Reviewer #1: The manuscript is not written well. The authors are rather sloppy and miss to specify important technical details.

Major criticism:

Section 2.3 is written rather sloppily. The brief description of RF is partly incorrect. The randomly chosen subsets of variables are generated for each candidate split in each tree.

Thanks for catching this error. We have corrected it in the first paragraph of 2.3.

This is a scientific paper. Provide proper references for the variable importance measure. Please provide a credible reference to the scientific literature rather than a company web page.

Duly noted. We have now replaced the link with two references.

Section 2.4 is in parts wrong which is most unfortunate and surprising. The authors should know better. I agree that a single split in test and training set should not be used and that multiple splits are far better. In Ref [24] Hawkins used a LOO-CV for model assessment which is a perfect, almost unbiased way to estimate the prediction error. Tropsha used LOO-CV in Ref [27] as fitness function for variable selection (i.e. not for model assessment but for model selection!). In this case, it overfits. But, if you do no hyper-parameter tuning (here: variable selection) but model assessment on fixed models, then LOO-CV cannot overfit. So the statements w.r.t. overfitting in section 2.4.1 are incomplete to say the least. The two applications of LOO-CV cannot be compared since LOO-CV serves a different purpose.

If multi-split external validation is done correctly, the partitioning impacts bias in the prediction error estimate but in no way overfitting. Overfitting is controlled by the way you do your model selection, i.e. it is controlled by the way internal validation alias hyper-parameter-tuning is done. The authors should always differentiate the purpose for which CV is used (model assessment vs. model selection) in order not to confuse readers.

Thank you for raising a very important point. Use of validation methods has been a contentious issue in QSAR literature over the last few years, and your comment clarifies it quite a bit. We make it clear in 2.4.1 that we use multi-split CV for model assessment, and include in a separate note after 2.4.1 a small discussion about the distinction between model selection/assessment.

Section 2.4.2: "Even though training the full model based on all might intuitively look like a correct thing to do, this naïve approach would use information from holdout compounds in the training step, thus synthetically inflating all model metrics."

Specify "on all"! What does that mean? All compounds? That would indeed by naïve. Nobody would do this since there would essentially be no holdout compounds anymore. Please state that precisely.

Indeed, we meant ‘all compounds’. We have made this clear in Section 2.4.2.

The term "Two-deep" is from John Tukey (The Handbook of Social Psychology, 2nd ed., Vol. 2, Eds.: G. Lindzey, E. Aronson, 1968, Page 147; ISBN: 0-201-04263-0) and not from references [8, 25, 29-31]. By the way, avoid excessive self-citations. There are far more researchers using and evaluating two-layered cross-validation schemes than the single group cited. There are also far earlier papers on the topic than those cited. Since the topic of two-layered CV is stressed here, a proper bibliography must be given.

We were indeed not aware of the origin of the term ‘two-deep’. To address this point, we have added your supplied reference to section 2.4.2, and added three more references by other researchers.

Random Forests by default uses bagging as the inner loop of re-sampling, thus bootstrapping, and not cross-validation. Does the R implementation provide cross-validation? How many trees were trained. How many splits in the outer loop were computed. How was the partitioning in the outer loop? Please be precise on your set-up and specify the technical details in full.

Thanks for this constructive suggestion. We have now added a full subsection in section 3 to include all technical modelling details.

Minor detail:

The references need to be adapted to the journal's style.

Reviewer #2: The main objective of this paper was to use computed molecular descriptors in the prediction of BBB entry of a diverse set of 415 chemicals, and find out influential descriptors with potential mechanistic interpretations. But, the work is basic and simple, and no obvious importance was found.

1. There are many machine learning (ML) methods for building QSAR models. The authors used only one of ML methods, RF. Why not use other ML methods and compare the different performance of ML methods?

2. Only the basic predictive performance parameter (AUC) and the top 20% lift were used. It is suggested to add the sensitivity (SE), specificity (SP), Matthews Correlation coefficient (MCC) and overall classification accuracy (CA) to the text including discussion.

We have added performance analysis for the other four metrics in Table 3 and Figure 1, and discuss the outputs subsequently.

3. According to the OECD principles for the validation of QSAR, any QSAR should be associated with a defined domain of applicability (AD). Please add AD calculation and analysis in the text.

4. The authors used the top 20% lift as the validation parameter. Why was it top 20% lift for validation instead of top 10%? Please enter in justification in the text.

5. The 'refs' (the second line from the bottom of page 13) didn't provide the reference number. Please add the corresponding references.

Thanks for catching this. We have added two references here.