Dear Editor,

Thank you very much.

We are grateful for your comment and the ones raised by the reviewers, all of which greatly enriched our perspective on the questions addressed in our manuscript.

I obtained two reviews for the work. It is clear that you have some interesting results in the manuscript, and yet, there is still much work needed for it to be published in MBE. I suggest you revise focusing on three aspects: (1) address all comments; (2) Try to get additional "impact" by adding more analysis; (3) Make an editing effort.

We first addressed all comments raised by the two reviewers. We more specifically focussed on stating the problem up front and elaborating on the added value of this work. We also performed more analyses, shown in supplementary materials, for assessing estimation accuracy for selection pressure (ω) and mutational bias (λ) under various scenarios (number of sites, branch lengths and topology). We also added a figure showing that MF can predict the accurate nucleotide composition at first, second and third codon positions, while the MG model fails to predict them (first figure in supplementary materials, panels A-F).

Finally, we also performed several rounds of proofreading on the manuscript.

Reviewer: 1

The authors present a new codon model that explicitly models the mutation bias observed in coding sequences. As a result, in addition to inference of the mutation bias, their parametrization also affects the inference of the selective pressure. The authors compare their model (referred to by the acronym MF) to the Muse and Gout, 1994 codon model (referred to by the acronym MG), in both a simulation study as well as on empirical datasets.

The authors motivate the development of their model by arguing that accurately modeling mutation bias can lead to better inference of selective pressure and probably other types of biological questions that codon models are applied to.

Although their results do show that their MF model better estimates the ground true mutation bias, they also show that although the MG model fails to achieve inference of the mutation bias as accurately as the MF model, this has a marginal effect, if any at all, on the inference of selection, both in the simulation study and in the empirical data. Therefore, the key question I'm struggling with is whether their heavily parametrized MG model (95 parameters) provides a meaningful added value, relative to what motivated them to develop this model.

Certainly, the mutational misspecification of the MG model does not strongly impact the estimation of ω as shown in our work. In itself, however, this novel result is important, as it is reassuring for a certain number of previously published analyses, in particular correlating dN/dS with life-history traits, in a context where GC content also correlates with life-history traits (Figuet *et al*, 2016; Bolívar *et al*, 2016).

But perhaps more fundamentally: by deriving a codon model where the underlying mutational process and the observed nucleotide frequencies are formally and correctly disentangled, we clarify the problem about 3x4 versus 1x4; which has been a long-standing problem. Many people are still using the 3x4 formalism, in spite of the fact that this has been shown to lead to inaccurate inference. Perhaps one reason is that people don't understand why a model such as 1x4, which does not correctly predict the nucleotide frequencies across positions, might do better than 3x4. In this respect, our work is important, because it gives the first clear explanation of how to correctly formalize this problem. Along those lines, we added a figure in supplementary materials (first figure, panels A-F) showing that the MF can predict the accurate nucleotide composition at first, second and third codon positions, while the MG model fails to predict them.

Finally, and as developed in the discussion, our work is an important first step. Conceptually: it gives interesting directions to make a more robust conceptual junction between mutation-selection and ω-based modeling approaches with promising developments, for example codon models with gBGC.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., & Galtier, N. (2016). Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Molecular Biology and Evolution,* 33(6), 1517–1527.
https://doi.org/10.1093/molbev/msw033

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., & Mugal, C. F. (2019). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20(1), 1–13.
https://doi.org/10.1186/s13059-018-1613-z

As minor points, I'm not sure I understand why the Muse and Gout 1994 model was used for comparison, given that there are more advanced models. They also do not do a very thorough job of at least referring to other relevant works. For example, there are models that attempt to estimate rates of synonymous substitutions, which therefore might make them less agnostic to teasing mutation bias and selection than the MG model.

Effectively, it is indeed true that we did not give a sufficiently clear justification of the use of the MG (1x4) model as a baseline. This choice stems first from previous works (Kosakovsky Pond *et al,* 2005; Rodrigue *et al,* 2008, 2010; Spielman & Wilke, 2015), who all showed that the MG model, with a 1x4 parameterization, is the best performing model (precision and estimator bias) for estimation of the underlying ω. This is why we solely used MG as a reference for comparison.
The manuscript has been edited to clarify these points.

Kosakovsky Pond, S. L., & Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution,* 22(5), 1208–1222.
https://doi.org/10.1093/molbev/msi105

Rodrigue, N., Lartillot, N., & Philippe, H. (2008). Bayesian comparisons of codon substitution models. *Genetics*, 180(3), 1579–1591.
https://doi.org/10.1534/genetics.108.092254

Rodrigue, N., & Philippe, H. (2010). Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends in Genetics,* 26(6), 248–252.
https://doi.org/10.1016/j.tig.2010.04.001

Spielman, S. J., & Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, 32(4), 1097–1108.
https://doi.org/10.1093/molbev/msv003

Finally, there are a few typos along the manuscript so I think it would benefit another round of proofreading. For example, in line 147 (Results section), the second appearance of 2NeP_fix^WS should probably be 2NeP_fix^SW.

We performed another round of proofreading and edited the manuscript.

Reviewer: 2

Latrille and Lartillot introduce a novel codon model framework for teasing apart mutation vs. selection effects in coding sequences. The introduction of the w tensor is a nice addition to the field (yes, this is a pun!), and it seems like a promising approach to identifying effects of mutation vs. selection from sequence data. I recommend with major revisions.

Main comments

1. While I think this paper is a helpful contribution to the field, the paper appears to be somewhat "slapped together." I strongly recommend a couple more rounds of editing to make sure the main points are clearly conveyed. For example...

> ○ Based on the Abstract and a decent amount of the Introduction, I thought this paper was going to be all about gBGC.
>
> ○ I really recommend explicitly stating the problem up front: We want to address identifiability issues when estimating fixation vs. mutation in coding sequences. This appears to be the central point of the paper, but the written manuscript does not explicitly emphasize this point as a central goal that motivates all discussion.

Undoubtedly we should focus on the problem statement. This comment reflects the suggestion of another reviewer. We updated the introduction and abstract to make our points more salient.

2. On page 10 where the authors claim support for their model based on LRT and AIC comparisons with MG94. However, this is potentially problematic, as explained in this paper - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011587. Comparing arbitrarily more complex models with the simplest homogeneous models generally lead to improvements in fit, from this part of that Abstract: "Here, we show that the single rate model of non-synonymous substitution is easily outperformed by a model with multiple non-synonymous rate classes, yet in which amino acid substitution pairs are assigned randomly to these classes. We argue that, since the single rate model is so easy to improve upon, new codon models should not be validated entirely on the basis of improved model fit over this model." Can the authors provide additional justifications for the use of their approach beyond improved fit compared to the simplest other option?

We agree that, in general, AIC and LRT statistics may not always be sufficient to claim support for the use of the richer model. Spielman & Wilke (2015) also argued that the best performing models in terms of precision and estimator bias (for estimating ω) are not necessarily the models selected by AIC or LRT.

Spielman, S. J., & Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, 32(4), 1097–1108. https://doi.org/10.1093/molbev/msv003

However, we think that, in the present case, AIC and LRT are correct and are saying something that is valid. Fundamentally, what the AIC is meant to measure is the balance between the stochastic error under the richer model and the systematic error induced under the simpler model by imposing the same rate for all amino-acid pairs. We agree that, in the absolute, the error on the rate estimates under the richer model may be large, in particular for smaller alignments (see supplementary materials). But on the other hand, the true rates are very different for different amino-acid pairs, and thus imposing the same rate for all of them, as does MG, induces an even larger systematic error.

Justification for the use of our the MF model in the present case is further corroborated by our simulation experiments, in which the MF model gives an accurate estimate of λ. To further examine this point, we have conducted simulations under smaller alignments, and even in that case, the estimated mutation bias under MF model is reasonably accurate, as we now show in the supplementary materials. Finally, On figure 2 we added a line representing the AT/GC predicted at equilibrium by the model, solely for the MF model the predicted AT/GC matches the observed AT/GC of the alignment

We edited our manuscript to cite the suggested paper and tone down the use of AIC and LRT as a claim for support:

The empirical fit to the data between the nested models, using AIC and Likelihood ratio tests (Posada and Buckley, 2004) favor the MF model compared to the MG model (table 1). This suggests that the additional parameters entailed by MF are warranted on these empirical data, at least when compared to MG. It has been argued previously that a higher fit under AIC or LRT may not always mean that the more complex model should be used (Delport et al., 2010). However, our simulations suggest that, in spite of the larger estimation error on the individual rates between all pairs of amino-acids on smaller alignments, the estimate of the mutation bias is always reasonably accurate, even on small alignments (supplementary materials).

In another simulation analysis, it has also been shown that better fitting models could sometimes lead to less accurate inference (Spielman and Wilke, 2015). This point was more specifically made concerning models such as 3x4. We concur with this argument, which is particularly relevant in the present context. The 3x4 model is typically better fitting than the 1x4 model (which is the default considered here through the MG model). Yet, and this is precisely one of the main points of the present work, 3x4 does not represent the correct way to model the processes that are creating the variation in nucleotide frequencies across the 3 coding positions and, for that reason should not be used, in spite of its higher fit. The MF model, on the other hand, gives the correct logical solution to this problem and our simulation experiments confirm that this leads to accurate estimation of the mutation bias. In summary, this is the conjunction of the higher fit observed here on empirical data with the logical arguments and the simulation experiments presented above that together justify the use of the MF model. Based on these justifications, we can thus interpret the estimate of λ_{MF} as reflecting the mutation bias, and the difference between $2N_e P_{fix}^{SW}$ and $2N_e P_{fix}^{WS}$ as suggesting that the fixation biases are different in the two directions also in the case of empirical data.

> ○ In Table 10, two lines are labelled "MG (2NeP{^WS_fix})", which is which?

Sorry for this mistake, the second line has been edited to $2N_e P_{fix}^{SW}$

> ○ In P-value calculations in Table 10, can the authors provide a more explicit basis for df=93 in the main text? This can/should be a brief addition, maybe to the table caption for example.

The full model contains 95 parameters with 75 amino-acid exchangeabilities and 20 amino-acid equilibrium frequencies. However, we have two constrains that reduce the degree of freedom :

- The sum of all 20 amino-acid equilibrium frequencies equals 1
- The sum of all 61 codon frequencies equals 1.

3. I am very concerned that this paper only considers one simulation phylogeny, and the divergence in this tree (and how divergence can be interpreted under HB98 simulations) is not discussed. While I generally do not like to ask for additional analyses, I feel that the simulations here are insufficient. More than 1 topology and more than 1 sequence length seem necessary for evaluating the proposed codon model.

We performed more analyses shown in supplementary materials with different sequence length (498, 996, 2490, 4980 and 9960 codon sites), with different branch lengths (decrease by a facto 2 and increase by a factor 2, 4, 8 to obtain saturation) and also with a different topology (90 mammals instead of 61 primates).
These analyses have shown that the number of sites does not influence the estimator's accuracy for mutational bias (λ), nor for selection pressure (ω).
However, whenever the branch length increases, saturation of sequences (multiple substitutions at the same site or identical substitutions in different sequence) contributes to biased estimators. Both the MG and MF models fail to give an accurate estimator for selection pressure (ω), while mutational bias (λ) is accurately estimated.

> ○ Regarding the simulation procedure, it's further not clear to me that these are true replicates vs. pseudoreplicates. On page 13, authors indicate that site-specific fitness profiles are randomly selected from Bloom 2014. How is this treated across replicates? If, for example, if which Bloom 2014 codon site is sampled at a given codon index differs across replicates, these are not true replicates. It's important to ensure there are true replicates in the simulation study presented here.

To build the fitness profiles, for each codon site z of our simulation, we assign randomly one the 498 fitness profile (sampling with replacement) experimentally determined by deep mutational scanning for 498 codon sites of the nucleoprotein in Influenza Virus strains (as human host). However, the replicates are obtained using the same fitness profiles as input, and the variability come from the stochastic mapping of substitutions along the phylogeny. Altogether, the replicates shown in the manuscript are indeed true replicates.
We adapted the manuscript to make this point clearer:
All simulations presented in the main manuscript are obtained using the same underlying tree topology and branch lengths of 61 primates from Perelman et al. (2011), along with the same 4980 codon sites with amino-acid fitness profiles resampled from experimentally determined profiles in Bloom (2017).

4. I'm pleased to see the GitHub links for reproducibility, but if the authors wish for others to use this method, the code needs more documentation. If nobody can use a new method, then it's not really a new method at all. Please carefully consider how you envision this method being used by the community, and update code repositories to support these uses.

Thank you for these suggestions.
We updated the main *readme.md* to focus solely on the installation procedure of Hyphy and then refers to two newly created *readme.md* for the empirical datasets and simulated datasets.
The *readme.md* in the simulated folder is focused in reproducing the figures shown in the manuscript using the program *SimuEvol*.
The readme.md in the empirical folder describes the procedure to perform analyses on your own datasets, and use as example the datasets shown in the manuscript.
Altogether, users can use the GitHub repository either to reproduce the results or to perform analyses on their own datasets.

Minor comments

1. I would be careful about describing all nucleotide composition and natural sequences as necessarily due to an equilibrium, since non-equilibrium dynamics in nature (if not always models) are to be expected. I'd tone this down to "interaction" or similar.

The term equilibrium was indeed ill-suited to highlight the fact that sequences are not optimized but are the result of a balance between different forces. We updated the abstract and the introduction without referring to equilibrium whenever not necessary.

2. I'm not convinced the mutation process should be blind to coding structure. I entirely agree that the partitioning of three codon positions into different mutational categories is not ideal, but to contend that the biological process of mutation is the same across all regions of a coding sequence strikes me as questionable. For example, consider something like actin, which is highly repetitive. I would imagine various molecular mechanisms in

such genes would affect mutation itself. Can the authors bolster this claim (made in Abstract)?

Admittedly, we would also imagine different molecular mechanisms for which genes would affect mutation itself (repetitions, methylation, correlation between chromatin opening and mutational opportunities, ...). Here, we simply meant that the mutation happening at the DNA level is blind to the genetic code occurring in ribosomes, and that there is no reason for the mutation process to be shared by all the first position in the genes, while at the same time being different than the process at all second and third positions.

The section was updated to:
These modulations across the three coding positions have been accommodated using the so-called 3x4 formalism (Goldman and Yang, 1994; Pond and Muse, 2005a), allowing for different nucleotide rate matrices at the three coding positions. However, this is also problematic. For instance, it has the consequence that synonymous substitutions, say from A to C, occur at different rates at the first and third positions. Yet, although modulations of the mutation process along the sequence cannot be excluded, most of the empirically observed compositional differences between positions are likely the consequence of selection, which is stronger at the first and second than at the third position. In principle, these selective effects should not directly impact synonymous rates. Thus, although the mutational process might be more complex, there is no reason to model it in terms of a 3x4 structure which conflates two levels of mechanisms that are not supposed to play together.

3. Page 2: The discussion of classical codon models is a bit fuzzy on GY vs. MG style (e.g. line 39 claims all classical codon models use a single parameter, which is certainly not true). I recommend some re-writes here, at author's discretion.

The section was edited to clarify and acknowledge the variety of phenomenological codon models. Moreover, this section now clearly states that we use MG as a reference for comparison since Spielman & Wilke (2015) have shown that MG is the best performing model (precision and estimator bias) to estimate the underlying ω.

4. Page 3 Line 88: The phrase "All this suggests" is rather a stretch - reading any amount of text before this paragraph, I never thought "ah, a tensor will solve this!" Please rephrase.

The section was updated to:
In order for a codon model to correctly formalize this subtle interplay between mutation and selection, the parameter responsible for absorbing the net effect of selection (i.e. ω) should not be a scalar, but an array of ω values (i.e. a tensor) unfolding along multiple directions.

5. Page 3 Line 93: what is "and this"? This comment brings up a larger issue I had reading the manuscript - I often had trouble identifying the correct antecedent in many sentence

constructions. A few more rounds of copy editing for grammatical clarity would be helpful. This phrase appears a few times in the manuscript too, and it's very challenging to understand.

We edited several occurrences of ‘this’ in the manuscript that admittedly were quite frequent.

6. Section 2.1 should be "Simulation experiments" (not "simulations")

Absolutely

○ In Section 2.1 line 109 - given the different models floating around in the paper, it would be helpful to be explicit about whether this is a global or per-site nucleotide frequency (presumably global).

The sentence was updated to:
We assume a simple mutation process with a global parameter controlling the mutational bias toward AT.

○ Page 3 line 116 should be "increases" not "increase"

Absolutely

7. In the caption of Figure 1, I'd rephrase "repeats" to "replicates" (also used caption fig 3, and throughout text).

We rephrased five occurrences of "repeats" throughout the manuscript.

8. Please explicitly define Nr vs. Ne. Also, in Figure 1 caption, text refers to Nr as effective population size. Which is it?

N_e refers to the absolute effective population size (number of individuals in a panmictic population). On the other hand, N_r refers to the change in effective population size relative to that obtained in experimentally determined fitness profiles (as in equation 13 of section 4.3 – Selection at the amino-acid level).

9. For consistency, refer to MG94 either as Muse & Gaut or MG. Paper currently goes back and forth.

Muse & Gaut has been kept only for the titles, otherwise it is referred as MG.

10. Please label axes and scale bars in Figure 3. Currently from caption + figure alone, it is not possible to interpret the first two panels.

11. Page 8 line 192, "tensor" not "tenser"

Absolutely

12. Page 9 line 214, "genes" not "gene"

Absolutely

13. Page 9 line 216, "alignments" not "alignment"

Absolutely

14. Page 10 line 249, the phrase "are inherently misspecified" is very loaded considering the many, many ways these models are misspecified. Please be more specific about the misspecification of interest.

The phrase was updated to:

However, current parametric codon models predict that the observed and underlying mutational biases should be equal. For that reason, they are inherently misspecified and are unable to tease apart opposing effects of mutation and selection correctly. As a result, they don't estimate the mutational process accurately.