prediction	7.9392
109	8.398	Training	7.8233	Prediction	7.6858	Training	7.8136	Prediction	7.7908
110	8.398	Training	7.9046	Prediction	7.7459	Training	7.8649	Training	7.8579
111	8.398	Training	7.7010	Prediction	7.5740	Prediction	7.6926	Prediction	7.6866
112	9.000	Training	8.2665	Prediction	8.0393	Prediction	8.1465	Prediction	8.1606

that residual-based model possesses poor external predictivity, hence should not be adopted to create QSAR models. In addition, as many as possible parameters should be reported for a QSAR model developed using single splitting method. Because, the true predictive ability of residual-based model was captured, only when many statistical parameters were calculated.

For some parameters viz. R^{2}_{tr}, $R^{2}_{\text{adj.}}$, R^{2}_{cv}, R^{2}_{LMO}, R^{2}_{Yrand}, *s, Kxx*, ΔK, $RMSE_{tr}$, $RMSE_{cv}$, CCC_{tr}, CCC_{cv}, MAE_{tr}, MAE_{cv}, and *F*, the performance of random splitting model is either statistically satisfactory or comparable with the other models. But, for some parameters viz. CCC_{ex}, $r^{2}_{m}\,av$, and r^{2}_{m}, the performance of the model is questionable. A large difference of 0.309 between R^{2}_{tr} (=0.674), and R^{2}_{cv} (=0.365) for sphere exclusion model reflects large inaccuracy of the model (Schuurmann *et al*., 2008) or overfitting (Kiralj and Ferreira, 2009). A probable reason could be either the small size of dataset-1 or size of training and prediction sets. But, the problem of large inaccuracy of model or overfitting is not visible for other models. Similarly, the very low value of *F* (=7.023) indicates low statistical reliability of the sphere exclusion model. A very surprising and rare observation for sphere exclusion model is the higher values of $Q^{2}{-}F^{1}$ (=0.706), $Q^{2}{-}F^{2}$ (=0.705), and $Q^{2}{-}F^{3}$ (=0.828) than R^{2} (=0.674), leading to the contrasting conclusion that the model is able to predict new data better than fitting available ones (Chirico and Gramatica, 2011; 2012).

For residual-based and random splitting models, $RMSE_{tr}$ and MAE_{tr} are lower than $RMSE_{ex}$ and MAE_{ex}, respectively. This indicates that the samples for which the models fit very well are present in the training set. Exactly reverse is true for the sphere exclusion model, for which $RMSE_{tr}$ and MAE_{tr} are higher than $RMSE_{ex}$ and MAE_{ex}, respectively. This observation points out one serious drawback of common practice followed in external validation, in which single split is performed to validate the model. If a researcher purposely selects training and prediction sets such that $RMSE_{tr}$ and MAE_{tr} are higher than $RMSE_{ex}$ and MAE_{ex}, respectively then, the model will be with lower internal predictivity but with high external predictivity. In such case, many parameters will give false positive results because of the purposeful selection of training and prediction sets. Therefore, one cannot merely rely on external validation based on single split; instead, boot-strap or multiple modeling (Masand *et al*., 2014) must be followed to develop a good number of statistically robust QSAR models with good external predictive ability.

As the number of compounds is same in the training and the prediction sets for the three models, the difference between R^{2} and Q^{2} should be comparable for all the models. But, different models have different variation indicating that the method of splitting has good influence on many statistical parameters.

Results for the dataset-2

Similar to the dataset-1, different statistical parameters viz. R^{2}_{tr}, $R^{2}_{\text{adj.}}$, R^{2}_{cv}, R^{2}_{LMO}, R^{2}_{Yrand}, *s*, R^{2}_{ex}, *Kxx*, ΔK, $RMSE_{tr}$,

Table 6 Experimental and predicted pEC_{50} by different models for dataset-3

ID	pEC_{50} (M)	Status	Pred. pEC_{50} (Original model)	Status	Pred. pEC_{50} RBM	Status	Pred. pEC_{50} RSM	Status	Pred. pEC_{50} SEM
1	5.0460	Training	5.2880	Training	5.0711	Prediction	5.2655	Training	5.2987
2	5.8860	Training	5.3162	Prediction	5.1067	Prediction	5.1656	Training	5.3758
3	6.0970	Training	5.4203	Prediction	5.1851	Training	5.4203	Prediction	5.4266
4	5.5380	Training	5.1536	Prediction	5.0015	Prediction	5.0639	Training	5.2397
5	4.8120	Training	5.1292	Training	4.9953	Prediction	5.0477	Prediction	5.1723
6	5.2150	Training	5.2416	Training	5.0927	Prediction	5.0486	Training	5.3319
7	5.0760	Training	5.3873	Training	5.1679	Prediction	5.3175	Training	5.4200
8	5.0860	Training	5.2170	Training	5.0278	Training	5.1048	Prediction	5.2875
9	5.1610	Training	5.2564	Training	5.0749	Training	5.1851	Training	5.2775
10	5.5530	Training	5.2207	Prediction	5.0306	Training	5.1083	Training	5.2915
11	5.3670	Training	5.3308	Prediction	5.1010	Prediction	5.2687	Training	5.3750
12	5.0660	Training	5.1657	Training	5.0194	Prediction	4.9778	Prediction	5.2299
13	5.0410	Training	5.2956	Training	5.0944	Training	5.2224	Training	5.3125
14	4.8570	Training	5.2932	Training	5.0920	Training	5.2728	Training	5.2583
15	5.0660	Training	5.3473	Training	5.1502	Prediction	5.3307	Prediction	5.3120
16	4.9960	Training	5.1086	Training	4.9515	Prediction	4.9761	Training	5.0904
17	5.1670	Training	4.9870	Prediction	4.8743	Prediction	4.8687	Prediction	4.9219
18	5.0560	Training	5.0659	Training	4.9428	Prediction	4.8331	Training	5.0495
19	5.2150	Training	4.9953	Prediction	4.8925	Prediction	4.9007	Training	4.9143
20	4.9630	Training	5.1377	Training	4.9818	Training	5.0357	Prediction	5.1872
21	5.6990	Training	5.2073	Prediction	5.0408	Training	5.1339	Training	5.2368
22	5.4090	Training	5.6696	Training	5.3335	Training	5.6231	Training	5.7631
23	5.9210	Training	5.6557	Prediction	5.3315	Training	5.6635	Training	5.7037
24	5.9210	Training	5.7666	Prediction	5.4245	Training	5.6428	Training	5.8705
25	5.0560	Training	5.6574	Training	5.3471	Training	5.7013	Prediction	5.6881
26	5.0130	Training	5.1211	Training	4.9909	Prediction	5.2146	Prediction	5.0824
27	5.0270	Training	5.2824	Training	5.1005	Training	5.3394	Prediction	5.2222
28	4.9210	Training	5.0851	Training	4.9867	Training	5.0342	Prediction	5.0581
29	6.0000	Training	5.3765	Prediction	5.1800	Training	5.4363	Prediction	5.3123
30	5.5230	Training	5.6885	Training	5.5625	Prediction	5.7878	Training	5.7320
31	5.9210	Training	5.7055	Prediction	5.4872	Prediction	5.7108	Prediction	5.6846
32	5.8540	Training	5.8074	Prediction	5.6322	Training	5.7752	Training	5.8473
33	6.0000	Training	5.6413	Prediction	5.4524	Prediction	5.6397	Training	5.6148
34	5.0920	Training	5.3899	Training	5.1492	Training	5.4014	Training	5.4303
35	5.4320	Training	5.3296	Prediction	5.1263	Prediction	5.3885	Prediction	5.3132
36	5.1610	Training	5.4215	Training	5.1878	Training	5.3298	Prediction	5.4626
37	5.1670	Training	5.3199	Training	5.2029	Prediction	5.4170	Training	5.2786
38	5.3770	Training	5.4780	Training	5.2955	Training	5.5066	Prediction	5.4217
39	5.4810	Training	5.5637	Training	5.3845	Prediction	5.5192	Training	5.5493
40	5.0920	Training	5.4044	Training	5.2601	Prediction	5.4386	Training	5.3439
41	5.0920	Training	5.0119	Prediction	4.8827	Training	4.9975	Training	5.0372
42	5.4090	Training	5.2179	Prediction	5.0396	Training	5.2640	Prediction	5.2032
43	5.3280	Training	4.9977	Prediction	4.8990	Prediction	4.8683	Prediction	5.0198
44	5.4560	Training	5.2855	Prediction	5.1040	Training	5.3574	Prediction	5.2528
45	5.4090	Training	5.5518	Training	5.2783	Training	5.5371	Training	5.6367
46	6.1550	Training	5.7172	Prediction	5.4084	Training	5.7770	Prediction	5.7592
47	5.7960	Training	5.8079	Training	5.4795	Prediction	5.6956	Prediction	5.9094
48	6.0000	Training	5.7385	Prediction	5.4377	Training	5.8305	Training	5.7611

Table 6 continued

ID	pEC_{50} (M)	Status	Pred. pEC_{50} (Original model)	Status	Pred. pEC_{50} RBM	Status	Pred. pEC_{50} RSM	Status	Pred. pEC_{50} SEM
49	5.0760	Training	5.0362	Prediction	4.9532	Prediction	4.9090	Training	5.1036
50	5.7210	Training	5.2562	Prediction	5.0786	Prediction	5.1665	Prediction	5.2963
51	5.1190	Training	5.2491	Training	5.1047	Training	5.0739	Prediction	5.3336
52	5.0460	Training	5.2651	Training	5.1029	Training	5.1386	Training	5.2798
53	5.1190	Training	4.9765	Prediction	4.8580	Training	4.8949	Prediction	5.0200
54	5.6020	Training	5.5567	Prediction	5.2637	Training	5.5449	Prediction	5.6006
55	5.1870	Training	5.3790	Training	5.1589	Prediction	5.2824	Prediction	5.4672
56	5.0410	Training	5.5276	Training	5.2577	Training	5.5819	Training	5.5568
57	5.2150	Training	5.1070	Prediction	5.0042	Training	5.2056	Training	5.0763
58	5.0270	Training	5.1689	Training	5.0449	Training	5.2921	Training	5.1264
59	5.1550	Training	5.1929	Training	5.0573	Training	5.1803	Training	5.1693
60	5.5380	Training	5.3921	Prediction	5.1635	Prediction	5.4864	Prediction	5.3224
61	5.4090	Training	5.5410	Training	5.3063	Prediction	5.6685	Prediction	5.6176
62	5.5690	Training	5.5278	Prediction	5.2954	Training	5.4907	Training	5.6434
63	5.4440	Training	5.7260	Training	5.4307	Prediction	5.9172	Training	5.7908
64	5.1800	Training	5.3365	Training	5.1632	Training	5.3221	Prediction	5.4211
65	5.7700	Training	5.6413	Prediction	5.3814	Prediction	5.7366	Training	5.6995
66	5.2680	Training	5.3039	Training	5.1497	Training	5.2291	Prediction	5.3851
67	5.4950	Training	5.4599	Prediction	5.2491	Training	5.5817	Training	5.4745
68	4.9870	Training	5.2660	Training	5.0696	Training	5.2030	Prediction	5.3385
69	6.9590	Training	5.7388	Prediction	5.4117	Prediction	5.7975	Training	5.8029
70	5.1310	Training	5.3843	Training	5.1923	Prediction	5.2749	Prediction	5.4800
71	5.0320	Training	5.5372	Training	5.2608	Prediction	5.5536	Training	5.5362
72	5.4090	Training	5.5461	Training	5.3585	Prediction	5.5900	Prediction	5.5134
73	5.6780	Training	6.0272	Training	5.6627	Prediction	6.2168	Prediction	6.0480
74	5.6020	Training	5.6556	Training	5.4393	Training	5.6306	Prediction	5.6522
75	6.3980	Training	6.0451	Prediction	5.6965	Training	6.2229	Training	6.0713
76	4.9430	Training	4.8893	Prediction	4.8292	Training	4.8592	Training	4.8908
77	5.0660	Training	5.2310	Training	5.0376	Prediction	5.2929	Training	5.2048
78	5.0660	Training	4.9741	Prediction	4.8859	Training	4.9271	Training	4.9982
79	5.1190	Training	5.0173	Prediction	4.8858	Training	5.1317	Training	4.9584
80	5.0920	Training	5.1228	Training	4.9395	Prediction	5.2270	Training	5.1068
81	5.0860	Training	5.5045	Training	5.2698	Training	5.4702	Prediction	5.5902
82	5.0810	Training	5.1924	Training	4.9968	Prediction	5.3423	Prediction	5.1544
83	5.1430	Training	5.3475	Training	5.1368	Training	5.3156	Prediction	5.4093
84	5.6020	Training	5.4963	Prediction	5.2000	Training	5.5783	Training	5.5088
85	5.5230	Training	5.3221	Prediction	5.1095	Training	5.2794	Training	5.3884
86	6.0460	Training	5.6648	Prediction	5.3883	Training	5.7883	Prediction	5.6897
87	5.0600	Training	5.1604	Training	4.9940	Prediction	5.2306	Training	5.0745
88	5.5380	Training	5.4862	Prediction	5.2545	Training	5.6187	Training	5.4473
89	5.2760	Training	5.0955	Prediction	4.9946	Prediction	5.0613	Training	5.0430
90	5.4950	Training	5.5162	Training	5.2800	Prediction	5.6582	Training	5.4595
91	5.1940	Training	5.3851	Training	5.1992	Prediction	5.3311	Prediction	5.4108
92	5.5380	Training	5.1659	Prediction	5.0812	Prediction	4.9477	Training	5.2390
93	6.0000	Training	5.5070	Prediction	5.2861	Training	5.6562	Prediction	5.5464
94	5.6580	Training	5.4254	Prediction	5.2085	Prediction	5.3977	Training	5.5069
95	5.9210	Training	5.7454	Prediction	5.4344	Training	5.8117	Training	5.8043
96	5.0270	Training	5.4707	Training	5.2993	Training	5.3705	Training	5.5809

Table 6 continued

ID	pEC_{50} (M)	Status	Pred. pEC_{50} (Original model)	Status	Pred. pEC_{50} RBM	Status	Pred. pEC_{50} RSM	Status	Pred. pEC_{50} SEM
97	5.5530	Training	5.6583	Training	5.3984	Training	5.8009	Prediction	5.6963
98	5.1190	Training	5.6214	Training	5.3299	Prediction	5.7446	Prediction	5.6112
99	5.2290	Training	5.2426	Training	5.0457	Training	5.3492	Prediction	5.2081
100	5.2520	Training	5.4367	Training	5.1699	Training	5.3379	Training	5.4874

Table 7 Comparison of statistical parameters for original model for dataset-1, 2 and 3

Statistical Parameter	DataSet-1	DataSet-2	DataSet-3
R^2_{tr}	0.709	0.841	0.410
$R^2_{\mathrm{adj.}}$	0.670	0.837	0.378
R^2_{cv}	0.597	0.827	0.344
R^2_{LMO}	0.723	0.843	0.419
R^2_{Yrand}	0.120	0.242	0.050
Q^2_{Yrand}	−0.185	−0.488	−0.076
s	0.250	0.242	0.300
Kxx	0.425	0.245	0.208
ΔK	0.025	0.209	0.027
$RMSE_{tr}$	0.233	0.238	0.291
$RMSE_{cv}$	0.274	0.248	0.307
CCC_{tr}	0.830	0.914	0.581
CCC_{cv}	0.763	0.906	0.537
MAE_{tr}	0.193	0.172	0.226
MAE_{cv}	0.226	0.179	0.239
F	18.50	190.388	13.047
r^2	0.601	0.827	0.347
r^2_o	0.429	0.793	−0.574
$1\text{-}(r^2/r^2_o)$	0.285	0.040	2.653
r'^2_o	0.597	0.827	0.344
$1\text{-}(r^2/r'^2_o)$	0.007	0.000	0.009
k	0.996	0.999	0.997
k'	1.001	1.000	0.999

Table 8 Comparison of statistical parameters for original, residual-based rational, random splitting, and sphere exclusion models for dataset-1

Statistical Parameter	Original model	Residual-based model	Random splitting Model	Sphere exclusion model
R^2_{tr}	0.709	0.855	0.801	0.674
$R^2_{\mathrm{adj.}}$	0.670	0.813	0.743	0.578
R^2_{cv}	0.597	0.732	0.640	0.365
R^2_{LMO}	0.723	0.871	0.815	0.709
s	0.250	0.137	0.241	0.320
R^2_{ex}	–	0.845	0.418	0.722
R^2_{Yrand}	0.120	0.223	0.238	0.231
Q^2_{Yrand}	−0.185	−0.443	−0.421	−0.433
Kxx	0.425	0.457	0.452	0.446
ΔK	0.025	0.036	0.015	0.017
$RMSE_{tr}$	0.233	0.118	0.207	0.275
$RMSE_{cv}$	0.274	0.159	0.279	0.384
$RMSE_{ex}$	–	0.441	0.344	0.200
CCC_{tr}	0.830	0.922	0.890	0.805
CCC_{cv}	0.763	0.859	0.795	0.621
CCC_{ex}	–	0.606	0.611	0.845
MAE_{tr}	0.193	0.089	0.161	0.238
MAE_{cv}	0.226	0.122	0.220	0.325
MAE_{ex}	–	0.394	0.260	0.169
$Q^2{-}F^1$	–	0.443	0.266	0.706
$Q^2{-}F^2$	–	0.097	0.266	0.705
$Q^2{-}F^3$	–	−1.039	0.451	0.828
r^2m	–	0.762	0.290	0.655
$r^2m\ av$	–	0.678	0.270	0.612
$r^2m\ de$	–	0.168	0.040	0.085
F	18.50	20.120	13.714	7.023
r^2	0.601	0.845	0.418	0.722
r^2_o	0.429	0.757	0.256	0.677
$1\text{-}(r^2/r^2_o)$	0.285	0.105	0.386	0.062
r'^2_o	0.597	0.836	0.324	0.713
$1\text{-}(r^2/r'^2_o)$	0.007	0.011	0.223	0.012
k	0.996	0.916	0.974	1.005
k'	1.001	1.089	1.021	0.993

$RMSE_{cv}$, CCC_{tr}, CCC_{cv}, MAE_{tr}, MAE_{cv}, and $r^2_m\ av$ indicate good predictive ability and robust statistical performance of the residual-based model than the other models. The high value of r^2_m (=0.744) for residual model, though lower than sphere exclusion model (=0.804), indicates good external predictivity. A very high F (=328.459) value for residual-based model than the other models (=112.835 for random splitting and 111.362 for sphere exclusion model) indicates very high statistical significance of regression model. Similar to dataset-1, a large difference between $RMSE_{tr}$ (=0.143) and $RMSE_{ex}$ (=0.405) as well as between MAE_{tr} (=0.109) and MAE_{ex} (=0.353) suggests low

Table 9 Comparison of statistical parameters for original, residual-based rational, random splitting, and sphere exclusion models for dataset-2

Statistical Parameter	Original model	Residual-based model	Random splitting Model	Sphere exclusion model
R^2_{tr}	0.841	0.945	0.856	0.854
$R^2_{\mathrm{adj.}}$	0.837	0.942	0.848	0.847
R^2_{cv}	0.827	0.934	0.836	0.834
R^2_{LMO}	0.843	0.947	0.859	0.855
s	0.242	0.148	0.212	0.227
R^2_{ex}	–	0.877	0.842	0.816
R^2_{Yrand}	0.242	0.052	0.053	0.051
Q^2_{Yrand}	−0.488	−0.087	−0.086	−0.089
Kxx	0.245	0.246	0.293	0.220
ΔK	0.209	0.231	0.201	0.219
$RMSE_{tr}$	0.238	0.143	0.205	0.220
$RMSE_{cv}$	0.248	0.157	0.219	0.234
$RMSE_{ex}$	–	0.405	0.287	0.266
CCC_{tr}	0.914	0.972	0.922	0.921
CCC_{cv}	0.906	0.967	0.912	0.911
CCC_{ex}	–	0.777	0.876	0.893
MAE_{tr}	0.172	0.109	0.149	0.166
MAE_{cv}	0.179	0.119	0.160	0.177
MAE_{ex}	–	0.353	0.197	0.187
$Q^2{-}F^1$	–	0.553	0.809	0.822
$Q^2{-}F^2$	–	0.442	0.809	0.810
$Q^2{-}F^3$	–	0.561	0.717	0.788
r^2m	–	0.744	0.689	0.804
$r^2m\ av$	–	0.803	0.581	0.720
$r^2m\ de$	–	0.118	0.217	0.168
F	190.388	328.459	112.835	111.362
r^2	0.827	0.877	0.842	0.816
r^2_o	0.793	0.877	0.649	0.768
$1\text{-}(r^2/r^2_o)$	0.040	0.000	0.229	0.060
r'^2_o	0.827	0.854	0.809	0.816
$1\text{-}(r^2/r'^2_o)$	0.000	0.026	0.039	0.000
k	0.999	0.953	1.000	0.992
k'	1.000	1.048	0.998	1.007

Table 10 Comparison of statistical parameters for original, residual-based rational, random splitting, and sphere exclusion models for dataset-3

Statistical parameter	Original model	Residual-based model	Random splitting Model	Sphere exclusion model
R^2_{tr}	0.410	0.621	0.527	0.478
$R^2_{\mathrm{adj.}}$	0.378	0.583	0.478	0.426
R^2_{cv}	0.344	0.516	0.430	0.365
R^2_{LMO}	0.419	0.634	0.537	0.495
s	0.300	0.134	0.279	0.305
R^2_{ex}	–	0.662	0.237	0.280
R^2_{Yrand}	0.050	0.093	0.090	0.092
Q^2_{Yrand}	−0.076	−0.144	−0.153	−0.142
Kxx	0.208	0.176	0.232	0.203
ΔK	0.027	0.084	0.028	0.079
$RMSE_{tr}$	0.291	0.127	0.263	0.288
$RMSE_{cv}$	0.307	0.143	0.289	0.318
$RMSE_{ex}$	–	0.524	0.353	0.306
CCC_{tr}	0.581	0.766	0.690	0.647
CCC_{cv}	0.537	0.704	0.625	0.575
CCC_{ex}	–	0.339	0.480	0.488
MAE_{tr}	0.226	0.099	0.203	0.212
MAE_{cv}	0.239	0.112	0.225	0.236
MAE_{ex}	–	0.458	0.280	0.257
$Q^2{-}F^1$	–	0.239	0.112	0.246
$Q^2{-}F^2$	–	−0.728	0.105	0.238
$Q^2{-}F^3$	–	−5.512	0.145	0.414
r^2m	–	0.516	0.151	0.238
$r^2m\ av$	–	0.302	0.113	0.127
$r^2m\ de$	–	0.430	0.077	0.220
F	13.047	16.382	10.909	9.158
r^2	0.347	0.662	0.237	0.280
r^2_o	−0.574	−0.093	−0.235	−0.602
$1\text{-}(r^2/r^2_o)$	2.653	1.141	1.991	3.150
r'^2_o	0.344	0.614	0.106	0.257
$1\text{-}(r^2/r'^2_o)$	0.009	0.073	0.554	0.082
k	0.997	0.916	0.995	1.006
k'	0.999	1.089	1.001	0.257

generalizability of the residual-based model. In addition, the lower value of CCC_{ex}, $Q^2{-}F^1$, $Q^2{-}F^2$, and $Q^2{-}F^3$ for residual-based model than random splitting model, and sphere exclusion model points out low true external predictivity of this model.

A conceivable reason for the lower values of $Q^2{-}F^1$, $Q^2{-}F^2$, and $Q^2{-}F^3$ could be the presence of prediction set objects near the boundary of the training set (Chirico and Gramatica, 2011, 2012; Consonni *et al.*, 2009, 2010; Schuurmann *et al.*, 2008). Again, these statistical parameters are sensitive to mean of training and prediction sets, a simple analysis of Table 11 reveals that the mean of the test and the training sets of residual-based model have higher difference than the rest (Chirico and Gramatica 2011, 2012; Consonni *et al.*, 2009, 2010; Schuurmann *et al.*, 2008). This observation once again confirms that the distribution of the test and the training set has important impact on performance of many statistical parameters. Thus, the residual-based model is scoring high for many parameters suggesting statistical robustness of this model,

Table 11 Mean of experimental $p\mathrm{IC}_{50}$ for prediction and training sets of various models for datasets 1–3

DataSet	Set	Original	Residual-based model	Random splitting Model	Sphere exclusion model
1	Prediction	–	4.8308	4.6389	4.6509
	Training	4.6397	4.4652	4.6404	4.6295
2	Prediction	–	7.5342	7.3803	7.3017
	Training	7.3867	7.2633	7.3921	7.4577
3	Prediction	–	5.6296	5.3598	5.3570
	Training	5.3778	5.1799	5.3925	5.3941

but some parameters raise doubts on its external predictivity.

Results for the dataset-3

Various statistical parameters viz. R^2_{tr}, $R^2_{\mathrm{adj.}}$, R^2_{cv}, R^2_{LMO}, R^2_{Yrand}, *s*, R^2_{ex}, *Kxx*, ΔK, $RMSE_{tr}$, $RMSE_{cv}$, CCC_{tr}, CCC_{cv}, MAE_{tr}, MAE_{cv}, r^2_m *av*, and *F* (see Table 10) indicate low predictive ability and poor statistical performance of all the models. But, a closer inspection of various models indicates that the performance of residual-based model is better than the other models. Some of the statistical parameters like R^2_{tr}, R^2_{LMO}, *s*, R^2_{ex}, $RMSE_{tr}$, $RMSE_{cv}$, MAE_{tr}, and MAE_{cv} are with acceptable values. However, the model possesses low internal and external predictivity. As stated earlier, it is not a useful model at all for the prediction and pattern recognition. The $Q^2{-}F^2$ and $Q^2{-}F^3$ are negative which indicates that the model is useless for external predictivity.

Comparison of performance of splitting methodologies and statistical behavior of statistical parameters

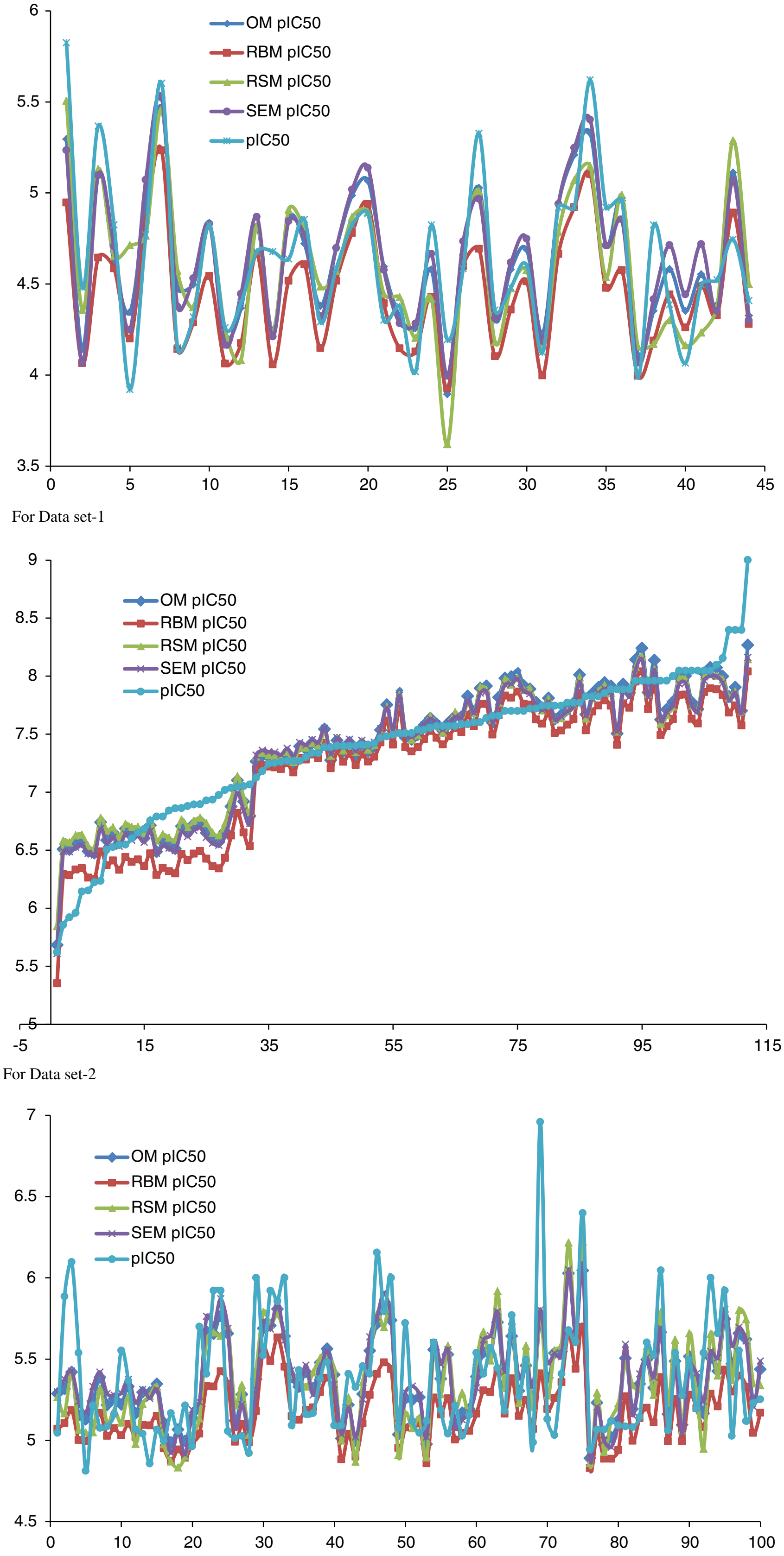
In the present analysis, information leakage was purposely performed for RBM. The descriptors were selected using the whole dataset, therefore, due to the information leakage, the selected descriptors must have captured the common structural features that influence the activity, and consequently, after splitting in any pattern/composition, the performance of RBM model for the all the datasets must be superior than SEM and RSM with respect to internal and external cross-validation parameters, i.e., must show high level of external predictivity with high validation score. Surprisingly, for RBM model, various validation parameters do not show expected behavior and values for all the datasets.

The random splitting models, for all the datasets, have varying performance; this could be due to the fact that during splitting the training or prediction set may not be covering the diversity of the whole dataset or the compounds are not close to each other. Repeating the random splitting several times is a good solution to arrive at best random splitting (Huang and Fan, 2011; Kiralj and Ferreira, 2009). Yet, as pointed out in a recent study, a QSAR model with high external predictivity for one prediction set does not necessarily indicate high accuracy for another external set (Huang and Fan, 2011). Therefore, precautions must be taken in using single random splitting.

Since, the performance of the RBM, RSM, and SEM models is varying, but by luck or due to rational splitting, the researcher may arrive at the training and prediction sets that indicate high external predictive ability, such situation, though, leads to a statistically robust but a misguiding QSAR model as observed in RBM. An easy and handy solution to this problem is to develop a model using undivided dataset and compare its performance with the other models. Herein, in all the datasets, the performance of original model, though not better than residual based and sphere exclusion models, is still statistically satisfactory. It is expected that a model developed with no prediction set will be most accurate and possess the highest coverage for external evaluation set. But, a recent study reports exactly opposite results in certain situations (Martin *et al*., 2012). Therefore, we recommend and accentuate reporting of a statistically robust QSAR model that is developed using undivided whole dataset and same set of descriptors, which were selected and used in splitting-based model. Then, such a model tells the true effect of inclusion of compounds in the dataset. That is, it is useful in understanding the effect of increase/decrease in size of dataset as well as for capturing less privileged yet useful structural features that govern the activity.

A higher value of R^2_{tr} for residual-based model in all the datasets than the rest of the models indicates a better fitting or explanation of variance (see Tables 8–10). Similar trend for R^2_{ex} for residual-based model for different datasets confers as if the residual based splitting is better method of splitting. Therefore, a QSAR modeler may consider the residual-based model superior than others. This apparent superiority can be attributed to the purposeful selection of the training and the prediction sets, that is, the method of splitting has significant impact on many statistical parameters. Moreover, a careful comparison of residual values for all the models (see Fig. 3) reveals that the difference between the experimental and predicted in many instances is large in case of residual-based model than the others. But, during the calculation of various statistical parameters either sum or average is used. Therefore, the statistical parameters are unable to recognize this serious pitfall. In fact, a QSAR model based on splitting method with an unusually robust training set R^2_{tr} of 0.8 or greater than R^2_{tr} of undivided set

Fig. 3 Difference between experimental and predicted $p\mathrm{IC}_{50}$ by various models for dataset-1, 2, and 3 (*X-axis*: Compound number and *Y-axis*: $p\mathrm{IC}_{50}/p\mathrm{EC}_{50}$; *X-axis*: Serial number of compound, *Y-axis*: pIC_{50} value)



For Data set-1

For Data set-2

For Data set-3

should be viewed with suspicion. Some parameters like CCC_{ex}, $Q^2{-}F^1$, $Q^2{-}F^2$, $Q^2{-}F^3$, and r_m^2 are more successful in identifying this crucial aspect. This can be ascribed to the method of calculation of these parameters (Chirico and Gramatica, 2011, 2012); (Consonni *et al.*, 2009, 2010; Schuurmann *et al.*, 2008).

$$Q^2 - F^1 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}i - yi)^2}{\sum_{i=1}^{n_{ext}} (yi - \bar{y}_{tr})^2}$$

$$Q^2 - F^3 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}i - yi)^2/n_{ext}}{\sum_{i=1}^{n_{tr}} (yi - \bar{y}_{tr})^2/n_{tr}}\,.$$

$$Q^2 - F^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}i - yi)^2}{\sum_{i=1}^{n_{ext}} (yi - \bar{y}_{ext})^2}$$

$$r_m^2 = r^2(1 - \sqrt{r^2 + r_o^2})$$

Thus, r_m^2 considers agreement between the actual and the predicted values as an essential factor to establish the true predictivity (Mitra *et al.*, 2010; Roy and Mitra, 2012). Thence, the statistical parameters viz. CCC_{ex}, $Q^2{-}F^1$, $Q^2{-}F^2$, $Q^2{-}F^3$, and r_m^2 reflect the factual performance of model regarding true external predictivity of a QSAR model. Therefore, these parameters should be used as criteria for selection of a consensus model, as in QSARINS v1.2. In QSARINS v1.2, MAE_{tr}, MAE_{ex}, $RMSE_{tr}$, $RMSE_{ex}$, CCC_{tr}, CCC_{ex}, $Q^2{-}F^1$, $Q^2{-}F^2$, $Q^2{-}F^3$, and some other parameters are used to find a consensus model.

In agreement with the previous reports, the trend of lower CCC with higher RMSE value is true for all the datasets (Chirico and Gramatica, 2011, 2012). However, the claim that the smaller the dataset size, the better the performances of r_m^2-EyPx and CCC compared to the other external validation measures was not observed for any of the dataset (Chirico and Gramatica, 2011, 2012). The similar values of $Q^2{-}F^1$and $Q^2{-}F^2$ for random splitting model for dataset 1 and 2 can be attributed to the fact that these parameters depend on agreement between the mean of the training and the prediction set values (Consonni *et al.*, 2009, 2010). For dataset 1 and 2, the mean of test and training sets values is very close to each other (see Table 11). A good difference between the mean of the undivided set and the training set values of the residual-based model for all the datasets indicates that the prediction set was not selected properly. Such a noticeable difference is absent in case of other models. This again indicates that residual-based method of splitting cannot be functionalised for splitting the dataset for external validation.

Since the whole dataset is involved in descriptor selection and model development, another point view toward the present approach is to consider it as a methodology to develop a model with good external predictivity using advantages of internal validation method. A model with good internal predictivity may or may not be good at external predictivity (Chirico and Gramatica, 2011, 2012; Consonni *et al.*, 2009, 2010; Gramatica 2013, Schuurmann *et al.*, 2008). In the present analysis, sphere exclusion model with higher values of $Q^2{-}F^3$, CCC_{ex}, and lower values of $RMSE_{ex}$, MAE_{ex} indicate good external predictivity of model.

Consonni et al. argued that increasing the mean of training set values increases Q^2 artificially (Consonni *et al.*, 2009). From Table 11, it is observed that the mean of the training set values for the random (for dataset- 1 and 2) and the sphere exclusion models (for dataset-1) is very close to mean of undivided set values; therefore, the value of Q^2 for these models should be close to Q^2 of the original model. However, for random splitting model, $Q^2 = 0.640$ for dataset-1, and $Q^2 = 0.836$ for dataset-2 are higher than that of the original model ($Q^2 = 0.597$). In addition, lower $Q^2 = 0.365$ for sphere exclusion model for dataset-1 than $Q^2 = 0.597$ for original model conflicts the finding of Consonni et al. The mean of the training set for residual model (=4.4652) is lower than mean of training set of undivided set (=4.6397). Therefore, for the sphere exclusion model, Q^2 should be lower than the Q^2 for original model, but the results are exactly opposite. Therefore, further studies are required to understand the effect of mean of training set on Q^2.

Conclusions

In conclusion, external validation based on single splitting is neither perfect nor absolutely accurate method of QSAR model validation as the statistical parameters can be influenced easily due to the biased and purposeful selection of the training and prediction sets. Moreover, the predictive ability of a QSAR model is sensitive toward the method of splitting and its manipulation is feasible. Thus, it is still insufficient to guarantee the true predictability of a QSAR model. The true external predictivity of any QSAR model cannot be decided on the basis of one or two parameters, that is, as many as possible statistical parameters should be calculated to judge the external predictivity. A good number of statistical parameters need to be calculated and presented to identify the true external predictivity of any QSAR model. We suggest and emphasize reporting of at least one statistically robust QSAR model that is developed using undivided whole dataset with appropriate cross validation.

In the present study, we presented a novel method for splitting the dataset for external validation. The residual method, though, generates statistically robust model but with low external predictivity. Further studies are in progress for the improvement of this method.

Acknowledgments We are thankful to QSARINS and RapidMiner developing teams for providing the evaluation and free versions of the softwares. One of the authors (VHM) is thankful to Dr. Paola Gramatica, Italy for providing QSARINS v1.2 and later versions.

References

Baumann K, Stiefl N (2004) Validation tools for variable subset regression. J Comput Aided Mol Des 18(7–9):549–562

Chirico N, Gramatica P (2011) Real external predictivity of qsar models: how to evaluate it? comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51(9):2320–2335

Chirico N, Gramatica P (2012) Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model 52(8):2044–2058

Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model 49(7):1669–1678

Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predictive ability by external validation techniques. J Chemomet 24:194–201

Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20(4):269–276

Gramatica P (2013) On the development and validation of QSAR models. Methods Mol Biol 930:499–526

Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem 34(24):2121–2132

Gramatica P, Cassani S, Chirico N (2014) QSARINS-chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J Comput Chem 35(13):1036–1044

Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44(1):1–12

Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross-validation. J Chem Inf Comput Sci 43:579–586

Hawkins DM, Kraker JJ, Basak SC, Mills D (2008) QSPR checking and validation: a case study with hydroxy radical reaction rate constant. SAR QSAR Environ Res 19(5–6):525–539

Huang J, Fan X (2011) Why QSAR fails: an empirical evaluation using conventional computational approach. Mol Pharm 8(2):600–608

Hwang JY, Kawasuji T, Lowes DJ, Clark JA, Connelly MC, Zhu F, Guiguemde WA, Sigal MS, Wilson EB, DeRisi JL, Guy RK (2011) Synthesis and evaluation of 7-substituted 4-aminoquinoline analogues for antimalarial activity. J Med Chem 54(20): 7084–7093

Kiralj R, Ferreira MMC (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J Braz Chem Soc 20:770–787

Kubinyi H (2002) From narcosis to hyperspace: the history of QSAR. Quant Struct Act Relat 21:348–356

Mahajan DT, Masand VH, Patil KN, Ben Hadda T, Jawarkar RD, Thakur SD, Rastija V (2012) CoMSIA and POM analyses of anti-malarial activity of synthetic prodiginines. Bioorg Med Chem Lett 22(14):4827–4835

Mahajan DT, Masand VH, Patil KN, Hadda TB, Rastija V (2013) Integrating GUSAR and QSAR analyses for antimalarial activity of synthetic prodiginines against multi drug resistant strain. Med Chem Res 22:2284–2292

Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, Tropsha A (2012) Does rational selection of training and test sets improve the outcome of QSAR modeling? J Chem Inf Model 52(10):2570–2578

Masand VH, Jawarkar RD, Patil KN, Nazerruddin GM, Bajaj SO (2010) Correlation potential of Wiener index and molecular refractivity vis-a‘-vis Antimalarial activity of xanthone derivatives. Org Chem 6(1):30–38

Masand VH, Jawarkar RD, Mahajan DT, Hadda TB, Sheikh J, Patil KN (2012) QSAR and CoMFA studies of biphenyl analogs of the anti-tuberculosis drug (6S)-2-nitro-6-{[4-(trifluoromethoxy) benzyl]oxy}-6,7-dihydro-5H-imidazo[2,1-b][1,3]oxazine(PA-824). Med Chem Res 21:2624–2629

Masand VH, Mahajan DT, Patil KN, Hadda TB, Youssoufi MH, Jawarkar RD, Shibi IG (2013) Optimization of antimalarial activity of synthetic prodiginines: QSAR, GUSAR, and CoMFA analyses. Chem Biol Drug Des 81(4):527–536

Masand VH, Mahajan DT, Gramatica P, Barlow J (2014) Tautomerism and multiple modelling enhance the efficacy of QSAR: antimalarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl. Med Chem Res

Mitra I, Roy PP, Kar S, Ojha PK, Roy K (2010) On further application of r m2 as a metric for validation of QSAR models. J Chemomet 24(1):22–33

Roy K, Mitra I (2012) On the use of the metric rm(2) as an effective tool for validation of QSAR models in computational drug design and predictive toxicology. Mini Rev Med Chem 12(6):491–504

Roy K, Roy PP, Leonard JT (2008) Exploring the impact of size of training sets for the development of predictive QSAR models. Chemomet Intel Lab Sys 90:31–42

Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17(5):4791–4810

Schuurmann G, Ebert RU, Chen J, Wang B, Kuhne R (2008) External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. J Chem Inf Model 48(11):2140–2145

Scior T, Medina-Franco JL, Do QT, Martinez-Mayorga K, Yunes Rojas JA, Bernard P (2009) How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem 16(32):4297–4313

Selassie CD (2003) History of Quantitative Structure-Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, 6 ed.; Abraham, D. J., Ed. JohnWiley&Sons, Inc.: 2003; Vol. 1

Sushko I, Novotarskyi S, Korner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Muller KR, Xi L, Liu H, Yao X, Oberg T, Hormozdiari F, Dao P, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Muller C, Varnek A, Prokopenko VV, Tetko IV (2010) Applicability domains for classification problems: benchmarking of distance to models for Ames mutagenicity set. J Chem Inf Model 50(12):2094–2111

Todeschini R, Consonni V, Mauri A, Pavan M (2004) Detecting "bad" regression models: multicriteria fitness functions in regression analysis. Anal Chim Acta 515(1):199–208

Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488

Turcotte V, Fortin S, Vevey F, Coulombe Y, Lacroix J, Cote MF, Masson JY, R CG (2012) Synthesis, biological evaluation, and structure-activity relationships of novel substituted N-phenyl ureidobenzenesulfonate derivatives blocking cell cycle progression in S-phase and inducing DNA double-strand breaks. J Med Chem 55(13):6194–6208

Van Drie JH (2007) Computer-aided drug design: the next 20 years. J Comput Aided Mol Des 21(10–11):591–601

Yuriev E, Agostino M, Ramsland PA (2011) Challenges and advances in computational docking: 2009 in review. J Mol Recognit 24(2):149–164