Beware of external validation! – A Comparative Study of Several Validation Techniques used in QSAR Modelling

Subhabrata Majumdar [1], Subhash C. Basak [2,*]

[1] University of Florida Informatics Institute, 432 Newell Drive, CISE Bldg E251, Gainesville, FL 32611, USA; smajumdar@ufl.edu

[2] Natural Resources Research Institute, University of Minnesota Duluth, and Departement of Chemistry and Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA; sbasak@nrri.umn.edu

***** Correspondence: sbasak@nrri.umn.edu; Tel.: +1-218-727-1335; Fax: +1-218-720-4328

Abstract:

Background: Proper validation is an important aspect of QSAR modelling. External validation is one of the widely used validation methods in QSAR where the model is built on a subset of the data and validated on the rest of the samples. However, its effectiveness for datasets with a small number of samples but large number of predictors remains suspect.

Objective: Calculating hundreds or thousands of molecular descriptors using currently available software has become the norm in QSAR research, owing to computational advances in the past few decades. Thus, for *n* chemical compounds and *p* descriptors calculated for each molecule, the typical chemometric dataset today has high value of *p* but small *n* (i.e. $n << p$). Motivated by the evidence of inadequacies of external validation in estimating the true predictive capability of a statistical model in recent literature, this paper performs an extensive and comparative study of this method with several other validation techniques.

Methodology: We compared four validation methods: leave-one-out, *K*-fold, external and multi-split validation, using statistical models built using the LASSO regression, which simultaneously performs variable selection and modelling. We used 300 simulated datasets and one real dataset of 95 congeneric amine mutagens for this evaluation.

Results: External validation metrics have high variation among different random splits of the data, hence are not recommended for predictive QSAR models. LOO has the overall best performance among all validation methods applied in our scenario.

Conclusion: Results from external validation are too unstable for the datasets we analyzed. Based on our findings, we recommend using the LOO procedure for validating QSAR predictive models built on high-dimensional small-sample data.

Keywords: Cross validation; leave one out (LOO) cross validation; *K*-fold cross validation; external validation; LASSO; chemical mutagens; aromatic and heteroaromatic amines

1. Introduction

With its humble beginning in the second half of the nineteenth century [1, 2] the field of quantitative structure-activity relationship (QSAR) has come a long way to its current state. QSARs are mathematical models which attempts to predict property/ biomedicinal activity/ toxicity of chemicals from their properties or calculated molecular descriptors. The three major pillars of QSAR are: a) Adequately large and good quality data on the dependent variable, i.e., physical property or bioassay data, b) Relevant descriptors (independent variables) that are capable of quantifying aspects of molecular structure related to the property/ bioactivity of interest, and c) Proper Statistical methods for model building. For a recent review of the topic, please see [3].

In the second half of the twentieth century, the linear free energy related (LFER) approach, also known as Hansch analysis, was introduced to the field of QSAR [4]. This approach uses various combinations of hydrophobicity (experimental or calculated), Hammett's electronic parameter (σ) and numerous steric descriptors as independent variables for correlation. Such property-property relationships (PPRs) or property-activity relationships (PARs) worked for the assessment of bioactivities of molecules belonging to congeneric sets. But in many cases, experimental physicochemical properties of many of the chemicals under investigation are not available [3, 5]. The PPR/ PAR approaches are not very useful in such situations. A practical approach that has gradually emerged in such data-poor situations is the use of properties that can be calculated directly from molecular structure without the use of any other experimental data. Topological, substructural, geometrical (3-D), and quantum chemical molecular descriptors belong to this group. For large sets of molecules, high level quantum chemical descriptors could be very demanding on computer resources. On the other hand, descriptors derived from topological aspects of chemical structures, e.g.; topological indices [6-8] and different types of substructures [9], have found wide applications in numerous QSAR studies. For a recent summary of these topics, please see the review by Basak [3, 10].

During the past half century or so there has been an important change in the landscape of available molecular descriptors (independent variable) for QSAR. Whereas in the 1950s a few QSAR descriptors, both experimental and calculated, were available, currently available software can calculate a large number of descriptors [11-16]. This makes the QSAR modeling situation rank deficient where the number of predictors (p) is much larger as compared to the number of data points to be modeled or dependent variables (n). Such a situation for the judicious and correct use of statistical methods for model building and validation [1, 2].

According to the OECD principles, one of the necessary criteria a QSAR model fit to be implemented in practice must satisfy is proper model evaluation [3]. In the last two decades or so, QSAR researchers have adapted to using either one of the three validation methods:

(a) Leave-one-out (LOO) cross validation: For each compound in the full dataset, its activity is predicted by a model built on samples excluding that compound.
(b) K-fold cross validation: The data is randomly split into K disjoint partitions. Each partition is taken as test set, and QSAR models built on samples outside that partition to predict activities of samples in the partition;
(c) External validation: The data is randomly split into two partitions only once: a larger training set and a smaller test set. QSAR model is built on the training set and evaluated on the test set.

Golbraikh and Tropsha [4] argued using empirical evidence that in some cases LOO cross-validation overestimates the predictive ability of a model but external validation does not. Indeed, a diverse body of literature exists on QSAR models evaluated using external validation, references to which are available in [5, 6]. On the other hand, Hawkins *et al* [7] showed through theoretical argument and empirical study that for small sample sizes, the cross-validated q^2 obtained from a LOO procedure is a better estimator of the true R^2 (i.e. proportion of variance in the response variable explained by the predictors) than an externally validated q^2.

As mentioned earlier, the typical QSAR dataset is High-Dimensional Low Sample Size (HDLSS). Although external validation is one of *the* widely used validation methods in the QSAR community, evidence has been mounting towards its inadequacy in prediction problems for HDLSS data. Furthermore, there is the added issue of nested cross-validation. Statistical procedures on HDLSS data involve a dimension reduction step (Principal Component Regression (PCR), Partial Least Squares (PLS)), variable selection step (forward selection in regression models) and/or tuning parameter selection (Least Absolute Shrinkage and Selection Operator (LASSO) regression [8] or machine learning methods). To ensure that holdout compounds do not influence the training step while doing cross-validation, these steps should be repeated each time a model is trained. This two-step procedure is called two-deep cross-validation [9] or double cross validation [10, 11, 12].

In general, recent methodological research on the efficacy of validation techniques suggests a repeated two-deep validation procedure, that either covers the data through disjoint partitions (i.e. *K*-fold), or averages results over multiple random splits of the data (from hereon referred to as *multi-split validation*), over single-split external validation. When total number of samples is *n*, and size of the training and test sets are n_1 and n_2, respectively (with $n_1 + n_2 = n$), Yang [13] shows that the multi-split validation leads to almost sure recovery of the true underlying statistical model when $n_1 \to \infty$ and $n_2/n_1 \to 0$ as $n \to \infty$. In their simulation setup, the multi-split method outperforms single-split external validation for small sample sizes ($n = 100, 200$). Both methods perform similarly in large sample simulations ($n = 400, 800$ and 1000). For prediction, [11] showed through analyzing simulated data as well as two chemical activities datasets that the multi-split validation provides a more unbiased estimate of prediction errors than external validation.

In probably what is the one of the most relevant work considering our focus on HDLSS data, [14] showed through simulation that when sparse regression methods like LASSO, Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) are used (we shall discuss LASSO, please refer to [15] for details on SCAD and MCP) for model building, and estimating predictive performance is the goal, LOO cross-validation tends to outperform either *K*-fold CV or multi-split validation.

Motivated by the general tone in the body of work mentioned above, which highlights inadequacies of external validation in QSAR model evaluation, in this paper we perform a comprehensive analysis of all the validation methods, i.e. LOO, *K*-fold, external and multi-split. Through empirical results we argue that external validation outputs from a *single*, *random* split of a small set of chemicals tend to be extremely unstable, owing to their dependency on the small, randomly chosen testing set. Based on the experimental results we recommend LOO as the preferable method of cross validation in our specific scenario.

Note that we do not suggest the LOO cross-validation as a panacea for all validation problems. The propensity of LOO for over-fitting is well-known [16, 13], it becomes too computationally demanding for large sample sizes, and in presence of additional information like the order of sample collection, other

validation methods are preferable [17]. Nevertheless, we argue that LOO, and other validation methods like *K*-fold or repeated split, are more plausible options in terms of stability of outputs *when the dataset being modeled is small* and consists of random samples, as compared to the single-split external validation method that is widely used for QSAR modelling.

The rest of the paper is organized as follows. Section 2 contains more details on all these datasets. We build QSAR models on these data using the LASSO regression method [8], which we elaborate on in Section 3. This section also contains further details about our validation methods, as well as the two-step validation scheme. We present and discuss the results obtained from this analysis in Section 4. The paper concludes with section 5, where we highlight the takeaways from the paper, and motivate future directions of research.

2. Data

In total, we use 300 simulated datasets and a well-known chemical activities dataset for the study.

2.1 Simulated data

For sample size *n* and number of descriptors *p*, we generate data from the multivariate linear model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (1)$$

With $\boldsymbol{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times p}, \boldsymbol{\beta} \in \mathbb{R}^p$, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ being the random error with $\epsilon_i \sim N(0, \sigma^2)$ for i = 1, 2, …, n and $\sigma > 0$. We fix n = 100, and consider three different values of p: 100, 500 and 1000. For a fixed p, we first generate rows of the matrix of descriptors $\mathbf{X}$ as independent and identical draws from a p-dimensional normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. We fix the entries of $\boldsymbol{\Sigma}$ as

$$\sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0.9^{|i-j|} & \text{if } i \neq j \end{cases}$$

There is often high correlation among chemical descriptors, and when modelling data on hundreds of such descriptors the intrinsic dimensionality of the descriptor data is often much lower than the actual dimension of the predictor space [18] [19]. We use the above correlation structure to simulate this scenario. For the coefficient vector $\boldsymbol{\beta}$, we set its first 10 entries as 1 and rest p – 10 entries as 0. Finally, we generate elements of $\boldsymbol{\epsilon}$ by setting $\sigma = 1$, calculate the response variable $\boldsymbol{y}$ from (1), generate 100 such independent datasets for each value of p, and repeat the process for different values of p.

2.2 Congeneric data of 95 amines

This dataset is due to Debnath *et al* [20]. It contains information on a congeneric set of 95 amine compounds: specifically values on 275 descriptors calculated for each compound, and their mutagenic activities on the *Salmonella typhimurium* strain TA98: as measured by the number of revertants per nmol (in log scale) when a sample compound is applied to a test culture.

<Insert table 1>

Table 1: Information on descriptor types in the congeneric amines data

Type	No. of descriptors	Description	Software used

TS	108	Sees the molecule as a graph with unweighted edges, and quantifies connectivity and adjacency of the vertices (i.e. atoms) in the graph	POLLY v2.3 [21], MolconnZ v4.05 [22]
TC	158	Encodes connectivity within the weighted molecular graph, with edges weighted according to specific physical or chemical properties	POLLY v2.3 [21], MolconnZ v4.05 [22], TRIPLET [23]
3D	3	Encodes shape-related properties of the full molecule	Sybyl v6.2 [24]
QC	6	Quantifies high-level electro-chemical properties of the molecule, e.g. dipole moment	MOPAC v6.00 [25]

This descriptor dataset contains four types of descriptors: topostructural (TS), topochemical (TC), three dimensional (3D) and quantum chemical (QC), in increasing order of computational complexity. **Table 1** presents detailed information about these different types of descriptors. Please refer to Table S1 in supplementary material for a full list of descriptors. There is evidence from earlier QSAR studies [26, 27, 28] that while predicting chemical activity through QSAR modelling, the computation-intensive 3D and QC descriptors are largely redundant in presence of a large number of TS and TC descriptors that are computationally easy to calculate. However, we analyze all four types of descriptors in this paper for the sake of completeness, and because the statistical model used explicitly involves variable selection to automatically filter out variables that are not predictive enough.

3. Statistical methods

3.1 LASSO regression

For the linear model in (1), the LASSO method proposed by Tibshirani [8] obtains an estimate of $\boldsymbol{\beta}$ by solving the following minimization problem:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \qquad (2)$$

Where λ is a tuning parameter. This calculates a constrained estimate of $\boldsymbol{\beta}$ by adding an additional penalty term $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j|$ to the squared error loss function. Notice that using $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \boldsymbol{\beta}_j{}^2$ in (2) instead yields ridge regression, which has been extensively used to build predictive models in QSAR [26, 9, 29].

The advantage of using LASSO is two-fold:

(a) Since the penalty term is non-differentiable at 0, the solution $\widehat{\boldsymbol{\beta}}$ is sparse, i.e. some of its entries are exactly set to zero. Thus, LASSO performs simultaneous variable selection and estimation of predictor effects;

(b) Unlike linear regression which gives a unique solution only when $n > p$, existence and computation of the LASSO solution does not depend on the relative size of n and p. Thus it is able to tackle high-dimensional regression problems with a large number of predictors but limited sample size (i.e. $n \ll p$).

The large number of descriptors and low intrinsic dimensionality of datasets that are typical of many modern QSAR datasets makes dimension reduction or variable selection an essential component of QSAR modelling. Since LASSO provides sparse solutions of the coefficient vector in (2), it provides a frugal yet statistically rigorous way of modelling QSAR data that is high-dimensional in nature. Some previous studies have effectively applied LASSO and its variants to build QSAR models [30, 31, 32].

3.2 Cross-validation techniques

We use the following cross-validation techniques to evaluate the predictive capabilities of LASSO models built on the simulated as well as congeneric amine dataset.

***K*-fold cross validation (*K*-fold cv):** We divided the samples randomly into *k* splits, take samples in a split as test set, train a QSAR model on samples outside the test set and predict activity of samples in the test set with that model. Finally, we repeat this for all splits to cover all samples.

Leave-one-out cross validation (LOO-cv): For a sample of size *n*, we train *n* models, each time taking a distinct sample in the test set to predict the activity of that sample. This can be interpreted as a *n*-fold cross validation.

External validation: We randomly chose 10 samples to be included in the test set. We train the model using other samples and predict the responses in the test samples using that model.

Multi-split validation: We repeat the external validation method 100 times over different random train-test splits of the data, and takes the average of any metrics obtained over all such splits. This has been introduced in the QSAR literature as Monte-Carlo Cross Validation (MCCV). It ensures asymptotically consistent recovery of underlying components in a model (e.g. number of predictors or principal components) while avoiding overfitted models [16].

We use Mean Squared Prediction Error (MSPE) and cross-validated q^2 as model evaluation metrics. Table **2** gives the exact formulae that we use to calculate these metrics for each validation method.

<Insert table 2 here>

Table 2: MSPE and q^2 formulae for all validation methods

Method	MSPE formula	q^2 formula	Notation
LOO	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2$	$\frac{n \times \text{MSPE}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	$\hat{y}_{i,-i}$ = prediction for i^{th} sample from model that excludes that sample. $\bar{y} = \Sigma_{i=1}^{n} y_i/n$
K-fold	$\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_i - \hat{y}_{i,-k})^2$	$\frac{n \times \text{MSPE}}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_i - \bar{y}_k)^2}$	n_k = no. of samples in k^{th} fold $\hat{y}_{i,-k}$ = prediction for i^{th} sample in k^{th} fold from model that excludes that fold. $\bar{y}_k = \Sigma_{i=1}^{n_k} y_i/n_k$
External	$\frac{1}{n_t}\sum_{i=1}^{n_t}(y_{ti} - \hat{y}_{ti,-t})^2$	$\frac{n_t \times \text{MSPE}}{\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_t)^2}$	n_t = no. of test samples y_{ti} = i^{th} test sample

			$\hat{y}_{ti,-t}$ = prediction for i^{th} sample in test set from model excluding test set. $\bar{y}_t = \Sigma_{i=1}^{n_t} y_{ti}/n_t$
Multi-split	$\frac{1}{S}\sum_{s=1}^{S} \text{MSPE}_s$	$\frac{1}{S}\sum_{s=1}^{S} \frac{n_t \times \text{MSPE}_s}{\sum_{i=1}^{n_t}(y_{sti} - \bar{y}_{st})^2}$	S = no. of random train-test splits y_{sti} = i^{th} test sample in s^{th} fold MSPE_s = MSPE from s^{th} split $\bar{y}_{st} = \Sigma_{i=1}^{n_t} y_{sti}/n_t$

<Insert figure 1 here>

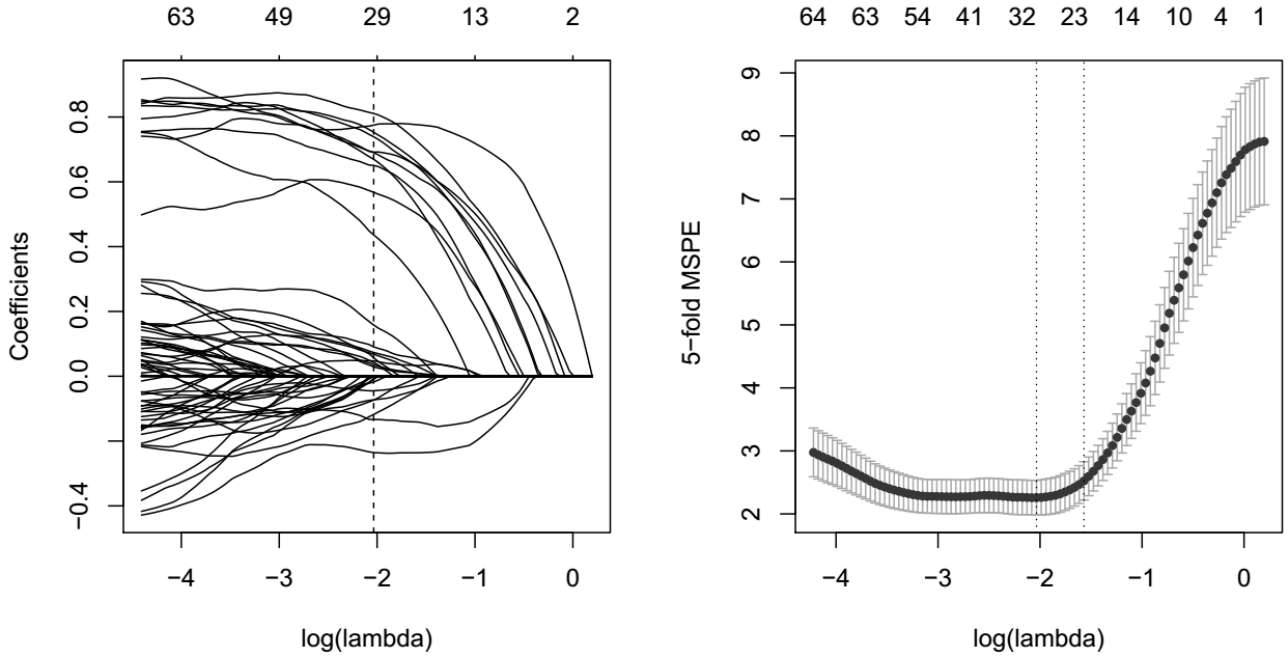

Figure 1: For a given value of λ, LASSO calculates coefficient vector with some values set to 0. The left panel plots these coefficient values for a range of λ values (in log scale) as calculated from the training dataset. Now 5 -fold cross-validation is done on this training dataset, and the right panel plots 5-fold cross-validation MSPE for these λ (with ± 1 standard deviation bounds). We choose the coefficient vector to the λ that minimizes this MSPE (dotted line close to 2 in both panels) as our final solution for $\boldsymbol{\beta}$.

The tuning parameter λ for the LASSO regression model in (2) is selected from a range of values using *K*-fold cross-validation. Here we shall take $k = 5$. Figure **1** illustrates this tuning parameter selection for a simulated dataset with $p = 100$. While implementing each of the validation methods mentioned above we need to make sure to incorporate this step every time a model is trained. In this situation, selecting the tuning parameters first on a model built on the full dataset and then predicting for different train-test splits might seem a more intuitive approach. However, this naïve approach uses information from the holdout compounds in the first step, thus providing an inflated estimate of the cross-validated q^2: which is termed as naïve q^2 [9].

Thus, we perform cross-validation twice: once to select the best tuning parameter from the training samples, and again to obtain q^2 values. As an example, for *K*-fold cross-validation the steps for this *two-deep cross validation* procedure will be as follows:

a) Randomly split data into *k* groups.
b) Consider samples in the first split as test set. Select the best tuning parameter by doing a 5-fold CV using the LASSO model in (2) on samples outside the test set.
c) Predict activities of compounds in test set using a LASSO model trained using the best tuning parameter.
d) Repeat steps (b) and (c) considering all other splits as test sets.
e) We now have predictions for all sample compounds. Calculate MSPE and q^2 values using these predicted values.

4. Results

We highlight the effect of increasing dimension of the predictor space on performances of all four validation methods through an extensive simulation study and a real data example. For the simulations we use three sets of datasets, corresponding to predictor dimensions p = 100. 500 and 1000. For each predictor dimension we generate 100 independent datasets, each with fixed sample size n = 100. We also perform this comparison of a congeneric dataset comprising of activities of 95 amine compounds.

4.1 Simulated dataset

For each of the validation methods applied, we report their q^2 and MSPE obtained using the two-deep method described above.

<Insert table 3>

Table 3: Performance of all validation methods on simulated data

q^2				
Number of predictors (p)		100	500	1000
5-fold cv		0.82(0.041)	0.71(0.111)	0.56(0.176)
LOO-cv		0.86(0.031)	0.8(0.076)	0.74(0.101)
External validation	Min	0.34(0.299)	0.23(0.342)	0.11(0.371)
	25th percentile	0.72(0.046)	0.7(0.109)	0.61(0.155)
	Median	0.79(0.035)	0.77(0.08)	0.69(0.136)
	75th percentile	0.84(0.026)	0.83(0.064)	0.76(0.112)
	max	0.95(0.015)	0.93(0.033)	0.89(0.069)
Multi-split validation		0.83(0.039)	0.75(0.089)	0.67(0.137)
MSPE				
Number of predictors (p)		100	500	1000
LOO-cv		1.8(0.321)	3.05(0.892)	4.69(1.333)
5-fold cv		1.54(0.271)	2.18(0.601)	2.74(0.74)
External validation	Min	0.26(0.135)	0.57(0.229)	0.72(0.245)
	25th percentile	1.23(0.235)	1.61(0.438)	2.11(0.631)
	Median	1.63(0.291)	2.24(0.588)	2.99(0.945)
	75th percentile	2.04(0.356)	3.07(0.754)	4.13(1.316)

	max	3.73(0.825)	6.72(1.96)	10.52(4.192)
Multi-split validation		1.65(0.286)	2.46(0.611)	3.33(1.017)

Table 3 reports values of the two metrics for the four validation techniques, considering three values of the predictor dimension *p*. For external validation on each dataset, we report minimum, 25th percentile, median, 75th percentile and maximum of MSPE and q^2 from the 100 train-test splits performed during the multiple external validation process. For each method, we compute average and standard deviations (in brackets) of all the metrics above across all 100 datasets for a value of *p*.

LOO-cv has the best performance across all predictor dimension and both metrics. All methods perform worse as dimension of the descriptor space grows, which is expected because of higher amount of noise introduced by more predictors.

The main issue with external validation, *which previous studies (e.g.* [4] [5]*) have not captured, is the high degree of variability in its performance depending on which subset of the full data is chosen as the validation sample*. The minimum and maximum values indicate that depending on the train-test split, the average two-deep q^2 from an external validation procedure can vary between 0.34 to 0.95 for p = 100, 0.23 to 0.93 for p = 500 and 0.11 to 0.89 for p = 1000. The instability of external validation becomes even more severe if we consider the minimum q^2 values and their high variance. The q^2 values from some random splits even turned out to be negative. This means that MSPE is more than the total sum of squares in the test set, indicating very high amount of noise in the fitted model, i.e. severe underfitting.

More than 50% of the external validation splits have worse performance than LOO-cv for both q^2 and MSPE across different values of *p*. This indicates that for higher number of predictors, LOO-cv is more likely to result in a QSAR model that is more predictive.

4.2 Amines dataset

We report results from the LASSO model validation analysis of the 95 compounds congeneric amines dataset in **Table 4**. For multi-split validation, we report the minimum, 25th and 75th percentiles, median and maximum here, and relegate results from all splits to Table S3 in supplementary material. In this case, both LOO and 5-fold cv have larger two-deep q^2 values than Multi-split validation, as well as half of the random external validation splits. The minimum q^2 value for external validation is as low as 0.15. One of the random train-test splits in external validation yielded a q^2 value of -0.003. This underscores a severe limitation of the external validation procedure: if such a split of a real-world dataset is used to validate a QSAR model, the whole modelling practice becomes nothing but a waste of resources.

<Insert table 4>

Table 4: Performance of all validation methods on 95 amines data

Number of predictors (*p*)		q^2	MSPE
LOO-cv		0.77	0.86
5-fold cv		0.73	1.05
	Min	-0.003	0.27
	25th percentile	0.65	0.61

External validation	Median	0.73	0.88
	75^{th} percentile	0.83	1.28
	max	0.94	2.01
Multi-split validation		0.71	0.97

5. Discussion

As stated by Johnson [33], the multiple caveats of the present scenario of QSAR modelling include incorrect statistical models used, overfitting, chance correlation and above all, the presence of multiple solutions. We think this is one of the main reasons that external validation results from the analysis of HDLSS data show so much instability over different train-test splits. Because of the small sample size and high number of predictors, local solutions on the rough response surface get masked by the high between-sample variation. Consequently, solutions obtained from most of the training models do not adequately capture variations in the test partition of the data: resulting in poor out-of-sample prediction performance.

QSAR modelling is extensively used in academia and industry setup for virtual screening of chemical compounds [1, 34]. These compounds often have lasting impact in human health and diagnostics and protection of the environment around us. In this situation, a *laissez-faire* use of external validation using small validation sets can have enormous consequences if the wrong compounds get selected in the screening procedure. Thus, it is difficult to overstate the importance of proper, stable and rigorous validation methods. This paper provides an objective assessment of the above problem in a specific modelling scenario across several relevant datasets, and suggests the leave-one-out cross-validation as a means to bypass the issues caused by single-split external validation *when the modelling goal is increasing predictive accuracy*. We intend to carry out detailed studies in future covering more chemometric datasets, modelling techniques, cross-validation methods and intelligent sample-splitting methods (e.g. D-optimal design) to explore the different nuances in the paradigm. We sincerely hope that the current work motivates other researchers in the field to explore in detail the validation aspects of QSAR modelling.

Conflict of Interest

The authors declare that there is no conflict of interest.

Supplementary Material

The supplementary material contains tables listing names of all descriptors in the 95 amine mutagens data, and external validation results from all train-test splits in the four datasets.

References

[1] S. C. Basak, "Philosophy of Mathematical Chemistry: A Personal Perspective," *Hyle- Int. J. Phil. Chem.*, vol. 19, pp. 3-17, 2013.

[2] S. C. Basak and S. Majumdar, "Current landscape of hierarchical QSAR modeling and its applications: Some comments on the importance of mathematical descriptors as well as rigorous statistical methods of

model building and validation: Volume 1," in *Advances in Mathematical Chemistry and Applications*, Bentham e-Books, 2016, pp. 251-281.

[3] S. C. Basak, D. Mills, D. M. Hawkins and J. J. Kraker, "Proper statistical modeling and validation in QSAR: A case study in the prediction of rat fat-air partitioning," in *Computation in Modern Science and Engineering, Proceedings of the International Conference on Computational Methods in Science and Engineering 2007 (ICCMSE 2007)*, Melville, NY, 2007.

[4] A. Golbraikh and A. Tropsha, "Beware of q2!," *J. Mol. Graphics Model.*, vol. 20, pp. 269-276, 2002.

[5] A. Cherkasov, E. N. Muratov, D. Fourches and others, "QSAR Modeling: Where Have You Been? Where Are You Going To?," *J. Med. Chem.*, vol. 57, no. 12, pp. 4977-5010, 2014.

[6] P. Gramatica and A. Sangion, "A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology," *J. Chem. Inf. Model.*, vol. 56, no. 6, pp. 1127--1131, 2016.

[7] D. Hawkins, S. Basak and D. Mills, "Assessing model fit by cross-validation," *J. Che. Inf. Comput. Sci.*, vol. 3, pp. 579-586, 2003.

[8] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. R. Statist. Soc. B*, vol. 58, pp. 267-288, 1996.

[9] D. Hawkins, S. Basak and D. Mills, "QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics," *Environ. Toxicol. Pharmacol.*, vol. 16, pp. 37-44, 2004.

[10] P. Filzmoser, B. Liebmann and K. Varmuza, "Repeated double cross validation," *J. Chemometrics*, vol. 23, pp. 160-171, 2009.

[11] D. Bauman and K. Baumann, "Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation," *J. Chemoinformatics*, vol. 6, p. 47, 2014.

[12] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Statist. Soc. B*, vol. 36, pp. 111-147, 1974.

[13] Y. Yang, "Consistency of cross validation for comparing regression procedures," *Ann. Statist.*, vol. 35, pp. 2450-2473, 2007.

[14] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *J. Econometrics*, vol. 187, pp. 95-112, 2015.

[15] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Ann. Appl. Statist.*, vol. 5, pp. 232-253, 2011.

[16] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemom. Intell. Lab. Syst.*, vol. 56, pp. 1-11, 2001.

[17] R. P. Sheridan, "Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction," *J. Chem. Inf. Model.*, vol. 53, no. 4, pp. 783--790, 2013.

[18] S. Majumdar and S. C. Basak, "Exploring intrinsic dimensionality of chemical spaces for robust QSAR model development: A comparison of several statistical approaches," *Curr. Comput. Aided Drug Des.*, vol. 12, no. 4, pp. 294-301, 2016.

[19] S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal and G. D. Veith, "Topological indices: their nature, mutual relatedness, and applications," *Mathematical Modelling*, vol. 8, pp. 300-305, 1987.

[20] A. Debnath, G. Debnath, A. Shusterman and C. Hansch, "A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella typhimurium TA98 and TA100," *Environ. Mol. Mutagen.*, vol. 19, pp. 37-52, 1992.

[21] S. C. Basak, D. K. Harriss and V. R. Magnuson, "POLLY v2.3," Copyright of the University of Minnesota, 1988.

[22] MolconnZ v4.05, Quincy, MA: Hall Ass. Consult., 2003.

[23] S. Basak, G. Grunwald and A. Balaban, "TRIPLET," Copyright of the Regents of the University of Minnesota, 1993.

[24] Sybyl Version 6.2, St. Louis, MO: Tripos Associates, Inc., 1995.

[25] J. Stewart, MOPAC Version 6.00, QCPE #455, Frank J. Seiler Research Laboratory: US Air Force Academy, CO, 1990.

[26] S. Majumdar, S. C. Basak and G. D. Grunwald, "Adapting interrelated two-way clustering method for quantitative structure-activity relationship (QSAR) modeling of mutagenicity/non-mutagenicity of a diverse set of chemicals," *Curr. Comput. Aided Drug Des.*, vol. 9, pp. 463-471, 2013.

[27] S. C. Basak, B. D. Gute and G. D. Grunwald, "A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters," in *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A. T. Balaban, Eds., Amsterdam, The Netherlands, Gordon and Breach Science Publishers, 1999, pp. 675-696.

[28] S. Majumdar and S. C. Basak, "Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets," *Curr. Comput. Aided Drug. Des.*, vol. 11, pp. 117-123, 2015.

[29] S. Nandi, M. Vracko and M. C. Bagchi, "Anticancer Activity of Selected Phenolic Compounds: QSAR Studies Using Ridge Regression and Neural Networks," *Chem. Bio. Drug Des.*, vol. 70, pp. 424-436, 2007.

[30] S. C. Basak, R. Natarajan, D. Mills, D. M. Hawkins and J. J. Kraker, "Quantitative structure-activity relationship modeling of juvenile hormone mimetic compounds for Culex pipiens larvae, with a discussion of descriptor-thinning methods," *J. Chem. Inf. Model.*, vol. 46, pp. 65-77, 2006.

[31] G. Ghasemi, S. Arshadi, A. N. Rashtehroodi and others, "QSAR Investigation on Quinolizidinyl Derivatives in Alzheimer's Disease," *J. Comput. Med.*, vol. 2013, pp. 1-8, 2013.

[32] Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih and M. Aziz, "High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO," *J. Chemometrics*, vol. 29, pp. 547-556, 2015.

[33] S. R. Johnson, "The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy)," *J. Chem. Inf. Model.*, vol. 48, pp. 25-26, 2008.

[34] S. C. Basak and S. Majumdar, "Editorial: The Importance of Rigorous Statistical Practice in the Current Landscape of QSAR Modelling," *Curr. Comput. Aided Drug Des.*, vol. 11, no. 1, pp. 2-4, 2015.