Resubmission Cover Letter - Chemically Aware Model Builder (camb): An R package for property and bioactivity modelling of small molecules

The manuscript has only changed very slightly, but we have made significant changes to the package in line with the issues put forward by reviewer #1. There are some formatting issues with the latex generated PDF (which uses the required template) but these were experienced by a co-author submitting successfully to the same journal who said that they were taken care of in the typesetting process on that occasion.

Reviewer #1:

* While I, personally, believe that the package essentially reinvents the wheel in a number of aspects (PCA, RMSE, feature selection etc) the package does not do anything wrong and could be useful for an R user who is new to QSAR modeling. At the same time, from an educational point of view I believe that packages such as this are somewhat of a disservice - for example, it encourages one to avoid having to learn the details of the caret package for feature and model selection.

We agree with this criticism. We tried to address this concern to some degree in the paper: “It must be noted that *camb* does limit practitioners to a limited but easily used workflow to begin with, however, each function can be used with non-default parameters to fulfil the more versatile needs of more experienced users. In fact, experienced users are encouraged to use the full vignettes capacity of the caret package as these authors recommend it highly.” This has now been changed to make this reviewer’s (and ours) much more explicit: “It must be noted that *camb* does limit practitioners to a limited but easily used workflow to begin with. Experienced users, or those that intend to practice machine learning in R extensively are encouraged to neglect this basic wrapper completely on their second training attempt and learn how to use the caret package from the caret related vignettes directly. These authors cannot recommend the recommend the caret package highly enough.”

The abstraction for very new users can be quite useful though as it reduces training and model use for new predictions to a small number of code lines. An example of where a new user may get caught up with *caret* (as is our experience already) is the application of the saved transformation to new data points requiring prediction. Our simplified wrapper abstracts these complications away. Users can always look into these simple wrapper functions to discover how they work on the inside.

* However, I was not able to install the package using the instructions described in the paper. I think this is a significant problem that prevents acceptance of the current paper. I'm running OS X 10.9.5 with R 3.1.1.

This issue should now be resolved. We replicated this problem on a macbook air running OS X 10.10.1 and then fixed it by incorporating Indigo’s changes that allow OS X 10.10 support. This allows it to compile Indigo and install the R package on our OS X 10.10 system and we suspect that it will also resolve reviewer #1’s installation issue too.

* I see that the package includes the sources of the gfortran compiler (!)

We weren’t sure this inclusion was necessary and it has now been removed.

* Another way to better organize the code would be to shift large JAR files (e.g., the CDK jar file) and other dependencies (Indigo), which don't change as rapidly as the main camb code base into a separate R package. This way, the dependency package can be installed once (and updated infrequently) while the main camb package can be updated more frequently. I assume the intent is to distribute the package via CRAN and in its current form (745MB source package) it is unlikely to be accepted on CRAN.  Even if it is not meant to be distributed via CRAN, a 745MB package is unusual for most R users.

Total git repo download size is now 24.70 MiB. While splitting the package into two parts is a good idea and will be considered should additional third party tools get added, considering this drastic reduction in size, we would prefer to have the package fetched as a single package for now which will make tutorial sessions that use it much quicker to set up. We don’t think that our package will be accepted on CRAN because it houses compiled code (they can’t be sure this doesn’t contain viruses so they don’t host it), so we had planned to have it obtainable from out GitHub repository. However, we will further investigate if it might be possible to get it accepted on CRAN as this would create an easier installation step for the new user.

* Finally, since the package is meant to be used as a tool, the lack of unit testing as a means to ensure that that the package does what it is supposed to do is is dissapointing. Before publication, the authors should ensure that unit tests (via RUnit, testthat or other public R unit testing frameworks) should be included in the package.

The package *testthat* has been used to check that the intermediate files generate through the logS prediction example workflow match those that have been generated already and saved as reference files. If reviewer #1 can suggest any additional types of unit tests we could add, we would be grateful.

Reviewer #2:

This reviewer’s comments were extremely positive and brought forward nothing to address. These comment appear below for ease of reference.

The authors have done a wonderful job of providing a single open source platform (a framework to interface with R package) that will enable researchers to carry out property and bioactivity modeling of small molecules. The "Chemically Aware Model Builder" (camb) provides either new functions or interface to already existing R-packages to carry out all essential steps in QSAR/QSPR: compound standardization, molecular descriptors calculation (for both small molecules and proteins), and variable selection, generation of predictive models, cross validation and visualization and prediction for external data set.   
  
The authors also provided two important case studies with complete steps along with the working codes: the first example illustrates how to build a QSPR model for "Aqueous solubility, using 1606 small molecules and 211 molecular descriptors".  The models were built using three machine learning methods: SVM, Random Forest and Gradient Boosting Machines (GBM).  The authors have also shown the utility of ensemble modeling techniques, greedy optimization and model stacking.  The second case study deals with the enzyme inhibition of COX-1 and COX-2 inhibitors.  In both cases, the validation of model was presented in accordance with the QSAR community.  As the authors claim, this open access camb package and the provided examples will serve as good template for teaching both undergraduate and graduate students in academic settings.