**Response to Reviewers**

We would like to thank the reviewers and editor for their helpful comments on our manuscript, and we have made substantial revisions to the manuscript to more clearly highlight the purpose of our study. We realize that the primary goal of our study was somewhat unclear. Our goal was not to propose a novel alignment algorithm, but rather, we aimed to obtain a comprehensive, high-quality dataset (alignment) and corresponding phylogeny for vertebrate biogenic amine receptors that can serve as a useful resource. In particular, in an analysis of this size, it is often unclear which sequences are worth retaining vs. which sequences may be inappropriate to include and thus confound analysis. Therefore, our primary contribution in this study was to develop an objective strategy, based on structural aspects, to determine which sequences should be confidently retained, and similarly to identify sequences that may be inappropriate to include. We have updated the manuscript substantially to better convey this overarching message.

**Comments by the Editor**

1. I think it's important to highlight the fact that "structure" is only indirectly informing your procedure—through predictions by GPCRHMM. Calling this "structurally-curated alignment" may convey the wrong idea. How about "HMM-informed alignment" so something along these lines?

We agree that protein structure enters our analysis indirectly. At the same time, we feel strongly that the important aspect of our analysis is that we consider structural information in our analysis, not that we use an HMM model. If we had a different, non-HMM-based computational tool to infer structural information, nothing else in our analysis and approach would change.

To address your concern, we have changed the title so that it now uses the term “structurally-informed” instead of “structurally-curated.” We believe this is a reasonable compromise between our respective viewpoints.

1. By design, the retained sequences have consistently predicted segments according to GPCRHMM. It is therefore natural that the resulting alignments look more compact. However, it's unclear that compact is necessarily better—at least not in an evolutionary sense (aligned residues => residues evolved from a common ancestral residue). Insertions that are specific to particular sequences necessarily lead to additional columns that look mostly "empty", which may offend our sense of aesthetics, but they might well be correct and in any case have little impact on downstream analyses (such as tree inference).

We completely agree that there is no *a priori* reason to expect that more compact alignments are better. We do not believe that gaps are inherently problematic. Rather, in this specific case, gaps simply reflect the underlying actual problem: Large gaps tend to appear when extramembrane domain regions align spuriously with transmembrane domain regions. We have clarified the manuscript to emphasize this distinction.

1. I agree with reviewer 2 that the likelihoods obtained from different MSAs (and/or with or without masked residues) are not directly comparable. I also agree with him that bootstrap support would be more meaningful (though note that bootstrap support is a measure of random error but is blind to systematic error).

We completely agree with the editor (and reviewer 2) on this point. Therefore, we have revised the manuscript to focus strictly on the masked structural alignment, and we only compare, via the widely-used likelihood ratio test (LRT), two phylogenies: one built without partitions, and one built with two partitions representing extramembrane and transmembrane domains. As these phylogenies were constructed with the same dataset, they represent nested models and thus their likelihood scores are indeed directly comparable.

Furthermore, while bootstrap values are indeed meaningful for assessing support of a given tree topology and model, we do not feel that bootstrap values have the ability to select the best tree among competing options. They indicate support for a given phylogeny and its corresponding model, and are therefore not directly comparable between phylogenies, particularly when there are topological differences. Therefore, we have computed bootstrap values for the both the partitioned and unpartitioned phylogeny, and we include them in the open-source github repository for readers to examine as desired. In the manuscript, however, we focus on the LRT comparison.

1. "By masking these positions with an ambiguous character, we ensured that each MSA column strictly contained residues belonging to the same structural domain." This is too strong. These are merely predicted to be from the same model; the words "strictly" and "ensure" implies certainty.

We have addressed this and similar word choice issues throughout the manuscript.

**Comments by Reviewer 1**

1. The title is somewhat misleading in that the alignment is based on structure predictions rather that the experimentally determined structure of the receptors. In addition it is not but filtered as, aside from the initial collection of the sequences, there is no manual intervention in the process.

See our response to a similar comment by the editor. We now use the word “informed” instead of “curated”. We believe this word choice resolves the issue.

1. The results rely utterly on the accuracy of Wistrand et al's GPRCHMM program. How accurate is this program? If this program has even a modest error rate, then so too will the alignments.

As noted in the manuscript, previous work by us (Spielman and Wilke 2013) has shown that this program provides excellent predictions in cases where actual structures are available. To further emphasize this point, we now include an additional figure (now Figure 1), and corresponding discussion, that maps GPCRHMM structural predictions onto known several biogenic amine receptor structures. As this figure demonstrates, domain predictions are exceptionally good, lending support to the validity of our final alignment.

1. Table 1 does not seem to be referenced in the text, aside from in the caption to Figure 4.

We thank the reviewer for noticing this omission, and we have included a reference to Table 1 in the text.

1. Results and Discussion, 1st paragraph: "We collected all sequences..." Perhaps "We collected all biogenic amines sequences..." to ensure the reader does not think that all GPRC sequences were collected? Also, typo confidentally should be confidently?

We have edited the sentence for clarity, and we have corrected this typo.

1. Results and Discussion, 2nd paragraph: does the determination of the consensus domains include the sequences about to be excluded? It appears the excluded sequences are included, as in Figure 1 the left-most 'M' in the consensus could be 2 or 3 out of 5 while the left-most 'E' could be 3 or 2 out of 5.

In general, the determination of consensus domains necessarily includes those sequences that will eventually be excluded. This strategy is necessary, as it is unknown which sequences will be removed before a consensus is determined from all sequences. We agree with the reviewer, however, that this figure was somewhat unclear. Thus, we added an ellipsis below the alignments in the flowchart to clearly indicate that the figure represents merely a subset of the entire alignment. We additionally added sentence to the figure caption clarifying this issue.

**Comments by Reviewer 2**

1. The URL in the link at the end of the Introduction is wrong. It should be:

<https://github.com/sjspielman/amine_receptors>

Thanks for pointing this out – we have fixed the URL.

1. When proposing a new approach for sequence alignment, it is necessary to compare its performance to previously published methods. In the case of the present study, the new method can (and should) be compared to methods that were designed for structural-aware alignment, especially the consideration of TM domains, and even specifically for GPCR proteins. The study by Hopf et al. 2012 has done exactly that (doi: 10.1016/j.cell.2012.04.012), based on alignment by HHblits (Remmert et al. 2012, doi:10.1038/nmeth.1818). MP-T (doi:10.1093/bioinformatics/bts640) was also developed for structural-aware alignment of TM proteins, and was found to be more accurate than both MAFFT and HHblits. Another program is TM-Coffee (doi:10.1186/1471-2105-13-S4-S1). Therefore, the publication of a new method requires comparison of performance relative to these previous ones.

We thank the reviewer for bringing these methods to our attention. When looking into these particular strategies, we discovered that HHBlits is not an alignment method but rather appears to be a tool for detecting homology between a given HMM profile and candidate sequences from a database, such as BLAST. In addition, we found that TM-Coffee and MP-T, while robust programs for aligning sequences with additional structural information, are ill-suited for our dataset. In particular, the MP-T usage manual states that the program is not suitable for more than 998 sequences, and TM-Coffee similarly states that it is not suitable for more than 150 sequences. Indeed, in the manuscript introducing TM-Coffee, the largest dataset examined contained fewer than 130 sequences, and the authors stated both that, “Regardless of the improvement reported here in terms of CPU, TM-Coffee remains a relatively slow method. One may argue whether the increased computation cost is worth the improvement reported here.” Thus, TM-Coffee is likely not able to process our large dataset of >3000 sequences.

Importantly, however, we now include references to both MP-T and TM-Coffee in our discussion of the limitations of structurally-aware alignment methods.

1. Which variant of MAFFT was used (e.g. L-INS-i?) and why?

We used the default MAFFT algorithm (FFT-NS-1). The more accurate L-INS-i algorithm would be too computationally expensive to process an alignment of the size we are dealing with. We have included this detail in the methods section of our manuscript.

1. In addition to masking residues that disagree with the structural domain prediction, badly aligned residues can be identified using tools such as HoT (<http://www.ncbi.nlm.nih.gov/pubmed/18229673>) and Guidance (doi: 10.1093/nar/gkq443), and if the aligner is switched - also T-Coffee (or TM-Coffee) consistency scores. It would be interesting to compare and combine the different predictors of alignment errors.

In general, we agree with the reviewer that these methods are useful approaches to identify poorly aligned regions. However, as mentioned in the introduction to this response, the primary purpose of our analysis is not to identify poorly aligned regions in a given alignment but rather to identify and remove confounding sequences in the dataset as a whole. Thus, comparing measures of alignment error is beyond the scope of our current analysis.

1. Page 3 paragraph 3: The conclusion that “our structurally-curated MSA featured far less error than did a structurally-naive MSA” is not supported by the presented results. Alignments with less gaps are not necessarily more accurate. Some alignments may be “over-aligned”, i.e. more compact than the true alignment.

We broadly agree with the reviewer on this point. See our response to Comment 2 of the Editor. We have clarified our discussion regarding the presence of gaps in the naïve alignment vs. structural alignment.

1. The prediction of secondary structure is not a strong indication of alignment quality, because it is merely a prediction. I suggest using actuall protein structures as a benchmark, which is the common practice in the field (e.g. Balibase and also in the abovementioned MP-T paper). There are multiple amine receptors in the PDB, which can be used for structural alignment. Then the accuracy of the proposed alignment can be assessed by comparison to these structural alignments.

As previously stated in our response to Comment 2 of Reviewer 1, we have included an additional figure (Figure 1) in our manuscript that explicitly demonstrates the high accuracy of GPCRHMM predictions using known amine-receptor structures from the PDB. Thus, we now show the reliability of these predictions, which in turn supports the accuracy of our final alignment.

1. “Structurally-aware MSA strongly improves phylogenetic inference”: The likelihood scores of trees built from the naive and structurally-curated alignments cannot be compared using AIC or in any other way because some of the sequences were removed in the filtering. Only likelihood scores of trees built for the same sequence data can be compared. Therefore, I suggest removing the same sequences also from the naive alignment (without re-aligning) and running raxml on that. The resulting tree of the same number of sequences will probably still be different from the tree that was built from the structurally-curated alignment. Then the likelihood scores may be compared. Another way to show the superiority of the tree is if bootstrap scores are higher than the naive tree.

We completely agree with the reviewer on this point. See our response to Comment 3 of the Editor.

1. MrBayes is usually superior to ML algorithms such as RAxML. I suggest applying it to the best alignment. If the run time is too long because of the large number of sequences, then I suggest sampling a smaller number using CD-hit (<http://bioinformatics.org/cd-hit/>).

We agree with the reviewer that Bayesian phylogenetic reconstruction can offer substantial improvements over maximum likelihood methods. However, as the reviewer suggests may be the case, our data set of 3039 sequences is far too large for a Bayesian analysis to accommodate. For example, as stated in the manual accompanying the PhyloBayes (a software that is quite comparable to MrBayes in terms of performance and general properties), “As for the number of taxa, the MCMC sampler seems able to deal alignments with up to 100 taxa reasonably well and has already been used on datasets with up to 250 taxa.”

Moreover, as the primary goal of our analysis is to generate a large, comprehensive dataset of biogenic amine receptors, we feel that sampling a smaller number of sequences to create a phylogeny is outside the scope of this study. While this strategy may prove useful for a smaller-scale analysis, reducing the size of our dataset for phylogenetic inference runs counter to our goal of generating a large, comprehensive alignment and corresponding tree containing all 3039 sequences.

1. Page 5 last paragraph: what was different between the models in RAxML for the different partitions? For the TM domains, a TM-specific matrix should be used.

We thank the reviewer for pointing out an area in the manuscript in need of clarification. While both models use the LG amino acid exchangeability matrix, each partition’s evolutionary model used a unique vector of amino-acid frequencies as well as rate heterogeneity categories (in RAxML’s CAT model). We have clarified this distinction in the manuscript.

Further, we agree with the reviewer that a TM-specific matrix would be more appropriate. Unfortunately, we are unaware of a robust TM-specific matrix, determined using a statistically rigorous maximum likelihood platform. While certain TM-specific scoring matrices have been derived (e.g. JTT-TM and PHAT), these matrices were constructed primarily for scoring alignments and were built from very small data sets using counting-methods (not likelihood). Thus, we feel that the safest strategy to capture TM-specific evolutionary features is to rely on TM-specific amino acid frequencies and evolutionary rates, as we have done. However, if the reviewer is aware of a robust TM-specific matrix, we are happy to consider it.

1. “or this clade represents an avian-specific diversification which the Xenopus tropicalis sequence resembles only convergently“ - convergence is very unlikely to mislead phylogeny reconstruction into joining neighboring clades that were not truly sister clades in their evolution. Also, if HRH-2 contains non-avian sequences than it is not possible that the new clade is an avian-specific diversification.

We agree, and we have rephrased this section of the manuscript accordingly.

1. As in the specific discussion of the TAAR clade, it would be interesting to note which species are present in each of the other subfamilies, and thus infer an approximate dating of their origin. E.g. if some clades are tetrapod-specific then they may be the result of functional specialization that occurred after the divergences of tetrapods. Others that include also fish species are more ancient.

We agree with the reviewer that this would be an interesting point on which to expand. Therefore, we have substantially reframed the section of our manuscript (including the TAAR-specific phylogeny shown in Figure 6) to focus on the taxonomic distribution of TAAR sequences.