Reviewers' comments 1:

Comments:

The authors try a few methods of random cross-validation on four small datasets (only one of which is “real”) and compare the goodness of prediction with those against LOO. LOO appears to get the higher q2 by a small amount.

I have several issues with this paper:

1. The conclusions about “stability” are fairly obvious.

With small datasets (say < 1000 compounds), the q2 of the test sets will depend on the specific random split. Therefore “external validation”

(one split) is probably not a very good method, and one should use “multiple splits” (doing several training/test splits). My experience is that all types of random splits with a given training/test ratio will give very similar results for larger datasets. In contrast, there is only one way to do LOO, i.e. LOO does not depend on any stochastic process, so it is always going to be “stable.”

2. The reason LOO appears better is probably not what the authors say.

The q2 of a test set will tend to be higher as the training/test ratio becomes higher. This is a simple “domain applicability” effect. When training/test is high, each compound in the test set is more likely to have a neighbor in the training set. In LOO the training/test ratio is maximally

high, so we expect maximum q2.

3. Getting the highest q2 for a test set is not the goal of validation. The point of validation is to guess what the q2 of true prospective prediction will be. For that one has to leave random splits altogether, because these tend to be much too optimistic. One publication discussing this is J. Chem.

Inf. Model. 2013, 53, 783-790. (That author uses R2, but it also applies to

q2.)

4. There are too few examples to say anything.

To demonstrate anything general, we would have to look at more realistic QSAR datasets of various sizes.

Reviewers' comments 2:

This paper attempt to address an important question --- is the widely used external validation effective and reliable in evaluation of the predictive power of QSAR regression models. This work compared a number of sophisticated validation methods including leave-one-out, K-fold, external and multi-split validation based on LASSO regression. The comparison was performed on both simulated and real datasets.

In fact, the current study can be considered as a follow of the highly cited paper by Golbraikh and Tropsha (Beware of q2!," J. Mol. Graphics Model., vol. 20, pp. 269-276, 2002). However, the Tropsha’s group has performed systematic studies on the topic of QSAR validation and they also published a number of references related to this topic. The authors should introduce and cite these works.

As a concept work, only one real dataset (95 congeneric amine mutagens) used in this study to evaluate the validation methods is not “comparative”

and “systematic”. The authors should consider to extend their analysis on some classic QSAR datasets, such as the ACE dipeptide inhibitors and the steroid compounds used by CoMFA.

It is expected that external test is unstable for random splitting of small sample sets --- the authors would evaluate other splitting methods such as D-optimal or experimental design.

More importantly, I do not agree the conclusion “we recommend using the LOO procedure for validating QSAR models built on high-dimensional small-sample data.” LOO is “stable”, but does not mean “reliable”.

The LOO can only be used as a necessary condition, but not sufficient condition, for a predictive QSAR model. In fact, the Monte Carlo cross validation (MCCV) seems a more rigorous method available for QSAR validation. The authors should consider the MCCV