JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

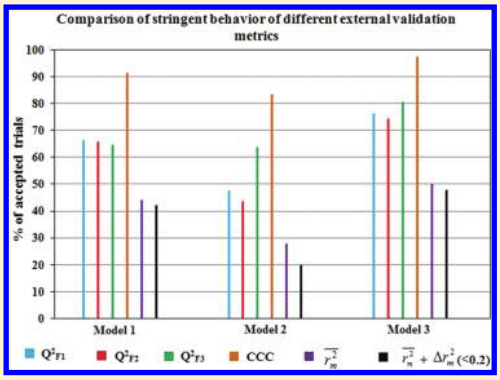
Comparative Studies on Some Metrics for External Validation of QSPR Models

Kunal Roy,* Indrani Mitra, Supratik Kar, Probir Kumar Ojha, Rudra Narayan Das, and Humayun Kabir

Drug Theoretics and Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

Supporting Information

ABSTRACT: Quantitative structure−property relationship (QSPR) models used for prediction of property of untested chemicals can be utilized for prioritization plan of synthesis and experimental testing of new compounds. Validation of QSPR models plays a crucial role for judgment of the reliability of predictions of such models. In the QSPR literature, serious attention is now given to external validation for checking reliability of QSPR models, and predictive quality is in the most cases judged based on the quality of predictions of property of a single test set as reflected in one or more external validation metrics. Here, we have shown that a single QSPR model may show a variable degree of prediction quality as reflected in some variants of external validation metrics like Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, CCC, and $\overline{r^2_m}$ (all of which are differently modified forms of predicted variance, which theoretically may attain a maximum value of 1), depending on the test set composition and test set size. Thus, this report questions the appropriateness of the common practice of the "classic" approach of external validation based on a single test set and thereby derives a conclusion about predictive quality of a model on the basis of a particular validation metric. The present work further demonstrates that among the considered external validation metrics, $\overline{r^2_m}$ shows statistically significantly different numerical values from others among which CCC is the most optimistic or less stringent. Furthermore, at a given level of threshold value of acceptance for external validation metrics, $\overline{r^2_m}$ provides the most stringent criterion (especially with Δr^2_m at highest tolerated value of 0.2) of external validation, which may be adopted in the case of regulatory decision support processes.



INTRODUCTION

Various structural representations of chemical compounds encode different types of chemical information that have definite quantitative relationships with properties exhibited by the compounds.[1,2] Statistical tools can be used to develop relationships between chemical information (in the form of numerical quantities called descriptors) and the properties. This kind of exercise is termed as quantitative structure−property relationship (QSPR) modeling.

One of the major applications of QSPR models is prediction of properties of untested chemicals thereby helping the prioritization plan for further synthesis and testing, leading to significant gain in resources in terms of material, manpower, money, and time.[3] For this purpose, a rigorous check of reliability of the models intended to be applied on new chemicals is an important issue. Validation strategies seek to explore the reliability of the developed models.[4−6] Among the popular validation techniques for QSPR models, two important ones are internal validation [mostly leave-one-out (LOO) validation or the jackknife test] and external or test set validation. In the case of the former, the original data set used for the development of the model is used for the validation purpose. One compound is omitted from the data set in each turn (in cases of LOO validation), and a new model is generated using the reduced set of data. This new model is used for prediction of the property of the omitted compound.[7−9] This continues until all the compounds have been omitted once from the data set. The jackknife test uses all available data for model development and validation, and thus, makes the model more reliable simply because of the use of a larger number of compounds than in the case where splitting of the data set is performed for external validation purposes.[8] This issue is more important in the case of a small data set, as a significant amount of information is lost due to omission of some compounds from the training set for external validation. Another method of internal validation is the leave-many-out method like *x*-fold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a data set is an astronomical figure;[10,11] in actual cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Because different selections will always lead to different results even for the same benchmark data set and same predictors, the subsampling test cannot avoid the arbitrariness. Though internal validation tools like the jackknife test have been criticized by some groups of authors,[5,12,13] these

Received: October 31, 2011
Published: December 27, 2011

dx.doi.org/10.1021/ci200520g | *J. Chem. Inf. Model.* 2012, 52, 396−408

have been increasingly and widely used by other investigators to examine the quality of various predictive models.[14] However, in general, external validation (or test set validation) has been considered as the most conclusive proof of reliability of the developed model in the QSPR literature. In the case of the test set validation, a new set that has not been used for model development is employed for prediction to check the reliability of the developed model.[12,13] In most cases, with truly new data being unavailable, the original data set is divided into a training set (used for model development) and a test set (used for checking reliability of the developed models). The same principles of model validation are also applicable for developing quantitative structure−activity relationship (QSAR) models for biological activity endpoints.[15,16]

In the QSPR literature, external validation of any model is mostly done on a test set selected from the original data set, and quality of the model is judged on the basis of different metrics of external validation calculated from the single test set.[12,17] Moreover, the size of the test set in most of the cases is only a limited fraction of the total data set size. In the present paper, we have studied reliability of this "classic" approach and comment on limitations of the conclusions drawn from the values of the external validation metrics at different threshold levels obtained from the predictions for a single test set. We have also compared different external validation metrics for their stringent behavior in terms of their numerical values. Here, we have focused on different metrics of external validation such as Q^2_{F1},[18] Q^2_{F2},[18] Q^2_{F3},[17] CCC,[19,20] and $\overline{r^2_m}$,[21] all of which may attain a maximum value of 1. The nearer to one the value of a metric is, the better the quality of the model in terms of external validation. However, the model can be considered of poor predictivity for obvious reasons if the value of a validation metric is lower than even 0.5. Along with $\overline{r^2_m}$, the Δr^2_m values[21] have also been noted. Details of these metrics have been discussed in the Supporting Information. The value of Δr^2_m should ideally be close to zero and conventionally taken to be lower than 0.2 for an acceptable model.[21] For external validation, a set of criteria proposed by Golbraikh and Tropsha[12] are also often considered. However, we have omitted these in our present comparison study as these do not represent any single metric.

MATERIALS AND METHODS

For the present work, we have chosen a large data set[22] with the adsorption capacities of 3483 diverse organic compounds with activated carbon in the gas phase (Table S1 of the Supporting Information). The present computational study has been conducted in three separate cycles for which we have selected three different training sets of 2000 compounds. In each case, the remaining 1483 compounds served as a pool of test set compounds. Note that in each of the three cycles of this study, we have kept the number of training set compounds the same, while altering the composition of test sets. Also, we have developed one single model in each cycle and validated the model externally using different test sets of varying size and composition to understand the impact of test sets in determining the quality of the external validation metrics. The objective of the present study is not to develop a sound model with good interpretability, rather it is to question the practice of external validation against a particular test set and to reach a conclusion about the quality of the model from the "classic" type external validation.

We have used a set of topological, structural, thermodynamic, and spatial descriptors[23] calculated using Cerius 2 version 4.10 software.[24] We have also used ETA descriptors calculated using Dragon software version 6.0.[25] The list of descriptors used in this study has been presented in Table 1. The whole descriptor matrix has also been uploaded in Supporting Information. For the selection of training set compounds, we have used three strategies in three cycles: cluster-based division (k-means clustering),[26,27] sorted response, and random division. In each cycle of the experiment, we have generated 10 test sets each of varying size, 100, 200, 300, 400, and 500, i.e., we have generated a total of 50 test sets for the single training set in each cycle of the experiment. The composition of the test sets of three cycles of the experiment is uploaded in the Supporting Information.

We have used a stepwise multiple linear regression (stepwise MLR)[28,29] approach for the development of the model in each cycle. We have used the objective function F-to-enter = 70 and F-to-remove = 69.9 in all three cycles to keep the number of descriptors appearing in the MLR model to a manageable size. We have not tried to optimize the equation quality in terms of the usual metrics[29] like R^2, adjusted R^2, and internal validation parameters[7−9,30] like Q^2 and predicted residual sum of squares (PRESS) as this was not the objective of our study. In a given cycle of the experiment, a single model was used for prediction of the response of the test set compounds, and accordingly, the quality of the model was judged in terms of different external validation metrics as listed previously. As there are 50 combinations of test set compounds for each training set, a single model is expected to show a range of external validation metric values for different test sets depending on composition of the test sets and test set size. As there were 10 test sets for a particular test set size, we applied two-way analysis of variance[29] (ANOVA) to explore any statistically significant differences among the external validation metrics and also among different trials. In this procedure, we have not considered Δr^2_m for obvious reasons.[21] We have also applied the least significant difference[31] (LSD) approach for multiple comparisons of the external validation metrics. In each cycle, we have also used average values of different validation metrics obtained from 10 trials for a given test set size for ANOVA to explore any possible impact of test set size.

RESULTS AND DISCUSSION

Three training sets were generated on the basis of cluster-based division, sorted response, and random division approaches. One stepwise MLR equation was generated from each of the three training sets. The models obtained from the cluster-based division, sorted response, and random division approaches show Q^2_{LOO} values of 0.823, 0.791, and 0.828 respectively. In each case, the model was externally validated using 50 test sets of different sizes, and the quality of external predictivity varied to a great extent as reflected in the values of various metrics depending on the composition and size of the test sets. The models are included in the Supporting Information. The results obtained from the three cycles of the experiment are shown in Tables 2, 3, and 4.

Results Obtained for Training Set 1 (Cluster-Based Division). The first training set is derived from the whole data set on the basis of a cluster-based approach. From the pool of the test sets, 50 test sets are chosen with 10 sets each at different test set sizes like 100, 200, 300, 400, and 500. The model derived from the training set is evaluated for its predictive quality using 50 test sets. Table 2 shows that variable

Table 1. List of Descriptors Used in the QSPR Analysis

category of descriptors	names and notations of the descriptors
topological	Balaban (Jx), Kappa shape (${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$, ${}^1\kappa_{am}$, ${}^2\kappa_{am}$, ${}^3\kappa_{am}$), Flexibility (φ), Subgraph count (SC-0, SC-1, SC-2, SC-3_P, SC-3_C, SC-3_CH), Connectivity (${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_p$, ${}^3\chi_c$, ${}^3\chi_{cbr}$, ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi^v_p$, ${}^3\chi^v_c$, ${}^3\chi^v_{ch}$), Wiener (W), Zagreb, E-state (S_sCH$_3$, S_dCH$_2$, S_ssCH$_2$, S_tCH, S_dsCH, S_aaCH, S_ddC, S_tsC, S_aasC, S_aaaC, S_ssNH$_2$, S_ssNH, S_tN, S_dsN, S_aaN, S_sssCH, S_dssC, S_aaO, S_sSH, S_sssssC, S_sssN, S_sOH, S_dO, S_ssO, S_sssS, S_aaS, S_sF, S_sCl, S_sBr, S_sI)
structural	MW, Rotlbonds, Hbond acceptor, Hbond donor
thermodynamic	AlogP, AlogP98, MR, MolRef, LogP
electronic	Dipole-mag
spatial	RadOfGyration, Jurs (Jurs_SASA, Jurs_PPSA_1, Jurs_PNSA_1, Jurs_DPSA_1, Jurs_PPSA_2, Jurs_PNSA_2, Jurs_DPSA_2, Jurs_PPSA_3, Jurs_PNSA_3, Jurs_DPSA_3, Jurs_FPSA_1, Jurs_FNSA_1, Jurs_FPSA_2, Jurs_FNSA_2, Jurs_FPSA_3, Jurs_FNSA_3, Jurs_WPSA_1, Jurs_WNSA_1, Jurs_WPSA_2, Jurs_WNSA_2, Jurs_WPSA_3, Jurs_WNSA_3, Jurs_RPCG, Jurs_RNCG, Jurs_RPCS, Jurs_RNCS, Jurs_TPSA, Jurs_TASA, Jurs_RPSA, Jurs_RASA), Area, Vm, Density, PMI_mag.
ETA indices (1st generation)	$\Sigma\alpha$, $\Sigma\alpha/N_v$, $\Sigma\varepsilon$, $\Sigma\varepsilon/N$, $\Sigma\beta_s$, $\Sigma\beta'_s$, $\Sigma\beta_{ns}$, $\Sigma\beta'_{ns}$, $\Sigma\beta$, $\Sigma\beta'$, η, η', η^{local}, $[\eta']^{local}$, η_F, η'_F, η_F^{local}, $[\eta']_F^{local}$, η_B, η'_B, $[\Sigma\alpha]_P/\Sigma\alpha$, $[\Sigma\alpha]_Y/\Sigma\alpha$, $[\Sigma\alpha]_X/\Sigma\alpha$
ETA indices (2nd generation)	$\Delta\alpha_A$, $\Delta\alpha_B$, ε_1, ε_2, ε_3, ε_4, ε_5, $\Delta\varepsilon_A$, $\Delta\varepsilon_B$, $\Delta\varepsilon_C$, $\Delta\varepsilon_D$, ψ_1, $\Delta\psi_A$, $\Delta\psi_B$, $\Delta\beta$, $\Delta\beta'$, $\sum\beta_{ns(\delta)}$, $\beta'_{ns(\delta)}$

ranges of values are obtained for different external validation metrics at a given size of the test sets. For example, for the test set size of 100, the value of Q^2_{F1} is as high as 0.869 (trial 7) and as low as 0.581 (trial 2). Again, in the case of trial 2, the values of Q^2_{F1}, Q^2_{F2}, and $\overline{r^2_m}$ are low (0.581, 0.579, and 0.566, respectively), while those of Q^2_{F3} and CCC are quite high (0.799 and 0.835, respectively). A two-way ANOVA suggests that statistically significant differences exist among the external validation metrics and also among the trials. A LSD analysis suggests that $\overline{r^2_m}$ is statistically significantly different from other metrics at $p = 0.05$, while there are no significant differences among Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}. The metric CCC is different from all other metrics, and it has the highest mean values for the 10 trials. It is interesting to note that performance of a single model for predicting different test sets varies to a large range in terms of the quality of external validation. Thus, it would be unfair to comment on the predictive quality of a model on the basis of the values of external validation metric(s) obtained from a single test set.

Similar observations are also made for other test set sizes. For example, at the test set size of 200, the value of Q^2_{F1} is very good for trial 19 (0.844) and very bad for trial 15 (0.182). For trial 15, though the values of Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and $\overline{r^2_m}$ are quite low (0.182, 0.133, 0.394, and 0.146, respectively), the value of CCC crosses 0.5 (0.521). ANOVA and multiple comparison suggest that $\overline{r^2_m}$ is significantly different from other metrics at $p = 0.05$, while CCC always shows an over-optimistic value.

Finally, we performed ANOVA on the mean values of the metrics for the 10 trials for each test set size. The results suggest that the external validation metrics are significantly different at $p < 0.05$ and also that there is a significant role of test set size in determining the quality of the model in terms of external validation. Multiple comparison using the LSD approach again confirmed $\overline{r^2_m}$ as the strictest metric in terms of its numerical value, which is significantly different from others among which CCC is the most optimistic or least precautionary.

Results Obtained for Training Set 2 (Sorted Response-Based Division). For the training set obtained using the sorted response approach, we created 10 test sets each for different test set sizes like 100, 200, 300, 400, and 500. The model developed from the training set using the stepwise regression approach was validated externally using 50 test sets as described above. The detailed results of the external validation experiment are shown in Table 3. The single model showed a varying pattern in the values of a single external validation metric as exemplified by trials 2 and 9 with Q^2_{F1} values of 0.887 and 0.157, respectively. It is shocking to note that a single model performs well in external predictivity as evidenced by the value of the metric Q^2_{F1} for the test set corresponding to trial 2, while it performs very poorly for another test set corresponding to trial 9. Similar observations may also be made considering other metrics of external validation and other test set sizes. Again, interestingly, for a single test set (trial 1), the model shows drastic variation in terms of the external validation metrics as evidenced from an acceptable value of Q^2_{F1} (0.727), an unacceptably low value of $\overline{r^2_m}$ (0.343), and a negative value of Q^2_{F3} (−1.416). A similar observation is also made at trial numbers 17, 18, 25, and 29 among others. Thus, it becomes quite difficult to comment on the predictive quality of the model on the basis of a single test set and a single validation metric. In the case of ANOVA applied to the data matrices

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396−408

Table 2. Results of External Validation for the Model Obtained from Training Set 1 (Cluster-Based Division)

		external validation metrics					additional metric[a]		slope and intercept terms			
test set trials	% of test set compounds outside the applicability domain	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	$\overline{r^2_{m\,(test)}}$	$\Delta r_m{}^2{}_{(test)}$	ANOVA and multiple comparison[b]	k	k'	Δm	Δc
						N_{test} = 100						
1	6.00	0.860	0.860	0.870	0.926	0.802	0.093	F_1 = 13.183 (*df* 9, 36)	1.011	0.982	0.125	0.147
2	3.00	0.581	0.579	0.799	0.835	0.566	0.243	F_2 = 27.383 (*df* 4, 36)	1.003	0.986	0.376	0.593
3	7.00	0.740	0.740	0.776	0.861	0.645	0.096	LSD (external validation metrics) = 0.03547 (p = 0.05)	1.006	0.981	0.130	0.163
4	6.00	0.763	0.762	0.662	0.867	0.667	0.167	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, $^2_{F2}$), $\overline{r^2_m}$	1.009	0.972	0.217	0.281
5	9.00	0.822	0.813	0.825	0.907	0.756	0.029		1.015	0.975	0.006	0.048
6	5.00	0.806	0.804	0.811	0.896	0.726	0.115		0.988	1.001	0.113	0.197
7	4.00	0.869	0.858	0.876	0.939	0.790	0.101		0.977	1.017	0.272	0.382
8	3.00	0.839	0.837	0.864	0.911	0.763	0.145		1.003	0.989	0.172	0.235
9	4.00	0.806	0.806	0.897	0.904	0.751	0.014		0.981	1.014	0.015	0.025
10	8.00	0.760	0.760	0.720	0.859	0.640	0.207		1.011	0.973	0.307	0.416
						N_{test} = 200						
11	7.50	0.798	0.794	0.754	0.888	0.713	0.111	F_1 = 45.033 (*df* 9, 36)	1.011	0.974	0.154	0.178
12	4.50	0.820	0.812	0.836	0.911	0.763	0.063	F_2 = 17.271 (*df* 4, 36)	0.983	1.009	0.114	0.108
13	3.50	0.757	0.756	0.849	0.879	0.681	0.012	LSD (external validation metrics) = 0.0507 (p = 0.05)	0.994	0.997	0.017	0.020
14	6.00	0.818	0.812	0.753	0.899	0.736	0.093	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F1}, Q^2_{F3}, Q^2_{F2}), $\overline{r^2_m}$	1.008	0.977	0.128	0.147
15	2.50	0.182	0.133	0.394	0.521	0.146	0.052		0.997	0.971	0.112	0.147
16	4.50	0.822	0.822	0.894	0.902	0.744	0.155		1.007	0.987	0.186	0.252
17	4.00	0.798	0.790	0.738	0.871	0.607	0.207		0.993	0.990	0.431	0.632
18	7.00	0.807	0.807	0.634	0.888	0.670	0.176		0.985	0.994	0.296	0.482
19	7.00	0.844	0.838	0.837	0.913	0.767	0.115		1.003	0.987	0.136	0.178
20	6.50	0.795	0.793	0.737	0.879	0.655	0.189		0.990	0.994	0.308	0.477
						N_{test} = 300						
21	6.33	0.781	0.781	0.760	0.884	0.696	0.083	F_1 = 32.469 (*df* 9, 36)	1.003	0.983	0.109	0.137
22	3.33	0.828	0.827	0.875	0.918	0.779	0.074	F_2 = 18.574 (*df* 4, 36)	0.988	1.005	0.000	0.000
23	4.67	0.549	0.545	0.454	0.754	0.437	0.080	LSD (external validation metrics) = 0.0501 (p = 0.05)	0.995	0.971	0.122	0.153
24	3.33	0.805	0.799	0.894	0.894	0.725	0.079	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$	1.012	0.982	0.114	0.131
25	5.00	0.807	0.800	0.646	0.879	0.629	0.192		0.984	0.994	0.384	0.599
26	8.33	0.863	0.862	0.844	0.930	0.807	0.021		0.996	0.994	0.023	0.033
27	4.00	0.787	0.786	0.792	0.884	0.699	0.126		0.998	0.990	0.146	0.209
28	3.00	0.352	0.329	0.582	0.657	0.287	0.032		0.979	0.999	0.034	0.081
29	3.00	0.809	0.807	0.778	0.885	0.659	0.184		1.001	0.986	0.346	0.499
30	5.00	0.866	0.865	0.863	0.924	0.769	0.125		1.013	0.979	0.243	0.317
						N_{test} = 400						

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396−408

Table 2. continued

test set trials	% of test set compounds outside the applicability domain	external validation metrics					additional metric[a]	ANOVA and multiple comparison[b]	slope and intercept terms			
		Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	$\overline{r^2_{m\,(test)}}$	$\Delta r_m{}^2{}_{(test)}$		k	k'	Δm	Δc
31	5.50	0.808	0.807	0.796	0.903	0.738	0.013	F_1 = 27.415 (*df* 9, 36)	0.996	0.992	0.018	0.022
32	4.25	0.809	0.806	0.813	0.898	0.729	0.076	F_2 = 24.636 (*df* 4, 36)	1.001	0.987	0.094	0.121
33	3.00	0.429	0.417	0.635	0.684	0.326	0.071	LSD (external validation metrics) = 0.0397 (p = 0.05)	1.003	0.977	0.121	0.153
34	6.50	0.815	0.813	0.736	0.895	0.711	0.166	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$	0.995	0.989	0.226	0.335
35	4.25	0.828	0.828	0.824	0.907	0.754	0.134		0.993	0.997	0.141	0.223
36	2.75	0.822	0.821	0.900	0.911	0.756	0.000		0.999	0.995	0.003	0.000
37	4.75	0.845	0.841	0.757	0.905	0.692	0.155		1.011	0.975	0.367	0.498
38	4.00	0.823	0.822	0.848	0.899	0.685	0.105		1.001	0.990	0.245	0.349
39	5.75	0.853	0.852	0.846	0.921	0.787	0.127		0.996	0.995	0.135	0.205
40	6.25	0.825	0.823	0.742	0.899	0.699	0.163		0.994	0.991	0.275	0.419
						N_{test} = 500						
41	4.80	0.808	0.808	0.810	0.902	0.737	0.030	F_1 = 57.019 (*df* 9, 36)	0.997	0.992	0.037	0.050
42	5.00	0.599	0.598	0.613	0.784	0.489	0.105	F_2 = 78.347 (*df* 4, 36)	0.982	0.995	0.120	0.203
43	3.20	0.834	0.833	0.812	0.903	0.716	0.154	LSD (external validation metrics) = 0.0276 (p = 0.05)	1.015	0.974	0.300	0.393
44	6.00	0.803	0.800	0.732	0.888	0.703	0.174	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}), $\overline{r^2_m}$	0.996	0.988	0.219	0.320
45	3.40	0.867	0.867	0.890	0.928	0.799	0.122		0.994	0.999	0.139	0.222
46	6.00	0.639	0.639	0.544	0.800	0.523	0.148		0.983	0.991	0.172	0.280
47	5.60	0.611	0.610	0.596	0.791	0.502	0.098		0.984	0.994	0.113	0.191
48	5.60	0.587	0.585	0.591	0.779	0.478	0.086		0.987	0.990	0.105	0.169
49	6.00	0.640	0.640	0.626	0.816	0.547	0.043		0.983	0.995	0.041	0.082
50	4.20	0.625	0.625	0.636	0.808	0.531	0.040		0.985	0.994	0.040	0.076
						mean (±s.e.) values at a particular test set size						
	N_{test} = 100	0.785 ± 0.026	0.782 ± 0.026	0.810 ± 0.024	0.890 ± 0.011	0.710 ± 0.024		F_1 = 51.855 (*df* 4, 16)				
	N_{test} = 200	0.744 ± 0.063	0.736 ± 0.067	0.743 ± 0.045	0.855 ± 0.037	0.648 ± 0.058		F_2 = 216.427 (*df* 4, 16)				
	N_{test} = 300	0.745 ± 0.052	0.740 ± 0.054	0.749 ± 0.046	0.861 ± 0.028	0.649 ± 0.052		LSD (external validation metrics) = 0.0149 (p = 0.05)				
	N_{test} = 400	0.786 ± 0.040	0.783 ± 0.041	0.790 ± 0.024	0.882 ± 0.022	0.688 ± 0.041		ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$				
	N_{test} = 500	0.701 ± 0.035	0.701 ± 0.035	0.685 ± 0.037	0.840 ± 0.018	0.603 ± 0.038						

[a]Not included in ANOVA. [b]F_1 corresponds to the F value between rows (i.e., test set trials in the case of a data array corresponding to a particular test set size or test set size in the case of the data array corresponding to the mean values). F_2 corresponds to the F value between columns (i.e., external validation metrics). Critical F values (p = 0.05): 2.153 (*df* 9, 36); 2.634 (*df* 4, 36); 3.007 (*df* 4, 16). [c]Two means not included within the same parentheses are statistically significantly different at p = 0.05.

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396–408

Table 3. Results of External Validation for the Model Obtained from Training Set 2 (Sorted Response-Based Division)

		external validation metrics					additional metric[a]		slope and intercept terms			
test set trials	% of test set compounds outside the applicability domain	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	$\overline{r^2_{m\,(test)}}$	$\Delta r_m{}^2{}_{(test)}$	ANOVA and multiple comparison[b]	k	k'	Δm	Δc
						N_{test} = 100						
1	9.00	0.727	0.666	−1.416	0.768	0.343	0.351	F_1 = 1.828 (df 9, 36)	0.933	0.908	0.894	1.449
2	6.00	0.887	0.852	0.690	0.919	0.747	0.127	F_2 = 0.883 (df 4, 36)	0.963	1.015	0.188	0.333
3	3.00	0.784	0.740	0.739	0.871	0.673	0.034	LSD (external validation metrics) = 0.3044 (p = 0.05)	0.968	1.015	0.005	0.056
4	3.00	0.771	0.746	0.851	0.884	0.700	0.147	ranked means[c] for external validation metrics (highest to lowest): (CCC, Q^2_{F1}, Q^2_{F3}, Q^2_{F2}, $\overline{r^2_m}$)	1.000	0.990	0.153	0.231
5	3.00	0.762	0.747	0.884	0.888	0.707	0.177		0.999	0.994	0.199	0.294
6	1.00	0.663	0.660	0.897	0.865	0.616	0.212		1.009	0.985	0.366	0.565
7	1.00	0.439	0.436	0.872	0.771	0.478	0.220		0.985	1.008	0.297	0.412
8	3.00	0.568	0.548	0.913	0.812	0.552	0.237		1.001	0.994	0.281	0.430
9	2.00	0.157	0.015	0.847	0.700	0.336	0.376		1.011	0.982	0.636	1.004
10	2.00	0.826	0.772	0.969	0.900	0.719	0.164		1.001	0.997	0.228	0.355
						N_{test} = 200						
11	9.00	0.760	0.747	−0.351	0.837	0.503	0.246	F_1 = 2.061 (df 9, 36)	0.958	0.968	0.587	1.025
12	5.50	0.757	0.745	0.669	0.868	0.658	0.055	F_2 = 3.633 (df 4, 36)	0.990	0.990	0.061	0.091
13	2.50	0.742	0.740	0.856	0.887	0.694	0.179	LSD (external validation metrics) = 0.2234 (p = 0.05)	0.999	0.993	0.227	0.345
14	2.50	0.762	0.761	0.919	0.898	0.701	0.166	ranked means[c] for external validation metrics (highest to lowest): (CCC, Q^2_{F3}), (Q^2_{F3}, Q^2_{F1}), (Q^2_{F1}, $\overline{r^2_m}$, Q^2_{F3})	1.009	0.987	0.260	0.418
15	2.00	0.639	0.616	0.924	0.828	0.574	0.143		0.996	1.000	0.175	0.258
16	1.00	0.303	0.177	0.899	0.728	0.383	0.346		1.004	0.991	0.556	0.859
17	2.00	0.176	−0.177	0.910	0.647	0.283	0.385		1.006	0.989	0.612	0.951
18	1.00	0.386	−0.095	0.942	0.686	0.310	0.398		1.008	0.989	0.707	1.102
19	1.50	0.475	0.448	0.817	0.783	0.501	0.283		0.996	0.993	0.351	0.542
20	2.00	0.797	0.619	0.884	0.870	0.558	0.229		0.996	0.999	0.621	1.027
						N_{test} = 300						
21	7.33	0.760	0.644	0.012	0.752	0.371	0.346	F_1 = 1.778 (df 9, 36)	0.974	0.940	0.752	0.945
22	3.67	0.625	0.405	0.795	0.748	0.430	0.184	F_2 = 2.644 (df 4, 36)	1.000	0.985	0.229	0.322
23	1.00	0.778	0.777	0.752	0.901	0.739	0.156	LSD (external validation metrics) = 0.1960 (p = 0.05)	0.980	1.007	0.204	0.281
24	0.33	0.804	0.803	0.718	0.906	0.750	0.060	ranked means[c] for external validation metrics (highest to lowest): (CCC, Q^2_{F1}), (Q^2_{F1}, Q^2_{F2}, $\overline{r^2_m}$,Q^2_{F3})	0.982	1.002	0.087	0.104
25	9.33	0.780	0.761	−0.026	0.854	0.564	0.214		0.954	0.986	0.436	0.767
26	5.00	0.689	0.683	0.762	0.857	0.639	0.151		0.990	0.996	0.176	0.253
27	2.67	0.681	0.681	0.891	0.867	0.642	0.204		1.010	0.984	0.285	0.459
28	2.67	0.151	0.098	0.851	0.694	0.355	0.367		1.005	0.987	0.494	0.764
29	9.00	0.763	0.747	−0.062	0.844	0.545	0.228		0.956	0.984	0.458	0.800
30	4.00	0.716	0.713	0.805	0.870	0.666	0.148		0.995	0.994	0.165	0.247

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396−408

Table 3. continued

test set trials	% of test set compounds outside the applicability domain	external validation metrics					additional metric[a]		slope and intercept terms			
		Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	$\overline{r^2_{m\,(test)}}$	$\Delta r^{\,2}_{m\,(test)}$	ANOVA and multiple comparison[b]	k	k'	Δm	Δc
31	5.25	0.753	0.721	0.211	0.817	0.486	0.272	F_1 = 1.605 (*df* 9, 36)	0.975	0.973	0.582	0.864
32	3.25	0.709	0.709	0.878	0.872	0.678	0.192	F_2 = 5.177 (*df* 4, 36)	1.014	0.980	0.201	0.343
33	1.25	0.527	0.432	0.880	0.779	0.493	0.287	LSD (external validation metrics) = 0.1137 (p = 0.05)	0.997	0.997	0.366	0.565
34	4.75	0.729	0.729	0.296	0.822	0.489	0.275	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, Q^2_{F2}, $\overline{r^2_m}$)	0.977	0.987	0.596	0.994
35	4.00	0.811	0.792	0.784	0.892	0.714	0.062		0.995	0.991	0.070	0.097
36	4.75	0.724	0.708	0.829	0.879	0.654	0.192		1.000	0.991	0.310	0.499
37	3.75	0.530	0.385	0.819	0.784	0.465	0.301		0.991	1.000	0.514	0.804
38	2.75	0.644	0.562	0.852	0.833	0.554	0.250		0.995	0.997	0.424	0.669
39	2.25	0.663	0.612	0.868	0.842	0.596	0.234		0.995	0.998	0.320	0.497
40	4.25	0.677	0.677	0.451	0.807	0.532	0.260		0.977	0.994	0.341	0.570
						N_{test} = 500						
41	5.20	0.755	0.735	0.362	0.828	0.506	0.262	F_1 = 2.043 (*df* 9, 36)	0.981	0.979	0.566	0.848
42	2.00	0.654	0.652	0.868	0.858	0.616	0.218	F_2 = 20.768 (*df* 4, 36)	1.002	0.991	0.329	0.510
43	3.60	0.736	0.719	0.741	0.841	0.607	0.192	LSD (external validation metrics) = 0.0587 (p = 0.05)	0.997	0.990	0.229	0.357
44	4.40	0.799	0.798	0.571	0.876	0.611	0.199	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F1}, Q^2_{F2}), (Q^2_{F3},$\overline{r^2_m}$)	0.988	0.988	0.433	0.683
45	3.80	0.703	0.701	0.628	0.829	0.587	0.225		0.985	0.994	0.252	0.401
46	2.80	0.740	0.739	0.646	0.847	0.604	0.223		0.985	0.995	0.304	0.484
47	4.60	0.722	0.722	0.658	0.844	0.614	0.192		0.987	0.994	0.213	0.339
48	4.60	0.666	0.666	0.598	0.809	0.545	0.201		0.987	0.991	0.236	0.372
49	3.20	0.774	0.774	0.739	0.869	0.651	0.199		1.002	0.984	0.290	0.418
50	2.80	0.778	0.778	0.746	0.872	0.654	0.197		0.992	0.994	0.276	0.425
						mean (±s.e.) values at a particular test set size						
	N_{test} = 100	0.658 ± 0.069	0.618 ± 0.077	0.625 ± 0.228	0.838 ± 0.023	0.587 ± 0.049		F_1 = 1.719 (*df* 4, 16)				
	N_{test} = 200	0.580 ± 0.072	0.458 ± 0.114	0.747 ± 0.125	0.803 ± 0.028	0.517 ± 0.048		F_2 = 14.709 (*df* 4, 16)				
	N_{test} = 300	0.675 ± 0.061	0.631 ± 0.069	0.550 ± 0.127	0.829 ± 0.023	0.570 ± 0.045		LSD (external validation metrics) = 0.0776 (p = 0.05)				
	N_{test} = 400	0.677 ± 0.029	0.633 ± 0.043	0.687 ± 0.083	0.833 ± 0.012	0.566 ± 0.028		ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F1}, Q^2_{F3}, Q^2_{F2}), (Q^2_{F2}, $\overline{r^2_m}$)				
	N_{test} = 500	0.733 ± 0.015	0.728 ± 0.015	0.656 ± 0.043	0.847 ± 0.007	0.600 ± 0.014						

[a]Not included in ANOVA. [b]F_1 corresponds to the F value between rows (i.e., test set trials in the case of a data array corresponding to a particular test set size or test set size in the case of the data array corresponding to the mean values). F_2 corresponds to the F value between columns (i.e., external validation metrics). Critical F values (p = 0.05): 2.153 (*df* 9, 36); 2.634 (*df* 4, 36); 3.007 (*df* 4, 16). [c]Two means not included within the same parentheses are statistically significantly different at p = 0.05.

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396–408

Table 4. Results of External Validation for the Model Obtained from Training Set 3 (Random Division)

test set trials	% of test set compounds outside the applicability domain	external validation metrics					additional metric[a]		slope and intercept terms			
		Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	$\overline{r^2_{m\,(test)}}$	$\Delta r^2_{m\,(test)}$	ANOVA and multiple comparison[b]	k	k'	Δm	Δc
						N_{test} = 100						
1	8.00	0.802	0.798	0.575	0.863	0.530	0.237	F_1 = 7.425 (*df* 9, 36)	1.003	0.971	0.710	1.037
2	3.00	0.876	0.862	0.820	0.919	0.714	0.138	F_2 = 10.145 (*df* 4, 36)	0.998	0.989	0.336	0.465
3	7.00	0.807	0.801	0.667	0.876	0.634	0.189	LSD (external validation metrics) = 0.0654 (p = 0.05)	1.025	0.955	0.458	0.579
4	2.00	0.805	0.794	0.799	0.883	0.681	0.184	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F1}, Q^2_{F3}, Q^2_{F2}), $\overline{r^2_m}$	0.993	0.996	0.249	0.407
5	3.00	0.853	0.838	0.844	0.919	0.796	0.045		1.027	0.964	0.012	0.069
6	2.00	0.484	0.483	0.706	0.785	0.500	0.226		0.997	0.984	0.269	0.417
7	6.00	0.812	0.809	0.638	0.900	0.737	0.091		0.984	0.997	0.085	0.163
8	1.00	0.890	0.883	0.904	0.937	0.819	0.109		1.002	0.993	0.143	0.213
9	0.00	0.684	0.676	0.906	0.843	0.605	0.050		1.007	0.987	0.041	0.090
10	0.00	0.856	0.839	0.966	0.916	0.771	0.075		0.998	1.000	0.080	0.124
						N_{test} = 200						
11	2.00	0.863	0.862	0.881	0.919	0.718	0.140	F_1 = 11.346 (*df* 9, 36)	0.997	0.996	0.322	0.494
12	1.00	0.828	0.820	0.850	0.903	0.743	0.115	F_2 = 27.284 (*df* 4, 36)	1.004	0.986	0.140	0.177
13	3.00	0.768	0.752	0.837	0.873	0.667	0.033	LSD (external validation metrics) = 0.0418 (p = 0.05)	0.999	0.992	0.044	0.059
14	3.00	0.766	0.761	0.853	0.858	0.626	0.214	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$	1.008	0.984	0.332	0.474
15	9.50	0.825	0.825	0.673	0.905	0.737	0.149		0.974	1.008	0.167	0.326
16	0.00	0.864	0.858	0.844	0.915	0.697	0.146		1.003	0.987	0.377	0.524
17	0.00	0.551	0.550	0.766	0.779	0.477	0.032		1.003	0.983	0.024	0.063
18	3.00	0.828	0.828	0.809	0.891	0.628	0.184		1.010	0.979	0.507	0.733
19	0.50	0.890	0.890	0.911	0.939	0.792	0.106		0.990	1.004	0.199	0.323
20	1.00	0.762	0.762	0.852	0.879	0.688	0.014		1.013	0.979	0.046	0.021
						N_{test} = 300						
21	1.67	0.862	0.861	0.880	0.921	0.749	0.133	F_1 = 14.581 (*df* 9, 36)	1.002	0.990	0.255	0.367
22	2.67	0.779	0.777	0.832	0.884	0.695	0.055	F_2 = 89.038 (*df* 4, 36)	0.998	0.992	0.066	0.094
23	5.67	0.801	0.801	0.772	0.889	0.712	0.172	LSD (external validation metrics) = 0.0187 (p = 0.05)	0.999	0.987	0.204	0.300
24	2.67	0.851	0.849	0.808	0.913	0.717	0.145	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$	0.988	1.000	0.282	0.446
25	1.33	0.719	0.718	0.773	0.841	0.611	0.163		1.009	0.977	0.216	0.284
26	6.67	0.842	0.842	0.832	0.919	0.779	0.038		0.997	0.993	0.043	0.060
27	4.00	0.823	0.822	0.788	0.904	0.747	0.145		0.990	0.998	0.149	0.247
28	4.67	0.818	0.815	0.806	0.909	0.755	0.032		0.990	0.999	0.041	0.052
29	5.67	0.816	0.811	0.846	0.914	0.764	0.142		1.002	0.989	0.156	0.264
30	2.33	0.845	0.844	0.877	0.911	0.735	0.146		1.006	0.987	0.257	0.363
						N_{test} = 400						
31	1.75	0.839	0.840	0.922	0.889	0.750	0.143	F_1 = 5.309 (*df* 9, 36)	1.000	0.991	0.212	0.303
32	3.00	0.844	0.837	0.971	0.910	0.659	0.137	F_2 = 58.479 (*df* 4, 36)	1.003	0.988	0.169	0.240

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396–408

Table 4. continued

test set trials	% of test set compounds outside the applicability domain	external validation metrics					additional metric[a]		slope and intercept terms			
		Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	$\overline{r^2_{m\,(test)}}$	$\Delta r^2_{m\,(test)}$	ANOVA and multiple comparison[b]	k	k'	Δm	Δc
33	4.75	0.840	0.839	0.759	0.910	0.731	0.147	LSD (external validation metrics) = 0.0301 (p = 0.05)	0.987	0.999	0.223	0.362
34	1.75	0.726	0.725	0.795	0.844	0.616	0.187	ranked means[c] for external validation metrics (highest to lowest): (Q^2_{F3},CCC), (Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$	1.008	0.980	0.240	0.322
35	1.25	0.777	0.763	0.865	0.862	0.641	0.209		1.004	0.989	0.292	0.431
36	3.25	0.837	0.837	0.898	0.891	0.684	0.176		0.997	0.990	0.282	0.425
37	2.75	0.775	0.773	0.930	0.868	0.709	0.166		1.003	0.985	0.198	0.277
38	2.00	0.837	0.837	0.935	0.901	0.728	0.157		0.997	0.992	0.219	0.327
39	3.50	0.855	0.855	0.907	0.910	0.690	0.174		0.999	0.988	0.202	0.292
40	2.50	0.791	0.789	0.904	0.867	0.695	0.176		1.001	0.986	0.252	0.366
						N_{test} = 500						
41	2.00	0.797	0.797	0.800	0.888	0.709	0.163	F_1 = 6.493 (df 9, 36)	0.999	0.989	0.187	0.273
42	2.80	0.803	0.803	0.831	0.893	0.720	0.123	F_2 = 279.783 (df 4, 36)	1.003	0.986	0.149	0.203
43	3.60	0.817	0.817	0.812	0.894	0.688	0.171	LSD (external validation metrics) = 0.0109 (p = 0.05)	0.997	0.992	0.291	0.440
44	2.20	0.821	0.820	0.841	0.901	0.741	0.145	ranked means[c] for external validation metrics (highest to lowest): CCC, (Q^2_{F3}, Q^2_{F1}), (Q^2_{F1}, Q^2_{F2}), $\overline{r^2_m}$	1.002	0.988	0.169	0.232
45	3.20	0.784	0.783	0.767	0.878	0.682	0.188		0.995	0.991	0.227	0.348
46	3.20	0.818	0.817	0.806	0.894	0.690	0.170		1.003	0.986	0.296	0.432
47	1.00	0.808	0.808	0.853	0.898	0.729	0.103		0.997	0.994	0.115	0.171
48	2.40	0.800	0.800	0.829	0.899	0.730	0.014		0.990	1.000	0.007	0.025
49	3.60	0.826	0.826	0.818	0.905	0.747	0.153		0.995	0.994	0.171	0.264
50	2.40	0.805	0.805	0.829	0.899	0.729	0.064		0.994	0.996	0.068	0.106
						mean (±s.e.) values at a particular test set size						
	N_{test} = 100	0.787 ± 0.038	0.778 ± 0.037	0.783 ± 0.041	0.679 ± 0.014	0.884 ± 0.035		F_1 = 3.456 (df 4, 16)				
	N_{test} = 200	0.795 ± 0.030	0.791 ± 0.031	0.828 ± 0.021	0.677 ± 0.014	0.886 ± 0.028		F_2 = 74.947 (df 4, 16)				
	N_{test} = 300	0.816 ± 0.013	0.814 ± 0.013	0.821 ± 0.012	0.726 ± 0.008	0.901 ± 0.015		LSD (external validation metrics) = 0.0240 (p = 0.05)				
	N_{test} = 400	0.812 ± 0.013	0.810 ± 0.014	0.889 ± 0.021	0.690 ± 0.007	0.885 ± 0.013		ranked means[c] for external validation metrics (highest to lowest): CCC, Q^2_{F3}, (Q^2_{F1},Q^2_{F2}), $\overline{r^2_m}$				
	N_{test} = 500	0.808 ± 0.004	0.808 ± 0.004	0.819 ± 0.008	0.717 ± 0.002	0.895 ± 0.007						

[a]Not included in ANOVA. [b]F_1 corresponds to the F value between rows (i.e., test set trials in the case of a data array corresponding to a particular test set size or test set size in the case of the data array corresponding to the mean values). F_2 corresponds to the F value between columns (i.e., external validation metrics). Critical F values (p = 0.05): 2.153 (df 9, 36); 2.634 (df 4, 36); 3.007 (df 4, 16). [c]Two means not included within the same parentheses are statistically significantly different at p = 0.05.

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396–408

derived at different test set sizes, insignificant values of F (at $p = 0.05$) are obtained in some cases (such as for $N = 100$, both F_1 and F_2 are insignificant at $p = 0.05$) because of the drastic variations of the values of the external validation metrics both across rows and columns, leading to lower confidence in determining the role of the factors like different validation metrics and different trials to the total variance. If we focus on the results of ANOVA applied on the mean values of different metrics obtained from 10 trials at different test set sizes, then it is observed that there is a significant difference among the validation metrics, while the impact of the test set size is not significant at $p = 0.05$. Multiple comparison of the ranked mean values of the external validation metrics shows that $\overline{r_m^2}$ is the strictest metric, while CCC is the most optimistic metric and also that $\overline{r_m^2}$ is significantly different from others except Q^2_{F2}.

Results Obtained for Training Set 3 (Random Division). For the training set obtained from the random division of the data set, 10 test sets each for the test set sizes of 100, 200, 300, 400, and 500 were generated, and the quality of the model generated from the training set was evaluated for predictive quality from these test sets. The results are shown in Table 4. Like the previous two cases, here also a single metric shows values ranging from low to high in different trials at a given test set size. For example, Q^2_{F1} shows a value of 0.890 at trial 8, while the corresponding value is 0.484 at trial 6. Again, at trial 6 the value of Q^2_{F3} is as high as 0.706. ANOVA of the data matrices at different test set sizes confirm significant differences among the external validation metrics and also significant contributions of trials (i.e., test set composition) in determining the quality of external validation. ANOVA applied on the mean values of the external validation metrics at a particular test set size confirms the impact of test size in determining the quality of external predictivity. Multiple comparison performed in all cases for this training set shows that $\overline{r_m^2}$ is significantly different from other external validation metrics, among which CCC is the most optimistic.

Additional Observations. Additionally, we observed some additional metrics like k, k', Δm, and Δc for the results obtained from each test set, and these are tabulated in Tables 2, 3, and 4. The metrics k and k' are included in the criteria of Golbraikh and Tropsha for external validation.[12] When observed responses are plotted in the y-axis, predicted responses are plotted in the x-axis, and the best fit regression line is drawn setting the intercept to zero, the slope of the regression line is termed as k. The metric k' can similarly be obtained by interchanging the axes. According to the criteria recommended by Golbraikh and Tropsha,[12] the values of k or k' should be close to 1 (ranging 0.85 to 1.15). Though, we have omitted the criteria recommended by Golbraikh and Tropsha for the present comparison analysis, it is interesting to note from the observations of Tables 2−4 that in all the cases the values of k and k' are near 1 (within 0.85−1.15) irrespective of the quality of external predictivity. Thus, it appears that the slope criterion (k or k') is not very suitable in determining the quality of the external prediction. The metric Δm represents the absolute difference of slopes of the best fit regression lines correlating observed and predicted responses (x-axis and y-axis, respectively, and also interchanging the axes). Similarly, Δc indicates the absolute difference of intercepts of the best fit regression lines correlating observed and predicted responses (x-axis and y-axis, respectively, and also interchanging the axes). It is interesting to note that for all three training sets, the metric Δr_m^2 bears high intercorrelation with each of Δm and Δc (Table 5). Thus, the metric Δr_m^2 signifies the impact of change of axes

Table 5. Intercorrelation (r) of Δr_m^2 with Δm and Δc

model no.	$r(\Delta r_m^2, \Delta m)$	$r(\Delta r_m^2, \Delta c)$
1	0.901	0.906
2	0.875	0.851
3	0.803	0.812

on the correlation between observed and predicted responses. In the cases of an ideal correlation (when all of the observed responses are exactly same as the corresponding predicted responses), the values of Δr_m^2, Δm, and Δc will be zero. The more the deviation of the predicted responses differ from the observed ones, the more the values of Δr_m^2, Δm, and Δc will deviate from zero.

Comment on a Previous Report of Relevance. Recently, Chirico and Gramatica[20] have compared various external validation metrics in a report where they have considered different thresholds for different metrics for comparison (0.6 for Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}, 0.5 for r_m^2, and 0.85 for CCC). This is completely an injustice to compare different metrics with different threshold values. One may consider a higher threshold for a particular metric to show it the most precautionary. We have done a comparison among different metrics with the same threshold values but repeated the calculations by varying the thresholds (0.5, 0.6, 0.7, 0.8, and 0.85). The results are shown in Figures 1, 2, 3, and 4, which suggest that $\overline{r_m^2}$ is the strictest metric at a given threshold level. If both $\overline{r_m^2}$ and Δr_m^2 are considered, the number of trials with an external validation metric value within the desired range attains a minimum value. On the other hand, CCC is an over-optimistic metric showing encouraging values even when most of the other metrics show poor values (for example, trials 2, 28, 33, etc. in Table 2, trials 9,17, 19, 28, etc. in Table 3, and trials 6, 17, etc. in Table 4). Considering the mean values of fractions of trials with an acceptable value of an external validation metric at different thresholds, CCC is found to be least precautionary and $\overline{r_m^2}$ (along with Δr_m^2 lower than 0.2) is confirmed to be the most stringent metric. It will not be out of scope to mention here that Chirico and Gramatica have mentioned in their paper[20] that calculation of r_m^2 does not consider slopes, whereas in the present paper, we have shown a direct relationship between Δr_m^2 and Δm values and also the redundancy of k and k' values. Further, Chirico and Gramatica have wrongly mentioned that the value of r_0^2 may be good even when the data points in the experimental/predicted graph do not match. When the data points on the ordinate values are 10 times the ones on the abscissa, r_0^2 can never be 1, as claimed by these authors.[20] The basic concept of r_m^2 originated from when predicted values deviate much from observed values, the value of r_0^2 will be inferior to r^2. The present work has clearly demonstrated that $\overline{r_m^2}$ along with Δr_m^2 provides a very stringent criterion of external validation.

TEST FOR APPLICABILITY DOMAIN OF THE MODELS

The applicability domain of a model refers to the chemical structure space in which the model makes predictions with a given reliability. We have checked applicability domains of the developed models (Table S2 in the Supporting Information)

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396−408

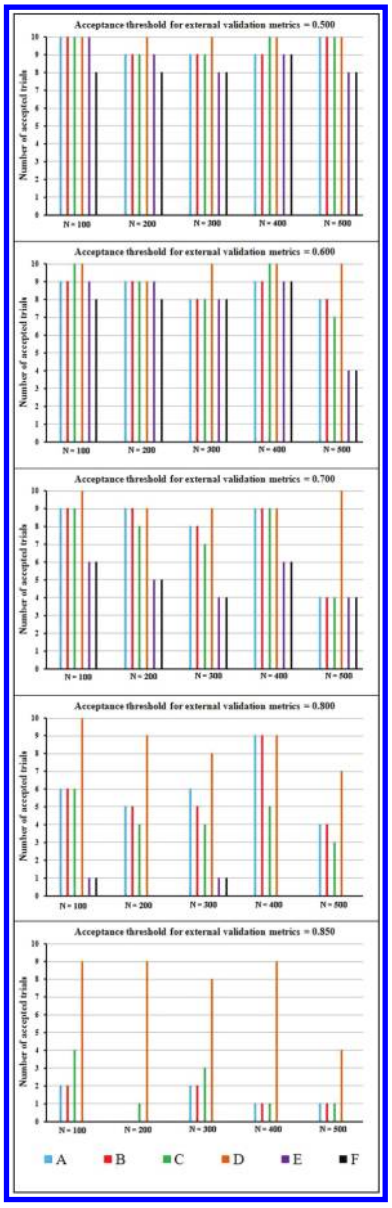


Figure 1. Comparative representation of numbers of test sets with accepted values of external validation metrics at different thresholds for model 1: (A) Q^2_{F1}, (B) Q^2_{F2}, (C) Q^2_{F3}, (D) CCC, (E) $\overline{r^2_m}$, and (F) $\overline{r^2_m}$ + Δr^2_m(<0.2).

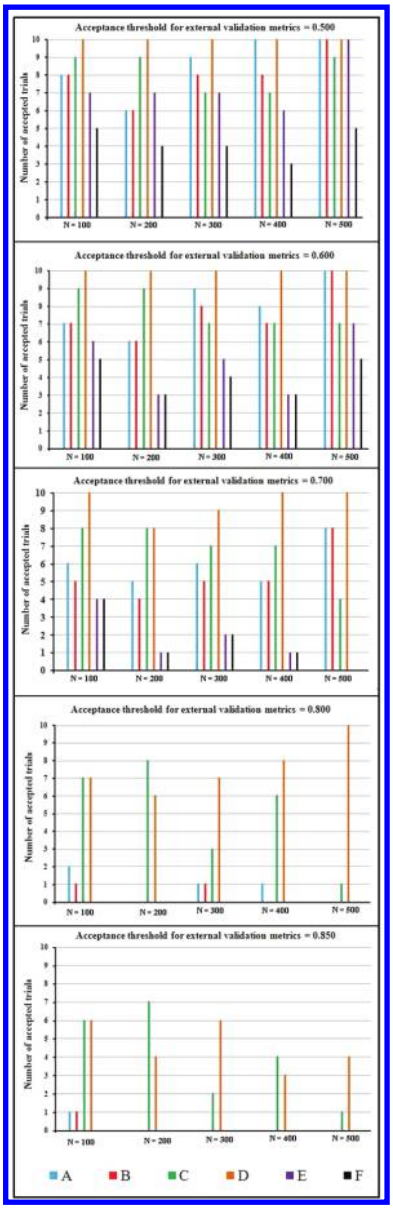


Figure 2. Comparative representation of numbers of test sets with accepted values of external validation metrics at different thresholds for model 2: (A) Q^2_{F1}, (B) Q^2_{F2}, (C) Q^2_{F3}, (D) CCC, (E) $\overline{r^2_m}$, and (F) $\overline{r^2_m}$ + Δr^2_m(<0.2).

using the leverage approach[32] and tested the fractions of test set compounds in different trials falling outside the applicability domains of the corresponding models (results shown in Tables 2−4). It is clear from the analysis that there is no significant influence of applicability domain in determining the quality of predictions, at least in this study. For example, Table 2 shows that though only 3% of the test set compounds (test set size = 100) are outside the applicability domain of model 1 in trial 2, the quality of external prediction is poor. Again, trial 5 for the same model (test set size = 100) shows a good quality of external predictions in spite of a higher fraction (9%) of test set compounds remaining outside the applicability domain of the model. This is in compliance with the observations made by Huang and Fan[33] that consideration of applicability domain alone is not sufficient for assessing a model's predictability on an external set.

OVERVIEW AND CONCLUSIONS

The present work has clearly shown that a single model may have a variable quality of external predictivity depending on the composition and size of test sets. Thus, it will be completely unethical to judge predictive quality of a model on the basis of the predictions found from a given test set as is usually practiced in the QSPR literature. In our opinion, more attention may be paid to the equation quality metrics and internal validation parameters in determining the quality of a model rather than making a conclusion exclusively based on predictions for a single test set, especially in the case of a test set of small size. If test set validation is done, multiple test sets of varying composition and size should be tried before making a decision on the predictive quality of the developed model. However, such approaches have been rare in the QSPR literature, which has given more importance to external

dx.doi.org/10.1021/ci200520g | J. Chem. Inf. Model. 2012, 52, 396−408

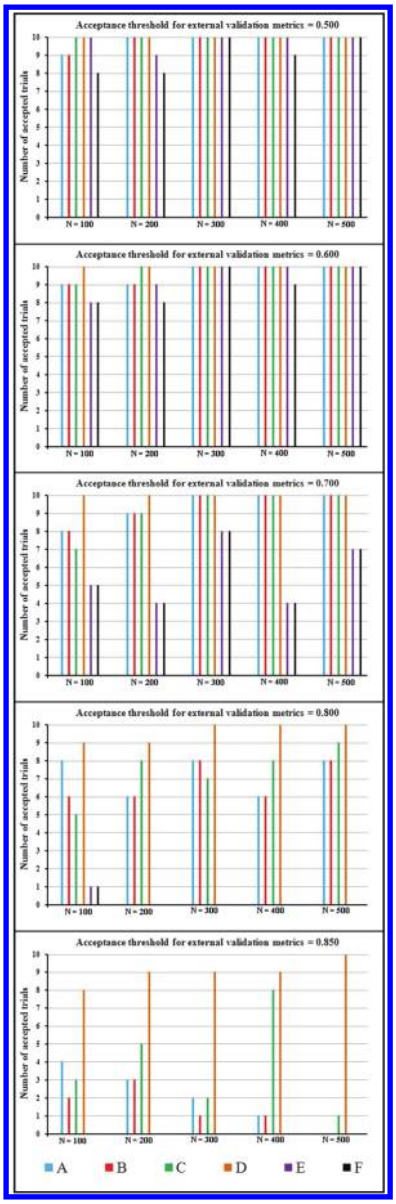


Figure 3. Comparative representation of numbers of test sets with accepted values of external validation metrics at different thresholds for model 3: (A) Q^2_{F1}, (B) Q^2_{F2}, (C) Q^2_{F3}, (D) CCC, (E) $\overline{r^2_m}$, and (F) $\overline{r^2_m}$ + Δr^2_m(<0.2).

validation in determining the quality of a model using a single test set. The present work has also shown that the metric $\overline{r^2_m}$ is statistically significantly different from other external validation metrics in most cases. On the basis of the numerical values of external validation metrics, CCC appears to be an over-optimistic metric, while $\overline{r^2_m}$ is the most conservative. It has also been found that $\overline{r^2_m}$ along with Δr^2_m (lower than 0.2) provides the most stringent criterion of external validation, and as for a minimum number of trials, these criteria are met at a given threshold value. Finally, it may be concluded that for regulatory decision support processes, external validation using multiple test sets is more desired, and $\overline{r^2_m}$ along with Δr^2_m may be used as the strictest criterion for accepting any model in terms of external predictivity. It will thus be a good idea to have the option of computing $\overline{r^2_m}$ along with Δr^2_m in various software

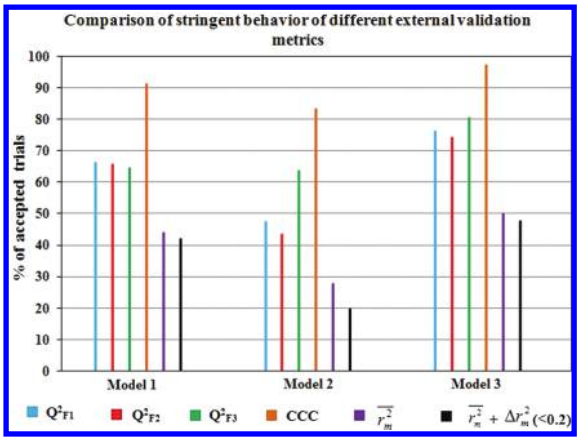


Figure 4. Representation of stringent behavior of different external validation metrics Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, CCC, and $\overline{r^2_m}$, $\overline{r^2_m}$ + Δr^2_m(<0.2): Fractions of total numbers of test sets with accepted values of external validation metrics (considering mean values obtained at different thresholds).

packages and Web servers used for developing QSAR/QSPR models.

■ ASSOCIATED CONTENT

Supporting Information

Details of the metrics used for validation and data set and computed descriptors. This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kunalroy_in@yahoo.com. URL: http://sites.google.com/site/kunalroyindia/. Phone: +91 98315 94140. Fax: +91 33 2837 1078.

■ ACKNOWLEDGMENTS

Financial assistance from the UGC, New Delhi and ICMR, New Delhi is thankfully acknowledged. We thank Dr. Supratim Ray for his help in computation of descriptors used in this work.

■ REFERENCES

(1) Gonzalez, M. P.; Teran, C.; Saiz-Urra, L.; Teijeira, M. Variable selection methods in QSAR: An overview. *Curr. Top. Med. Chem.* **2008**, *8*, 1606−1627.

(2) Helguera, A. M.; Combes, R. D.; Gonzalez, M. P.; Cordeiro, M. N. Applications of 2D descriptors in drug design: A DRAGON tale. *Curr. Top. Med. Chem.* **2008**, *8*, 1628−1655.

(3) Worth, A. P.; Bassan, A.; De Bruijn, J.; Saliner, A. G.; Netzeva, T.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Eisenreich, S. The role of the European chemicals bureau in promoting the regulatory use of (Q)SAR methods. *SAR QSAR Environ. Res.* **2007**, *18*, 111−125.

(4) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(5) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure−activity or structure−property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20*, 241−266.

(6) Roy, K. On some aspects of validation of predictive quantitative structure−activity relationship models. *Expert Opin. Drug Discov.* **2007**, *2*, 1567−1577.

(7) Wold, S.; Eriksson, L. In *Chemometrics Methods in Molecular Design*; Waterbeemd, H. V. D., Ed.; VCH: Weinheim, 1995; pp 309−318.
(8) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(9) Debnath, A. K. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001.
(10) Chou, K. C.; Shen, H. B. Cell-PLoc: A package of web servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols* **2008**, *3*, 153−162.
(11) Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236−247.
(12) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(13) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Diversity* **2002**, *5*, 231−243.
(14) Chou, K. C.; Shen, H. B. Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1−16.
(15) Du, Q. S; Mezey, P. G.; Chou, K. C. Heuristic molecular lipophilicity potential (HMLP): A 2D-QSAR study to LADH of molecular family pyrazole and derivatives. *J. Comput. Chem.* **2005**, *26*, 461−470.
(16) Du, Q. S.; Huang, R. B.; Wei, Y. T.; Du, L. Q.; Chou, K. C. Multiple field three dimensional quantitative structure−activity relationship (MF-3D-QSAR). *J. Comput. Chem.* **2008**, *29*, 211−219.
(17) Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194−201.
(18) Schuurmann, G.; Ebert, R. U.; Chen, J.; Wang, B.; Kuhne, R. External validation and prediction employing the predictive squared correlation coefficient: Test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140−2145.
(19) Lin, L. I. Assay validation using the concordance correlation coefficient. *Biometrics* **1992**, *48*, 599−604.
(20) Chirico, N.; Gramatica, P. 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.* **2011**, *51*, 2320−2335.
(21) Ojha, P. K.; Mitra, I.; Das, R. N.; Roy, K. Further exploring r^2_m metrics for validation of QSPR models dataset. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 194−205.
(22) Lei, B.; Ma, Y.; Li, J.; Liu, H.; Yao, X.; Gramatica, P. Prediction of the adsorption capability onto activated carbon of a large data set of chemicals by local lazy regression method. *Atmos. Environ.* **2010**, *44*, 2954−2960.
(23) Todeschini, R.; Consonni, V.; Mannhold, R. *Molecular Descriptors for Chemoinformatics*. Wiley: Weinheim, 2009.
(24) *Cerius2*, version 4.10; Accelrys, Inc.: San Diego, CA, 2005, http://www.accelrys.com/cerius2.
(25) *DRAGON*, version 6.0; Talete srl: Milano, Italy, 2011, http://www.talete.mi.it/products/dragon_molecular_descriptors.htm.
(26) Everitt, B. S.; Landau, S.; Leese, M. *Cluster Analysis*; Edward Arnold: London, 2001.
(27) *SPSS*; IBM: New York, 2009, http://www-01.ibm.com/software/analytics/spss/.
(28) Darlington, R. B. *Regression and Linear Models*; McGraw-Hill: New York, 1990.
(29) Snedecor, G. W., Cochran, W. G. *Statistical Methods*; Oxford & IBH: New Delhi, 1967.
(30) Wold, S. Validation of QSAR's. *Quant. Struct.−Act. Relat.* **1991**, *10*, 191−193.
(31) Bolton, S. In *Remington: The Science and Practice of Pharmacy*; Troy, D. B., Ed.; Lippincott Williams & Wilkins: New York, 2006; Chapter 12, pp 127−161.
(32) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−1375.
(33) Huang, J.; Fan, X. Why QSAR fils: An empirical evaluation using conventional computational approach. *Mol. Pharm.* **2011**, *8*, 600−608.