**Summaries**

Reviewer: 1   
Summary: The author suggest a combination of semi-automated experimentation and machine learning to optimize RBS sequences in an iterative fashion (design-build-test-learn cycles).   
  
Reviewer: 2   
Summary: 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, with the final one achieving up to 34% higher TIR than the benchmark sequence.   
  
Reviewer: 3   
Summary: The manuscript from Zhang and colleagues describes the application of active learning in order to predict and recommend best RBS sequences for an iterative DBTL cycle. The authors showed that 4 experimental iterations involving a 10% of the full design space was enough in order to identify RBS combinations with high performance.

**Strengths**  
  
Reviewer: 1   
Strengths: The work is a combination of previously published tools and methods. It is difficult to point to unique assets and/or new findings/tools/methods.   
  
Reviewer: 2   
Strengths: This is an important demonstration of the potential that machine learning has as a tool for designing RBSs. The paper is of high quality, technically solid, well written and mostly clear. I believe that the contribution of this work is valuable and is worth being published in Synthetic Biology.   
  
Reviewer: 3   
Strengths: Reading this study is really enjoyable and shows both a machine learning and synthetic biology mind.   
  
Results are positive and convincing, i.e., better RBS were successfully designed through the described approach.   
  
The authors made use of several iterations of a full DBTL cycle in a biofoundry, involving LEARN stages.   
  
**Major Comments**

Reviewer: 1   
More review: - In general, the descriptions in the manuscript lack critical detail (both for experimental and computational works), which makes it impossible to reconstruct or repeat what the authors precisely did. The figure captions are not sufficient to understand the content of the figure (see for instance Fig. S7). The methods section is very much underexplained given the complexity of applied methods (automation etc.).   
  
- It is unclear why the authors choose very strong RBSs as a design goal. This challenge has been amply solved in prior studies and in most optimization cases (e.g. in metabolic pathway optimization) the goal is certainly not to have the strongest RBSs (here fine tuning is key to avoid metabolic burden). Also the claim that the authors have found RBSs “exceeding the currently known strong RBSs” (quote from abstract) is unjustified and also meaningless to a large degree. RBSs are know to be highly dependent on the context (CDS and upstream 5’-UTR). The authors have at most found RBSs that are very strong in their specific context. A comparison to other studies is questionable, there are simply no strong standard RBSs. A similar argument may be made for the claim that the authors have identified “an extensive and reliable library of novel RBSs“ (p.2, line 42-43): It is unclear what the authors mean by “extensive” and “reliable” here, and what the “novelty” of the RBS is. Designing novel RBS is by no means a challenge or an achievement given the extremely large sequence space (any RBS that is designed at random has an extremely high chance to never have been tested before).   
  
- the “benchmark RBS” that is used/treated as a standard for very strong RBSs is found also for instance in the pET plasmid series. This RBS may in many cases be quite strong, but its use as a benchmark for very strong RBSs is questionable due to context dependence. It is by no means a surprise or a significant achievement that the authors find stronger sequences than this one amongst 450 tested sequences.   
  
-  Testing “just” 450 different variants to obtain strong RBSs seems like a lot of effort, especially in view of the fact that there are many predictions algorithms around for RBSs that can be used to design RBSs with quite high confidence. If that goal was to simply increase/maximize the expression level, one could also simply swap the promoter from pLac to pT7, which is much stronger, or by changing the copy number of the used plasmid (unfortunately there is no reference for the plasmid “pBbB6c-GFP” the authors used in this study, but it seems to be a low copy plasmid originally from the Keasling lab).   
  
- the two sequences that are stronger than the “benchmark” are only measured in a single replicate; given typical standard deviations in such experiments (compare also the experimental error for the “benchmark” in SI Fig. S2), it is at least questionable whether the small TIR increases for these two sequences (claimed 34% and 15%) are statistically solid.   
  
- In practical terms, it would have been more helpful to focus on how well the generated models predict the entire range of RBS strengths (from weak to strong). The R^2 values reached in this study are significantly worse than in comparable studies which reach 0.8 and beyond (as the authors correctly mention).   
  
- directly related prior art is not appropriately introduced: for instance in Ref25 a combination of high-throughput data generation and machine learning is introduced as a means to predict RBS behavior with high accuracy; similarly, Salis and coworkers have used ML to improve a biophysical model for the prediction RBS strength (this latter reference is entirely missing: https://pubs.acs.org/doi/abs/10.1021/acssynbio.0c00394); given the very high degree of similarity to this study, it seems surprising that the authors forgot (or chose not to?) explicitely introduce these studies as prior art. Importantly, these studies reach much higher prediction accuracies than the underlying work.   
  
- It is unlikely that the designed RBSs will behave in a similar fashion in combination with coding sequences other than the used GFP gene or in other contexts. Therefore, any claim that the identified RBSs could serve as some kind of new standardized parts or would have utility in future studies is invalid.   
  
Reviewer: 2   
More review: N/A   
  
Reviewer: 3   
More review: A major concern is that the authors have not described how they would deal with a multi-gene pathway with multiple RBS. Is this something doable within their workflow? What are the challenges for multigene pathways?   
  
Is this study really a ML-guided DBTL cycle? Experimental design of RBS might be seen only as one step between Design and Build. The impact of RBS on the final outcome is often only relative, as other factors might have a higher influence on the final performance of strains, such as those in the fermentation process.   
  
LEARN: I don't see the point of showing R2 up to 0.546, this is far below reasonable values.   
I think that the valuable result is the one in Fig 3A or 3B, where high TIRs were achieved through different iterations.   
  
LEARN: Not clear why Hamming distance might impact the budget. As the authors are using Machine Learning, predicted sequences that are Hamming distant might be possible, not necessarily just those of single edit.   
  
**Minor Comments**

Reviewer: 1   
without review: Page 2:   
- line 15: DBTL (Not DTBL)   
- line 23: unclear meaning of “perceived prediction error”   
- line 30: spell check “Escherichia coli”   
- line 40: The argument that the workflow for RBSs can be later translated to more complicated systems is speculative and likely not true. For instance, promoters or proteins are governed by entirely different mechanisms/principles. Whether this studies has useful implications for these systems remains to be demonstrated. The authors should relativize this statement to that end.   
  
Page 7:   
- the workflow description is very generic/inconcrete   
- k is mentioned in the Figure caption, but not used in the figure (Fig. 2)   
- line 54: the 20 bp sequence is only 19 bp long (T missing); the authors should specify what they mean by “core RBS”       
  
Page 8:   
- line 30: what is the “original, consensus sequence”?   
- the authors should explain in more detail what the “Bandit-0” dataset is, how it was generated, and why they selected this   
-   
  
Page 9:   
- the authors should specify the SD sequence they use. Is that the core SD motif AGGAGG put into the context of the “benchmark”?   
  
Page 13:   
- Lines 17 -19: The authors should specify how their approach compares favorably to Ref25. In this work, only one round is used in a very simple experimental setup, which none the less leads to about 1000-fold more datapoints.  Here, the authors use multiple rounds and complicated lab automation to generate only a small fraction of the data. The prediction accuracy reached in Ref25 is R^2 greater than 0.9 whereas the authors here reach only 0.27. Therefore, it is unjustified to claim that this work compares favourably to Ref25. Similar arguments may be made for other works in the context of RBSs.   
  
Reviewer: 2   
without review: I have some comments/questions (see attached file) which, if addressed, I believe would improve the clarity of the manuscript:

1. It is not completely clear from the beginning what do the authors refer to with the term Bandit and what with Round. It might be helpful to explain it somewhere in the beginning of Section2.2. It would also be advisable to clarify that Round refers to DBTL round and use either “round” or “Round” throughout the paper.

2. Section 2.2.1 – Please clarify how TIR is calculated, in particular the derivative part. Anexpression in a formula may be helpful.

3. Section 2.2.2 – The authors introduce their GPR approach and define the design space of 20-base RBS sequences. This may be misleading, as it turns out the design space for the algorithm is consisting of 6-base long sequences, which is explained at the end of Section 3.1. Only the BPS-NC, BPS-C, UNI, PPM methods consider 20-bp sequences, not the GPR approach that is employed for generating recommendations.

4. Page 4, line 32 – The assumption on the unknown mean and variance is in contrast to the one given in Supplementary, where a zero mean is assumed. It also turns out that α is the noise standard deviation, mentioned at the end of the section. Please clarify. I’d also suggest to clearly define yi (observations, labels).

5. Page 4, line 38 – The posterior mean and standard deviation, of what? Please say explicitly.

6. Page 5, last paragraph – I suggest to be more strict in using the term observation. For example,it is used both in “observed points (i.e. features)” and “true label, i.e. observation”.

7. Page 5, last paragraph – Given the batch approach for generating recommendations, it would beuseful to clarify whether the training should be, or not, performed for each point in the batch,i.e. 90 times, so to calculate predicted standard deviation and update the UCB score.

8. Figure 1 – “...mean of±1.96 standard deviation”. If “of” is not a typo here it is confusing.

9. I strongly suggest to increase font size in all figures’ axes, titles, labels.

10. Section 3.1, Figure 2 – Please explain TIR(n)=TIR(n-1) in the caption or make a reference in the text. Also, in caption n, k –>n,k.

11. Section 3.1, last paragraph – The authors chose to modify the 6-bp core of the sequence with their approach, motivated by larger impact on TIR when compared to single nucleotide changes in the non-core region. However, although the impact is larger, in all cases it is towards lower values of TIR. Having in mind that the objective of the study is to increase protein expression, the question arises of why they decided not to consider non-core modifications.

12. Figure 3 caption, line 37 – I suggest to write in full “position probability matrix” instead of the second abbreviation.

13. Figure 3 caption, line 46 – Two consecutive “the”.

14. Figure 3C – What exactly different exploration/exploitation colors for points within a round mean? This is not clear as exploration-exploitation parameterβis fixed for each round. For example, how is the color of a point, for which both predicted mean and standard deviation are high, determined?1

15. Page 10, line 15 – Please give more details on “our embedding function”.

16. Page 10, line 26 – It is not clear what is the difference between Fig. 4 and Fig. S4 – the titles are the same, except of metrics’ values.

17. Page 11, lines 43-44 – I suggest to make the last sentence a bit more clear.

18. Please introduce the STD and GFP abbreviations when mentioned for the first time.

Supplementary

1. Page 18 – Please define explicitly X,y.

2. After Eq. (4) – typo in covariance.

3. “...prior on the...” –> “covariance of the prior on the...”

4. The part “diagonal matrixα2Ion the kernel matrix” is not clear, Eq. (5) suggests it is a sum.

5. I suggest to write explicitly the expressions for the predictive distribution fory∗.

6. Eq. (8) – Seems like the left hand side should have x,x′ instead of X,X′. Please check.

7. It is not clear what exactly is φ specl(x).

8. Seems liked should be l just before Eq. (9) in Φ Specd (X). Please check.

9. Section A.2.1 – It is not clear what actually is the φ function. The text suggests it is ad-dimensional function – what isd?

10. It is not clear how Eq. (17) follows from Eq. (16).

11. Unit norm definition – Please check the subscripts m,d.

12. Figure S1:– What exactly are the rolling averages? Why they don’t start from the same position for all bases?– There is a self-reference to Fig. S1 within the caption.– What is “set X of sequences”?– It would be useful to mention in the caption that these are single nucleotide changes.

13. Figure S2 – “Dash line is...” seems copied from another caption.

14. Figure S3 – It is not clear what are the projected values  
  
Reviewer: 3   
without review: Figure 3A. I am not sure if random selections should always led to such low TIRs compared with the benchmark because of the low dimension of the design space (4096). Have the authors tried to statistically validate the results? Comparison with a random sampling, etc.   
  
Figure 3C. Exploitation results are significantly higher than exploration results. I would expect exploration results higher but perhaps it makes sense. Could the authors comment on this?   
  
I believe that t-SNE is not actually defined.