I would like to thank the reviewers for their comments on the original submissions. I have prepared a revised manuscript that address the comments and my responses are detailed below.

**Reviewer: 1**

**(1) Please, note that there are numerous typographical errors throughout the manuscript (which is quite disturbing).**

**(2) Chosen methods and descriptors should at least be referenced. For example, no references are provided for BCI and CDK.**

The manuscript has been thoroughly proof read and references included where appropriate

**(3) The derivation of RF models is essentially not described (except of making reference to default parameters settings in R). It remains largely unclear how these models were developed.**

The text has been updated to be more specific on the development of the random forest model.

**(4) The selection of training and test sets should be clearly specified in the Methods section. For example, how were the RF models trained and evaluated for which predictions of pairwise SALI values are reported in Figure 2? It is hoped that the reported SALI values were not predicted for the training data.**

We have updated the text to note the training/test split and have include results for both sets in the text. However, due to the similar peformance of models built using the whole dataset (and the fact that the RF approach trains and tests on out-of-bag data implicitly), we used models built using the entire dataset (save for some hold out molecules) in the latter half of Section 4.

**(5) No support is provided for the assumption that the RF models are relatively robust even for datasets of small size. In fact, the author later on states several times that limited model quality might be due to the use of small compound sets. Appropriate validation will be essential here.**

We have removed this statement as it is not necessarily a function of the random forest approach, and more a function of the nature of the dataset. Small datasets, if appropriately constructed can lead to good RMSE's. In our case, we investigated this by sampling from the Cavalli pairwise

dataset, using different sample sizes. As shown in the attached figure, model performance is clearly a function of sample size. Importantly, for certain small samples, the model performance can be quite good. But of course, this is due to luck.

**(6) There is only limited statistical assessment/validation of the models. Statements made in the text that R2 values are reasonable and the models perform relatively well are difficult to reconcile. Rather, Table 2 indicates that the R2 values were rather poor in several instances.**

The text has been updated to read more quantitatively and avoid ambiguous statements

**(7) It appears that there are systematic prediction errors for pairs. At low value ranges (that are not relevant for activity cliff assessment), SALI values are consistently over-predicted. By contrast, at high value ranges (that are relevant), SALI values are consistently under-predicted, in part significantly. This is not investigated, although the apparent errors at high value ranges are the perhaps most critical features of the reported RF models.**

Yes, there do appear to be such systematic errors - at this stage it's not clear what is the cause. However we have noted that while it is clearly evident for the Cavalli dataset, it is not as obvious for the ChEMBL datasets and in some cases is the reverse of what we see in the Cavalli case. For example, for the Dai dataset, we observed that lower SALI values are overestimated (whereas the Cavalli predictions are underestimaetd at low SALI values).

We hypothesize that the descriptors employed do not differentiate between observations with low and high SALI values very well, leading to overestimation of small values and underestimation of high SALI values. We have justified this by comparing descriptor distance distributions between the group of observations with low SALI values and those with high SALI values. The two distributions nearly completely overlap, supporting this hypothesis.

**(8) In the section Extending a landscape, predictions are finally reported for only three (!) molecules taken out of each data set. Carefully put, the predictions are heterogeneous. On the basis of the results reported for this little bit of evaluation, one would be hard pressed to make a case for the ability of the approach to predict activity cliffs with any certainty.**

We have updated this section to include several more hold out molecules. It should be noted that this section examines the model from a different viewpoint compared to Section 4. Specifically, Section 4.1 develops models in which we remove molecules from the original dataset. Given the size of the

original datasets, removing more than 10 or so, leads smaller pairwise datasets. Simultaneously, the predictions for this validation stratgey are not just for the 10 hold out molecules - it will be 10 $\times m$ where $m$ is the number of the molecules in the original datasets. Thus the number of predictiosn range from 250 to 340 values. Summarizing more than a few of these hold out molecules becomes cumbersome.

**(9) For the prediction of activity cliffs, SALI values must be calculated for compound pairs. This is the central idea of the approach. Accordingly, fingerprint descriptors and potency values must be combined in some ways.**

**(9.1.) It is nowhere stated how activity values are treated for compounds forming a pair. Are activity differences calculated at the pair level? If so, how are activity differences compared for the calculation of pair-based SALI scores - as a difference of differences? It should be noted predicted potency values in Figures 4-6 are essentially all over the place.**

The entire premise of the proposed method is that one evaluates SALI values for a dataset that has observed activities. We state specifically (Equation 1) how the activities are combined when molecules are considered pairwise.

It's not clear what the reviewer means by *"... how are activity differences compared for the calculation of pair-based SALI scores ..."*. The SALI value itself represents a pair of molecules and in this calculation, we consider the absolute activity difference. We do not consider pairs of pairs

**(9.2.) To obtain a fingerprint descriptor for a pair, the author aggregates fingerprints of individual molecules, finally, by simple averaging**

In fact, we do not generate pairwise fingerprint descriptors. Fingerprints are only used to evaluate the similarity term in the calculation of the SALI value.

**For comparisons of pairs, Tanimoto similarity is then calculated for averaged fingerprints, which presents an artificial assessment of pair similarity (one can easily come up with a few hypothetical compound similarity relationships that make this assessment questionable at best). As a consequence, activity cliffs are considered for which Tanimoto similarity calculations yield values between 0.2 and 0.3!**

I believe that these comments are based on the discussion in Section 3.1 and it appears that this

section of the paper was not clear.

Firstly, we have never mentioned the averaging of fingerprint descriptors. Rather, the averaging is applied to the topological descritpors (i.e., the X variables). Specifically, we generate a descriptor vector for a pair of molecules by averaging the (topological) descriptor vectors of the individual members of that pair.

The goal of Section 3.1 is to indicate that even though the Y variable and the X variables encode structural information (the former indirectly and the latter directly), there is little correlation between the individual (aggregated, topological) descriptors and the Y variable (where sructural information is included via fingerprints). To quantify this, we evaluate the Pearson correlation between each (aggregated, topological) descriptor and the Y variable and show that the maximum value is very low ($R^2 < 0.15$).

This discussion does not involve fingerprint-based Tanimoto similarities at all. In fact, if one does indeed consider the pairwise Tanimoto values of the molecules within a dataset, we do in fact observe an appreciable number of pairs wth $T_c > 0.8$.

In refering to pairs with $T_C$ values between 0.2 and 0.3 – it is true that such pairs are not really activity cliffs in the original sense of the term. As described below, defining a threshold SALI, above which a pair represents an activity cliff, is rather subjective.

**Thus, on the basis of these similarity calculations, one could never decide which level of similarity might be cliff relevant. In this respect, the descriptions of activity cliff by the author are quite telling he speaks of predicted versus true activity cliffs, of relatively accurate activity cliffs, and of most significant activity cliffs that are in fact not very significant activity cliffs in an absolute sense, and so on.**

I agree that the wording of this discussion is somewhat sloppy. The text has been updated to be more quantitative.

Yet at the same time, the concept of an activity cliff is subjective. In many cases, it is akin to saying *"I'll know it's an activity cliff when I see it"*. For example, a pair of molecules exhibiting a $T_c = 0.9$ and 1000x difference in acivity is obviously an activity clif. Would a pair with $T_C = 0.85$ and 50x difference in activity not be an activity cliff? From this point of view, I do not think it unreasonable to refer somewhat qualitatively to the "cliffness" of an activity cliff.

Certainly, with respect to the previous comment, a pair of molecules with $T_c = 0.29$ would likely not be regarded as a cliff - I specifically note that such a case is not a significant cliff. To address this aspect of SALI values, I have presented plots of dissimilarity versus activity ratio (Figures 7 and 9) which allow one to more directly determine "significant" versus "insignificant" cliffs.

**Furthermore, there might be many instances of activity and fingerprint similarity relationships between pairs of compounds that might yield high pair-based SALI scores dominated by either the activity or similarity component, without forming a true activity cliff. The conclusion of this reviewer is that it is currently not possible to capture activity cliffs with any degree of certainty on the basis of the pair calculations, as reported.**

While it is true that the SALI definition itself allows for high SALI values to be generated for pairs of molecules that would be small/insignificant cliffs (highly similar structures with small difference in activity), the methodology does let one rank pairs of molecules and as indicated in Figures 7 and 9, one can compare the strucural similarity and activity difference (or ratio) in 2D, allowing one to separate "true" cliffs from apparent ones.

In addition, Figures 8 and 10 do suggest that the approach does capture actual activity cliffs - it so happens that the datasets considered do not contain too many of them, to allow one to provide a large scale validation.

As noted in Section 5, predictive performance can likely be increased by a larger, more diverse descriptor pools as well as alternative modeling methods. The goal of this paper is to suggest that the proposed pairwise methodology does let as address the identifcation of activity cliffs in a prospective manner.

**(10) It also remains largely obscure how an activity landscape should be extended following this approach by compounds for which no activity information is available. Should the activity first be predicted by an RF model and then used for pair-based SALI score calculations? One would hope not.**

The original description was misleading. In fact, the approach does not allow one to obtain predicted activity values, due to the fact that the sign of the difference in activities in the SALI formula is ignored. While one could create another model in parallel, that predicts signed activity differences (and thereby determine whether the activity derived from the predicted SALI value should be less than or greater than that of the molecule with observed activity), this would probably not

be very reliable! The manuscript has been updated to remove the dicussion on extending SAR landscapes.

**Reviewer: 2**

**1. The manuscript is poorly written. There are many typos, grammatical errors, missing references, and incomplete descriptions of methods and results.**

The text has been thoroughly proof read to remove typos and incorrect grammar. References have been added and methods have been described more clearly. It's not clear which descriptions were incomplete. However, the methodology section now reads more coherently, and some aspects are now clearer based on other reviewer comments.

**2. Qualitative assessments of results are not acceptable. The authors should give statistical evaluations and corresponding metrics for models and their predictions.**

This was an oversight on our part. The use of qualitative terms has been removed and results and comparisons are quantitatively characterized. For example, the random forest models built to predict SALI values are now compared to Y-scrambled models to ensure that the the results are not purely due to chance.

**3. The tests sets are too small to make any meaningful conclusions.**

**4. SALI values needs to be determined for pairs of molecules - is that not the essence of the methodology?**

Yes. For a given dataset, with measured activity values, one can indeed evaluate the SALI values. The proposed methodology takes these "measured" SALI values for a training set and then predicts the SALI values for a new molecule with each member of the training set. As with traditional QSAR models, a complete validation would require one to measure the activity of the new molecule and evaluate the actual SALI values. In absence of experimental activities for the test set, we resort to a training/test split, which is an accepted approach in QSAR modeling.

**5. There is no discussion/direction of how to use the methodology in new applications.**

Section 4.1 in the original manuscript specifically addresses how one might use the proposed methodology. This has been made more explicit in the updated manuscript.