## Editorial Office Required Revisions

1. TABLE OF CONTENTS GRAPHIC MISSING: Table of Contents (TOC) Graphic is missing from the manuscript: Upload the TOC and save as a "Graphic for Manuscript" file or place it at the end of the manuscript. Please review the guidelines for the TOC graphic at the following link: http://pubsapp.acs.org/paragonplus/submission/toc\_abstract\_graphics\_guidelines.pdf?  
   
2. References: List more than 1 author before truncating with "et al."  
   
3. SUPPORTING INFORMATION: Please add a full header at the top of the page of the Supporting Information file, which includes: Title, Full Author List, and Author affiliations (exactly as they appear in the manuscript, including postal codes).  
   
4. SUPPORTING INFORMATION: Please number all pages in the Supporting Information file in the following format: S1, S2, S3, etc.  
   
5. SUPPORTING INFORMATION PARAGRAPH MISSING: Please provide a brief description of the contents of the supplementary material in non-sentence format. This section should appear directly before the Acknowledgement and Reference sections. The appropriate format is as follows: Supporting Information. Brief statement in non-sentence format listing the contents of the material supplied as Supporting Information.

**All this has been added / fixed.**

## Reviewer: 1

The authors evaluated the performance of machine learning guided design of ribosome binding sites (RBSs).They used Bayesian optimization, via Gaussian processes (GPs) and the upper confidence bound (UCB), to iteratively improve the design of RBS translation initiation rate (TIR). As a proof-of-concept for small-budget engineering, the authors explore ~10% of the full RBS sequence design space. Over four rounds of model-driven design, the authors engineered three RBS sequences with improved TIR over wild-type levels. They also compare this performance to alternative, non-model-driven approaches, which generally did not discover any RBS designs with performance better than wild-type The authors used a weighted kernel with shift (WDS) kernel for the GP.

Major comments

1. The limited dataset size makes it difficult to evaluate the expected performance of the approach. The authors are interested in evaluating the performance of model-driven design with limited experimental budgets. This is an important question, as the field in general is becoming increasingly interested in model-driven approaches to engineering.

But while the research interest is the performance on limited budgets, this is not sufficiently assessed with the data provided. Specifically, the authors have evaluated the empirical performance of a single run of the UCB algorithm. Given that these runs will be stochastic (e.g. measurement noise varies across experiments, resulting in different rankings of RBS designs) - the results of this single run are not representative of the performance on average. This is similarly reflected in the theoretical justifications for the UCB algorithm, which only guarantee expected performance [Srinivas et al., 2009].

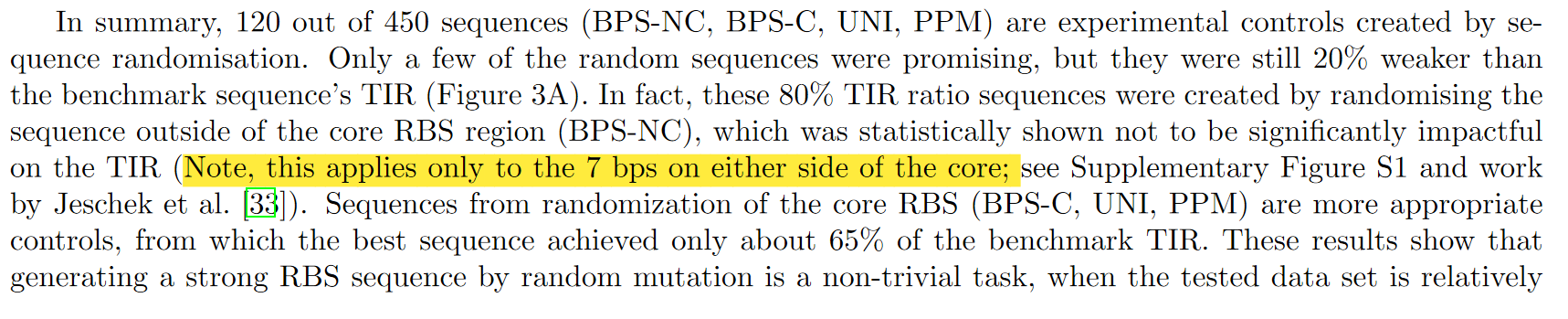
This is in particular important considering only three RBS designs improve over the wild-type sequence. So, it is difficult to know to what degree the algorithm "got lucky" versus reliable performance in future applications. This limited return is also somewhat unexpected compared to previous studies applying BO to bioengineering - where far fewer designs per iteration were necessary and most designs leading to improvement [Romero et al., 2012].

Finally, there are a number of design decisions (e.g. bandit batch size, exploration-exploitation paramter βt) that could influence the outcome of the algorithm. So, it is critical to determine whether these differences are an aspect of the targeted biological part (RBSs) or some other issue with the algorithm.

To address this concern, I would recommend one or both of the following revisions (with the first being the most ideal):

1. Measure the full 4096 RBS sequences for TIR. This seems within the capabilities of modern high-throughput measurement technology, and would provide complete information about the performance of potential RBS designs. Then, the authors could run simulations of their algorithm with different randomized training sets to evaluate the performance of the approach on average. Expected performance (in terms of regret) could then be evaluated. This would also provide the community with a valuable resource in the form of a benchmark dataset for testing future model-driven design algorithms.

**After careful consideration, we decided not to measure the full set of 4096 sequences due to the budget and time constraints. While the reviewer is right to point out the stochastic nature of the Bandit approaches, we would like to reiterate that a quarter of our measured sequences are actually experimental controls to show that an undirected stochastic search would be unsuccessful (the random set). However, we have decided to further investigate the efficacy of our approach using the second recommendation (please see below).**

****

2. If time and resources are not available for further experimentation, I would recommend similar evaluation of the algorithm under simulations with alternative datasets. These could come from datasets published on similar design targets (e.g. https://academic.oup.com/nar/article/48/

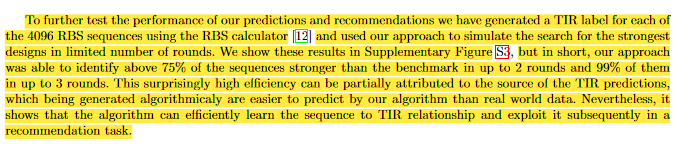
18/10602/5911751) or generating simulated designs from existing RBS calculators (https://

www.frontiersin.org/articles/10.3389/fbioe.2014.00001/full). While these would have potential disadvantages compared to (1), e.g. representing potentially different design goals for other experimental datasets or incomplete modeling of RBS dynamics for existing calculators, this would at least provide evidence for the actual performance of the algorithm when repeated

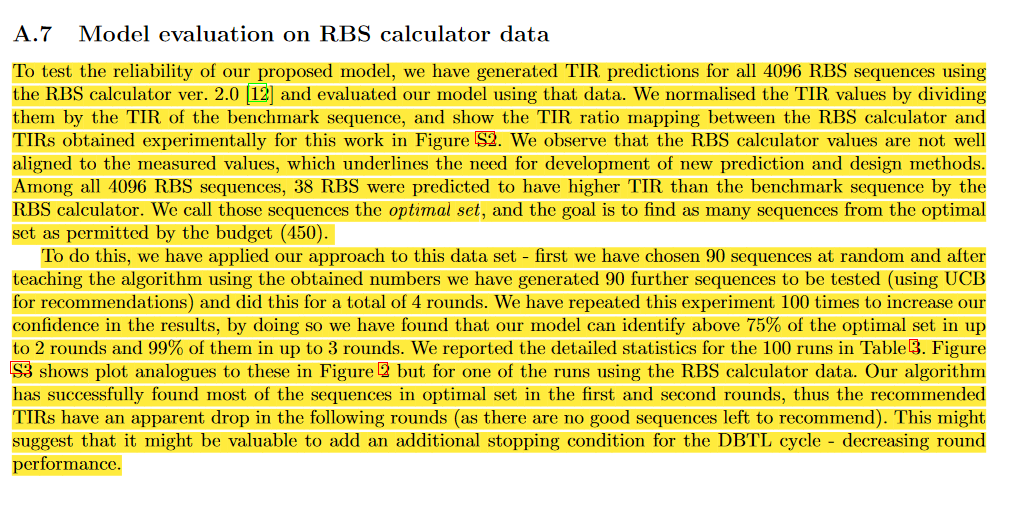
multiple times.

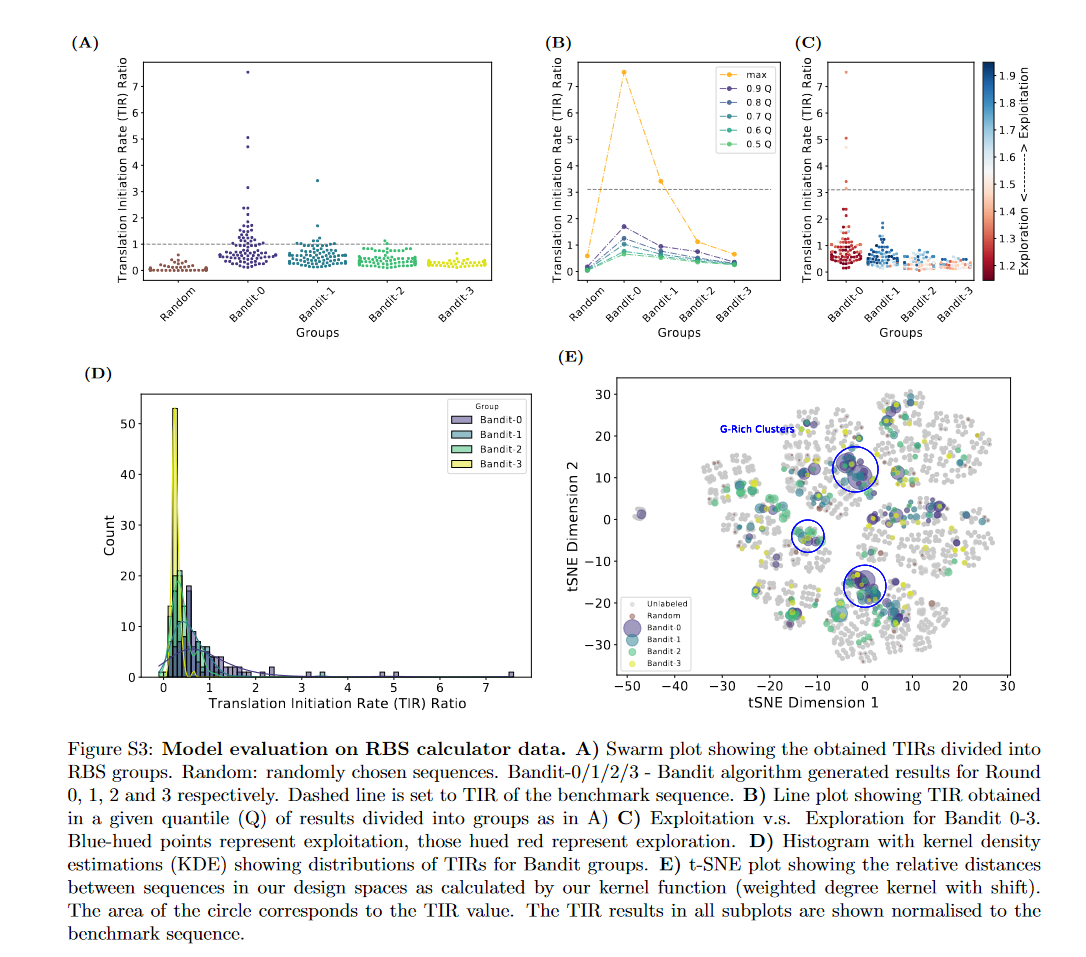
**Thank you for this suggestion. We have carefully considered the two suggested possibilities and found that:**

1. **Evaluating a bandits - based algorithm on data generated using a different policy (recommendation rule) can be biased and influenced by different design goals, as stated by the Reviewer. The dataset published together with Ding et al. 2020 provides only data for a classification task (not all data is available). We were able to receive more data from the corresponding author that allowed us to run a regression task using our approach, but because of the very small scale of the received data (400 sequences from 4^10 design space), it was impossible to properly evaluate our algorithm using this data set. At the same time, we were unable to identify other, more suitable published datasets.**
2. **The second approach (RBS calculator) is suitable for additional evaluation of our approach. We have now generated TIR values for all of our sequences using the Salis RBS Calculator (ver. 2.0) and run 100 synthetic experiments using this data that simulated our ML approach aimed at finding the strongest sequences from the set. In short, these experiments confirmed high efficacy of the algorithm in finding the strongest sequences. Our model can identify (on average from the 100 experiments) above 75% of the RBS sequences beyond benchmark up to 2 rounds and 99% up to 3 rounds. We have described the results of this in the Discussion section:**



**We have also generated additional plots describing these results in the supplementary document:**

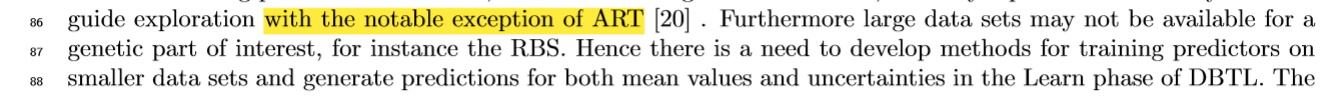




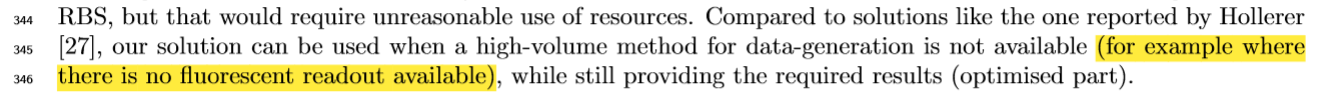
2. The proposed experimental budget is not small compared to the search space

Related to the previous comment, is the fact that the percentage of the variant sequence space (~10%) is not particularly small when compared to current experimental capabilities. It is generally more reasonable to expect much less coverage of the variant search space than the authors observe here. For example, the full variant search space when considering the complete 20bp region of the RBS would be on the order of 420 ≈1012. Current massive scale sequencing experiments are capable of measuring 106 — 107 total variants, so the upper bound on experimental coverage is much less than 10% in most cases. So, the discussion of low-budget engineering is not just in the context of total number of measurements - but also the relative percentage of search space observed. Given the recovery of only three variants improving over the reference, it’s unclear whether this approach would yield positive results in cases where a much smaller percentage of the variant search space can be sampled.

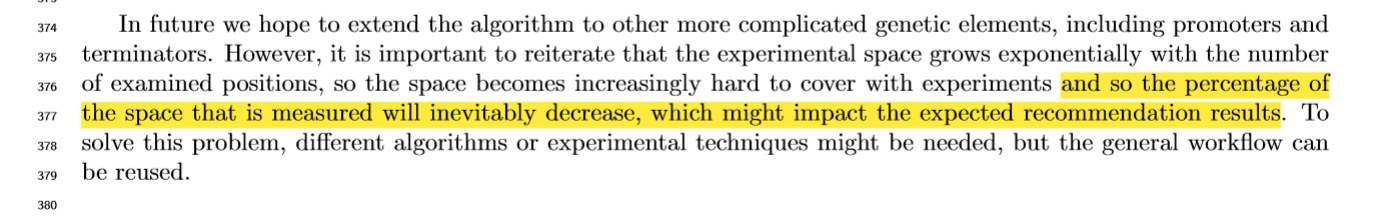
**The Reviewer is correct in their assessment – the higher the percentage of the experimental space that can be measured the better the expected predictions (generally speaking). It is also absolutely correct that there are now methods available that allow for 106 – 107 variants to be tested. However, throughout our manuscript we do make a point that our method, “as is”, is applicable to smaller studies, where the coverage of measured results is higher than what can be expected when the other methods are used. For example:**



**Additionally, the previously mentioned higher-volume methods are, for the most part, available only when there is a fluorescent signal that can be connected to the tested quality, for example a GFP fusion with a biosensor. On the other hand, our method can be used when a more advanced analytical methods, like HPLC analysis, must be used. We have pointed out this fact now in discussion:**



**Finally, in the final paragraph of the Discussion we do discuss the points raised by the Reviewer, where we now added a sentence pointing out the percentage problem to the reader:**



3. The prose of the manuscript is not of publication quality and requires improvement.

**Thanks for the suggestions. We have fixed all points mentioned below. We have additionally carefully updated many parts of the manuscript.**

There are grammatical mistakes throughout the manuscript that must be addressed. Some examples are (with issue highlighted in red):

• "In each round, we select some data points for exploration of the areas in which we have few or none tested data points." (lines 249-250)

**Fixed.**

• "Figure 4B shows the edit distance [that] is required for positive change. . . " (line 269)

**Fixed.**

• "One naive approach is to However, this approach may. . . " (line 509)

**This unfinished sentence has now been removed.**

There are also many imprecise or incorrect uses of statistical terminology. Some examples:

• ". . . generate predictions for both mean values and uncertainties. . . " (line 88) and ". . . we used

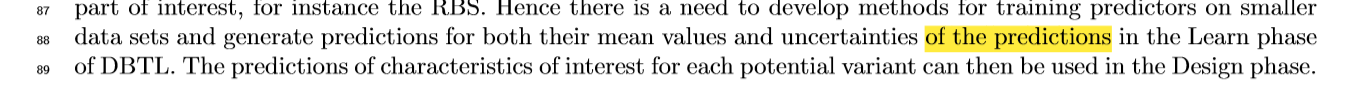
Gaussian process regression to predict the TIR mean and standard deviation" (line 293)

– It would be more appropriate to describe this as prediction of TIR by the GP, with the posterior

predictive distribution having a mean and standard deviation. Under the current phrasing, there

is ambiguity to whether the authors refer to the posterior of the GP or the TIR measurements.

**We agree with the Reviewer, both sentences have now been made more precise:**



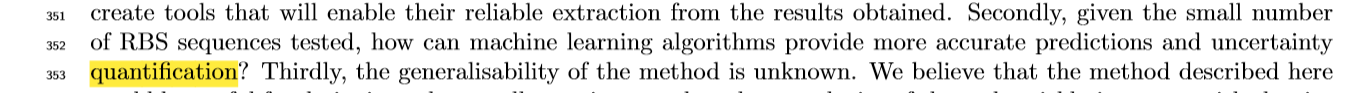
**and**



• ". . . how can machine learning algorithms provide more accurate predictions and uncertainty measurements?" (lines 350-351)

– Uncertainty in this context is not measured, but quantified.

**Another great point, this has been fixed in this revision:**

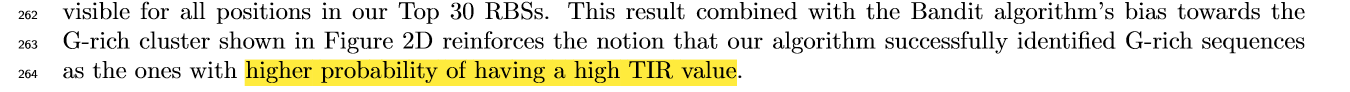


• ". . . our algorithm successfully identified G-rich sequences as the ones with high TIR probability"

(Lines 262-263)

– What is "high TIR probability"? Instead, should be "enrichment for higher TIR" or similar.

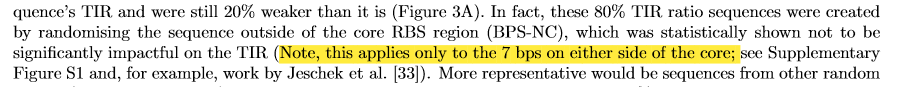
**This has been fixed as well:**



Minor comments

1. The authors claim that regions outside of the core RBS region do not impact TIR (lines 196-197), but according to figure S2 these mutations decrease TIR relative to the background (dotted line in Figure S2). Their statistical test (Welch’s t-test) only compares the effect on TIR between core and non-core regions. But this is not the same as testing the effect of non-core regions compared to the background. From the figure, it seems quite clear that the non-core regions do impact TIR (although not as much as core-regions).

**Thank you for making this point. The authors agree that they can’t make claims about statistical significance of changes made outside the 20bp region of the RBS. In the manuscript we do not claim that the changes made in the 14bp non-core region do not impact TIR, we claim that they are not statistically significant (in the terms of the Welch’s test that we believe is appropriate in these circumstances). To clarify that we mean only the 20 bp region when we talk about the statistical significance in lines 196-197, we have modified them as follows:**



2. The tSNE in Figure 2e is difficult to comprehend, as the distances contributing to the plot are from the WDS kernel and are therefore not easily understood. An alternative would be to plot the hamming graph (e.g. connecting variant nodes when they differ by one-mutation).

**We have chosen to show t-SNE plot generated based on the WDS distances, because these are the distances that are used by our algorithms to process the sequences and by extension are a better representation of how the algorithms “see” the space.**

**However, we do agree that t-SNE plot based on Hamming distances might be useful for readers and so we have added it in the supplementary materials:** **Chart, scatter chart

Description automatically generated**

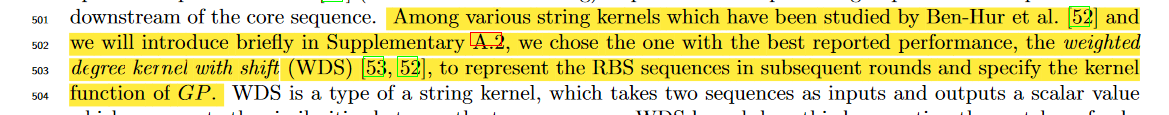
Also, this figure also does not show that the "RBSs recommended by Bandit groups have covered the majority of the design space" (lines 225—226). Instead, it shows an approximate representation (via the kernel) of the variants provided to the figure. Since this only included variants that were measured during experimentation, this will not show a complete representation of the sequence space. If unmeasured sequences were included in the plot as well, I suspect that the embedding would show certain regions of the sequence space are unexplored.

**We have shown the unlabelled RBS sequences in light grey in Fig 2(E) which the reviewer might have missed. We have tested this colour on a number of displays and they show up correctly. We reiterate that the plot contains all 4096 sequences.**

3. The Gaussian process model makes poor predictions for large TIR changes. This can be seen in Figure 3, where the true TIR is poorly predicted for changes of large magnitude (e.g. far from zero). This generally indicates that GP has insufficient capability to predict TIR for RBS sequences "far" from the training data (as quantified by the kernel). While this may be universally true for models of this data, alternative approaches (e.g. different GP kernels), should be tested.

**The poor performance of the algorithm prediction when the test data is “far” from the training set is a general challenge for all machine learning methods. This is due to the inability of the GP (and also other machine learning predictors) to predict values that it has not “seen” before. In other words, the algorithm can’t predict something “unexpected”. However, even if the values are not predicted with high precision, the algorithm is successful in predicting the ranking of the sequences, which is ultimately our goal.**

**Regarding the choice of the kernel - we chose a kernel that based on the literature should perform best in this exercise. We have made it clear in section 4.2.2:**



4. The order of the results sections is non-intuitive. Placing the section on learning (2.3) after design

makes the manuscript difficult to follow because the bandit selections depend on the GP model - but

it is not introduced until the following section.

**This is actually a very interesting point. The authors have tried both approaches (Learn first and Design first) when preparing the manuscript and have decided on the current order for a number of reasons. We decided to start with Design results, since it allows us to:**

1. **Explain to the reader how and what type of data we have generated**
2. **Focus on the most important results (performance of the recommendations) first, which allows for better flow of the article**
3. **Introduce the reader to the concept of the Multi-armed algorithm first, which we believe to be the more important in our implementation.**

**However, to further address Reviewer’s comment, we wanted to point out that in the Methods section we introduce Learn first (in 4.2.2) and Design second (in 4.2.3), which we think is the better flow for the Methods section.**

5. I do not understand the argument of the paragraph starting on line 238. There are many topics introduced: small versus large-scale modeling, evaluating bandit selections, ranking versus prediction. But, I cannot follow how these details contribute to the results. It appears the authors wish to justify the use of Spearman rank correlation in evaluating the predictive accuracy of the GP model. The references used to justify this argument (38 and 39) do not appear to come from the literature on active learning/Bayesian optimization. If the purpose is just to explain how the models predictions were evaluated, this should be made more clear.

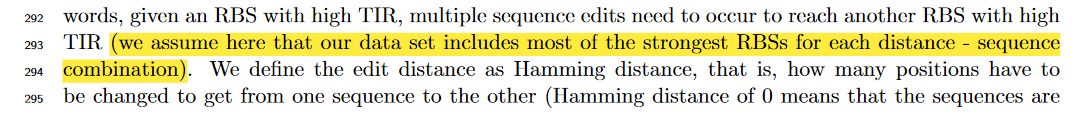
**The Reviewer is correct, this paragraph is mostly to explain why the evaluation has been done via Spearman's rank correlation coefficient and not Pearson correlation coefficient (r) or Coefficient of determination (R2). In life sciences it is a standard practice to use the latter two and so we felt it is important to explain why it is not suitable to do so in this case. We also think it’s important to explain to the reader why we are more interested in getting the ranks of RBSs right, rather than their specific numerical values. We have modified the paragraph in question to make it flow better based on these comments:**

Text

Description automatically generated

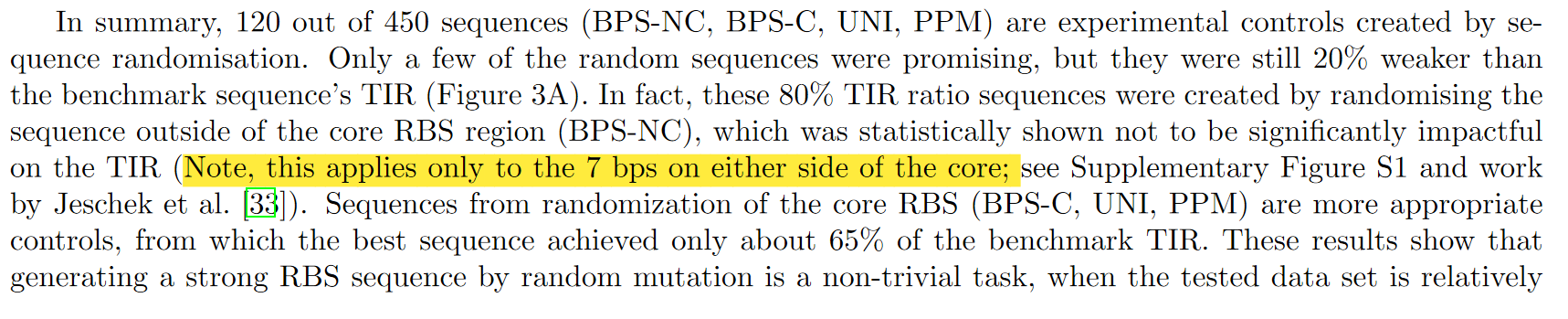
6. As I understand it, Fig 4b quantifies the maximum TIR for and RBS of a given hamming distance for each sequence on the y-axis. But this would only be empirical, as the full RBS sequence space is not measured. So, the maximum TIR at a given distance may be higher, but has not been measured in the current dataset. If this is the case, then the authors must better explain and justify this information - as it stands now this does not provide the information suggested (the hamming distance to a TIR of specified value).

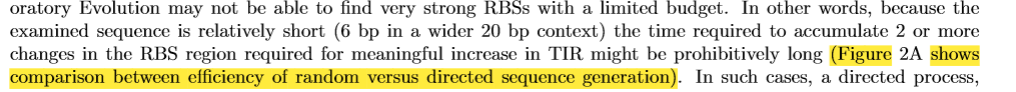
**The Reviewer makes an excellent point here. We agree that since we have not seen the whole experimental space our original text needed to be modified to include a disclaimer saying that we assume that we have found most of the relevant sequence – Hamming distance combinations:**



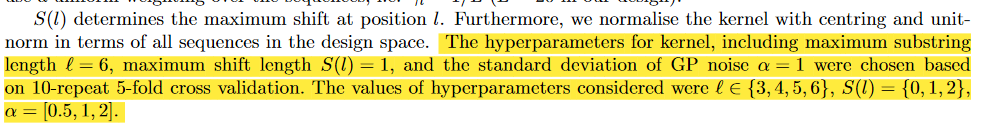
7. I do not agree with the argument on lines 278-281 that the search space is prohibitively difficult to explore with alternative methods. If the argument is that the full 20bp region cannot be as easily explored, this is true - but the analysis of this work has not explored that region either. So, it’s not a fair comparison to say UCB has outperformed randomized search where it’s fairly plausible that a library randomized only on the 6bp core region would be successful in identifying the same or similar high-performing variants as UCB.

**In our figure 2A we show how random generation of sequences performs versus UCB. The UCB guided approach is visibly (and measurably) more efficient in finding strong RBS sequences than the random approaches. As in the response to Reviewer 1, we now updated the paragraph of results to stress the fact that a quarter of our measured sequences are a library of randomized controls.**

**Given this, the authors still argue that the UCB guided approach is more desirable, even for the fairly small region of 6 bps. We have directed the readers to that figure in the revision to make this point clearer:**

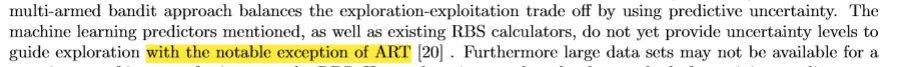


8. The authors perform cross-validation to select the hyperparameters of the WDS kernel (e.g. `, S(`)). But, there is no information about what values of the hyperparameters were considered or what values were ultimately chosen for the model.

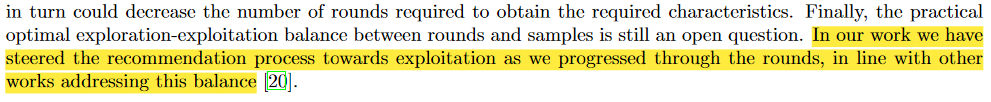
**Thank you for pointing this omission out. We have now added the details in Section 4.2.2 as follows:** 

Reviewer: 2  
  
Recommendation: Publish after minor revisions noted.  
  
Comments:  
The authors provide an approach for ML-guided experimental design of bacterial RBSs, with the aim of improving protein expression. They used Gaussian Process Regression to map genotype (core RBS sequence) to phenotype (translation initiation rate, TIR, measured through GFP fluorescence) and batch Upper Confidence Bound-based approach for generating recommendations of sequence designs to be tested in vivo. Four DBTL cycles, which integrated the ML approach with automation and high- throughput data generation, were performed, producing a data set covering about 10% of the whole design space. The final DBTL cycle achieved up to 34% higher TIR than the benchmark sequence.   
This is an important demonstration of the potential that machine learning, combined with automation, has as a tool for guiding the RBS optimization. The paper is of high quality, technically solid, well written and clear. The authors do a good job of discussing limitations and possible extensions. I believe that the contribution of this work is valuable and is worth being published in ACS Synthetic Biology, after addressing a few minor issues.  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
- Page 3 line 8 - “The machine learning predictors mentioned, as well as existing RBS calculators, do not yet provide uncertainty levels to guide exploration.”  
  
One of the ML method the authors cite here [20] does provide uncertainty associated to the predictions and use it to guide the design of experiments in the next cycle. The authors should also discuss how their exploration-exploitation approach relates to the one used by Radivojevic et al.

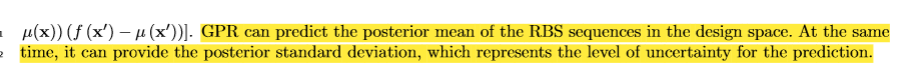
**We thank the Reviewer for this very important point. We have updated the line to:**



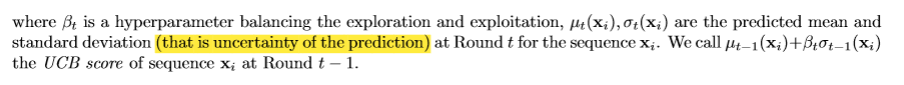
**Similarly, we have added the following sentence in our Discussion to indicate that our approach to setting the exploration – exploitation balance has been similar to the one in ART:**

  
  
- Page 13 line 55 - “GPR can predict both the posterior mean and standard deviation of the RBS sequences in the design space”  
  
Although in GPR the predicted standard deviation depends only on features, this shouldn’t be confused with the fact that standard deviation that is being calculated is in fact a measure of uncertainty in predictions of the label variables.

**The authors agree that this is an important distinction to make. We have reworked this and the following sentence to make this point clearer:**

  
  
- Page 14 line 47 - “are the predicted mean and standard deviation at Round t for the sequence xi”   
  
Likewise, here the authors should correctly state that the mean and standard deviation are for TIR values associated to an RBS sequence.

**Again, the authors agree and made this point clear in the revision:**

  
  
- Page 14 line 53 - “One naive approach is to” (unfinished sentence)

**Thank you, this has been fixed and the sentence has been removed from the revision.**