**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.*

**Thank you to the reviewers and the editor for recognising the benefits of our integrated approach for using machine learning to design ribosome binding sites, that produced TIR that is 34% higher than benchmark with only 4 experimental iterations.**

**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.

**First of all, we wanted to thank the Reviewer 1 for their work on this review, we appreciate the reviewing effort and hope to address all your concerns.**

**Our approaches’ uniqueness is not in each individual method, but in the cross disciplinary integration of methods into a holistic workflow. We agree with Reviewer 1 that pieces of our workflow, namely Weighted Degree Kernel with shift, Gaussian Process Regression, Multi-armed Bandits and high-throughput DBTL cycling have been published as separate methods before. However, we are the first to show how they can be combined into an efficient RBS design tool and demonstrate that it works well together. This work fits into the rising trend of similar works showing use of machine learning based methods in synthetic biology, e.g. compare with works by Jervis *et al.*** [1]**, Costello & Garcia Martin** [2] **and Hollerer *et al.*** [3]***.* Compared to the cited works, our unique proposition is showing how using even relatively small dataset effective RBS designs can be rationally generated using machine learning and automated DBTL cycles.**  
  
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.

**Thank you, we appreciate your support and your help with improving this manuscript.**  
  
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.

**Thank you, we appreciate your support and your help with improving this manuscript.**  
  
**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.).

**We have added numerous additional details to the methods section both in the main paper and in the supplementary materials.** **We believe that the amount of details present in the current revision should allow for reliable reproduction of our methods and the amount of detail is in line with other similar publications** [1]–[4]. **We have also expanded the supplementary figure captions to explain their contents better.**

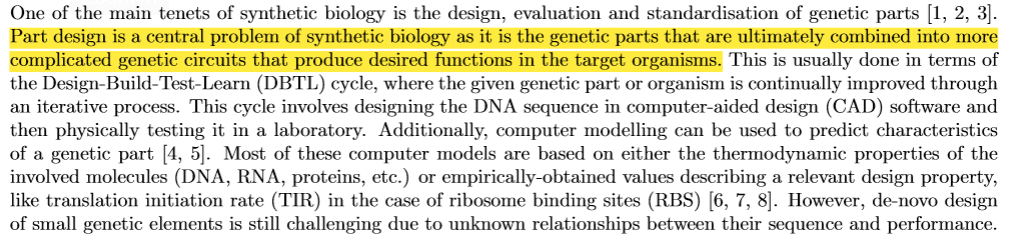
**We added the following details to the Laboratory Methods section:**

* **Volume of reverse primer used in the PCRs**
* **Mass of plasmid template DNA used in the PCRs**
* **Details of liquid handling for the preparation of the PCR mastermix**
* **Details of liquid handling for the preparation of ZAG electrophoresis samples**
* **Relative amounts of DNA parts in assembly reactions**
* **Details of liquid handling for the preparation of DNA assembly reactions**
* **Reworded the sentence describing DpnI digestion**
* **Details of liquid handling for E. coli transformation**
* **Details of liquid handling for the preparation of glycerol stocks**
* **Number of replicates for test strain cultures**
* **Details of liquid handling used to inoculate test strain cultures**
* **Reworded the sentences describing IPTG induction of test strain cultures**
* **Reworded the sentences describing the microplate spectrophotometry method**

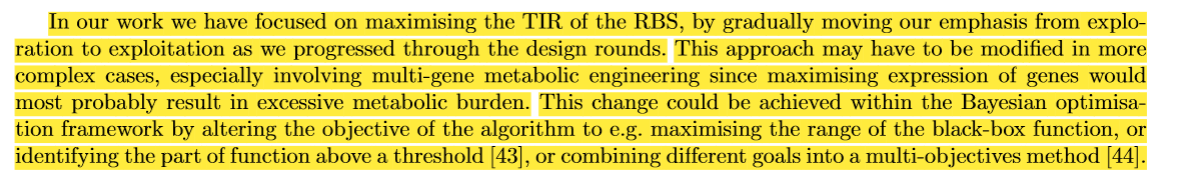
- It is unclear why the authors choose very strong RBSs as a design goal. This challenge has been amply solved in prior studies and in most optimization cases (e.g. in metabolic pathway optimization) the goal is certainly not to have the strongest RBSs (here fine tuning is key to avoid metabolic burden). 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).

**This comment confounds several different concerns. Due to its complex nature, we have tried to break down our answer into points addressing specific questions:**

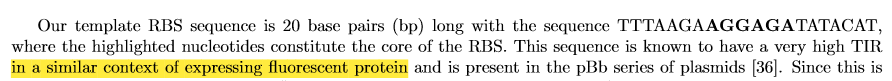
**a) “**It is unclear why the authors choose very strong RBSs as a design goal **“ - We have chosen maximising the TIR as our design goal to showcase how the Bandits algorithm can be used to address exploitation and exploration balance in a genetic part optimisation task, in our case that was the TIR of the RBS. We have extensively discussed why part design is an important problem in the first paragraph of the manuscript, where we now added a sentence explaining the need for part design:**



**We agree with the sentiment that optimisation will not usually mean maximisation in more complex metabolic engineering uses, but it is an appropriate function for a relatively simple case like ours, where the focus is put on the part itself (compare, for example, work by Jervis *et al.*** [1] **with that of Hollerer *et al.*** [3]**. The former focuses on RBS use in a wider context, while the latter focuses on the part itself, similar to our work). Additionally, as mentioned throughout the text, our algorithm can be used to find a range of TIRs, we are reiterating this point now with the new paragraph in the discussion section:**



**b) “**The authors have at most found RBSs that are very strong in their specific context.” - **We agree that the TIR of an RBS will be context dependent. In this study we strived to create easily reproducible conditions, which have been commonly used to study and characterizes number of types of genetic elements, including, but not limited to, as part of the iGEM competition (see the recent work by Beal *et al.*** [5]***)*. In our text, we do not suggest a standard, high strength RBS exists, we merely say that: “This [our benchmark RBS] sequence is known to have a very high TIR and is present in the pBb series of plasmids”. We have modified the beginning of the last paragraph in Section 3.1 to:**



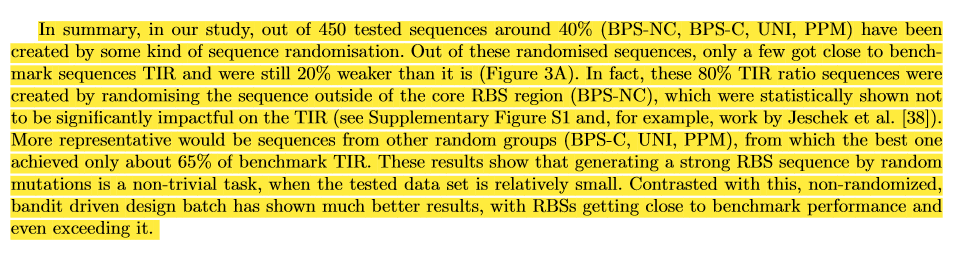
**c) “**A similar argument may be made for the claim that the authors have identified “an extensive and reliable library of novel RBSs“ (p.2, line 42-43): It is unclear what the authors mean by “extensive” and “reliable” here, and what the “novelty” of the RBS is” - **There is currently a very small selection of commonly used RBS libraries, including, but not limited to, the iGEM public collections** [6]**. However, using them is often problematic, as the reported characteristics numbers for these RBSs have been acquired using different and sometimes unreliable methods. In contrast, our library has been created using standardised automated methods which contributed to relatively low coefficients of variation (‘reliable’). Additionally, compared to other widely known libraries which usually have up to 100 entries, ours has 445 RBSs (‘extensive’). We have updated the first paragraph of the ‘BUILD & TEST: Characteristics of the tested sequences’ section to reflect this:**



**d) “**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 authors do not agree that designing novel RBS is a trivial task - finding new RBSs by randomly generating a DNA sequence may be computationally easy, but finding ones with the desired qualities is hard. For example, Hollerer *et al*. in their work show relative TIRs for over 300,000 randomly sampled RBSs in Figure 3b** [3]**. There, one can see that around 75% of randomly generated RBSs perform worse by at least 25% compared to the strongest sequences. Also, please refer to our Figure 3A where we show that randomly generated RBSs perform significantly worse than the ones designed using our algorithm.**  
  
- 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.

**In our study, out of 450 tested sequences almost 40% have been created by sequence randomisation. Out of these randomised sequences, only a few got close to benchmark sequences TIR and were still 20% weaker than it is (Figure 3A). In fact, these 80% TIR ratio sequences were created by randomizing the sequence outside of the core RBS region, which were statistically shown not to be significantly impactful on the TIR (see supplementary Figure S1 and, for example, work by Jeschek *et al.*** [7]**). More representative would be sequences from other random groups, from which the best one achieved only about 65%** **of benchmark TIR. Contrasted with this, non-randomized, bandit driven design batch has shown much better results, with RBSs getting close to benchmark performance and even exceeding it. The authors believe that this is evidence enough that finding RBSs stronger than our benchmark (which, as stated before, is known to be very strong in similar context) to be a significant achievement (contrasted with simply randomly mutating the sequence). This is not, however, a surprise – the goal of this work was to show that this is achievable using ML.**

**To point out this fact to the readers, we have added the following paragraph to our subsection “ DESIGN: Performance of the recommendation algorithm”:**

  
  
-  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).

**While there are several computational approaches for designing RBSs, we respectfully disagree that these can be used with high confidence. There are works published as recently as late last year trying to address this fact, see Reis and Salis** [8] **or Hollerer *et al.*** [3]***.* In fact the problem has been persisting for years, see, for example work by Reeve *et al.*** [9]**.**

**The authors do agree with the comment that if the goal of the work was simply to maximise the expression level, it would be easier to simply swap the promoter. However, this was not the goal of this work. Instead, our goal was to show how machine learning can be paired together with automated, high-throughput laboratory methods to design RBS, in this case by maximising it. The exact design goal can be modified by subsequent users within the Bayesian Optimisation framework by altering the objective of the algorithm.**

**Finally, the authors wanted to point out that they did reference the plasmid in the last paragraph of the “The experimental workflow” subsection (3.1).**  
  
- 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.

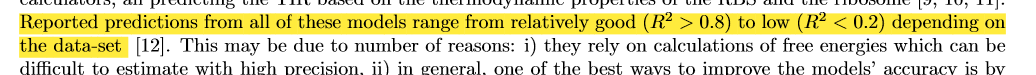
**All sequences have been tested in 6 replicates, not a single one, this is now specified in the 2.1.2 subsection:**



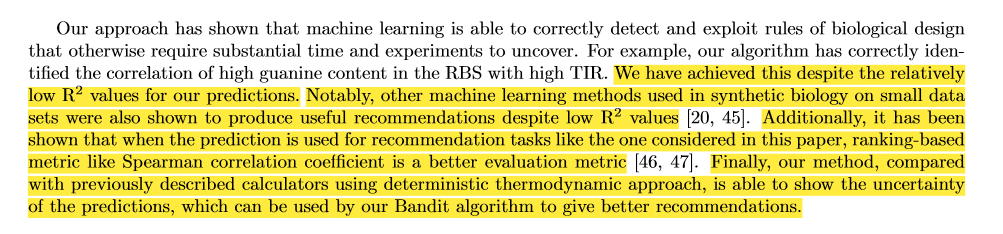
**The raw results for the 134% RBS can be found at row 364 and for the 115% RBS at row 368 of the Results\_Masterfile.xlsx file found in our supplementary file Code.zip\synbio\_rbs\data . Coefficients of variation (mean/standard variation) for these two RBSs are 7.1% and 6.1% respectively, which we would describe as very low for RBS samples, compared, for example, with the data compiled by Reis and Salis for their study**[8]***.***

- 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).

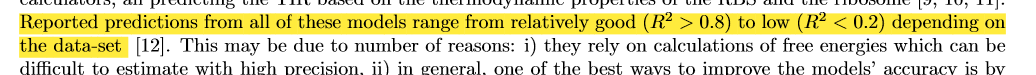
**The reported R^2 value of above 0.8 is actually the best case scenario, others studies for these models show much lower R^2 values depending on the used data set, we have updated the second paragraph of Introduction section to reflect that:**



**Additionally, the authors would like to point out that as long as the R^2 value is relatively low, it shouldn’t be viewed separately to other metrics shown in the study. Our ultimate goal is to be able to identify RBS sequences significantly higher than the benchmark sequence within fixed budget and our pipeline has successfully achieved that goal. When the prediction is used for recommendation tasks like the one considered in this paper, ranking-based metric like Spearman correlation coefficient is a better evaluation metric [10], [11]. Intuitively, in each round, as long as we can rank RBSs properly and distinguish RBS sequences with high from the ones with low TIRs, we can recommend RBS sequences exploitatively, even if the true label is not predicted with high precision. Other machine learning methods used in synthetic biology on small data-sets were also shown to produce useful recommendations despite low R^2 values [4], [12]. Finally, our predictor GPR, compared with previously described calculators using deterministic thermodynamic approach, is able to show the uncertainty of the predictions, which can be used by our Bandit algorithm to give better recommendations. We have updated our Discussion section to reflect these additional points.**

  
  
- 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.

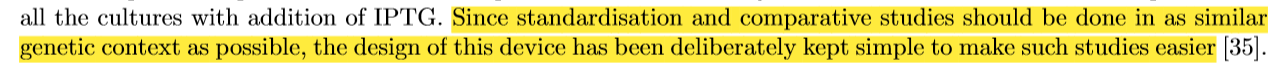
**Thank you for the suggestion to include the work. The authors did not choose to omit the work pointed out, rather they have decided to cite the original RBS calculator paper from Salis lab as the most relevant work from the said group. However, authors do recognize that the Reis and Salis paper does show the most recent advance in Salis RBS calculator and so did include it in the current revision of the manuscript under [12].**



**The work in question can’t be directly compared to ours in its goals. In the newly cited work Machine Learning is used for model evaluation, whereas in our work we present a new ML approach towards designing RBS, where ML is used directly used to learn from the data, predict the design space and guide recommendations. Additionally, we show how the round-to-round increase in data volume (by dividing our data set querying into distinct rounds) improves the model, whereas in the cited work all the data is presented to the model all at once.**

**Similarly, ref 25 (ref 26 in the current revision) can’t be directly compared to ours. It requires data sets orders of magnitude (over 300,000 RBSs) bigger than ours (450 RBSs) and uses neural networks, which is a different approach to ours. They also do not use the cycling approach to teaching the algorithm that we have presented in our work.** **Finally, our approach allows for much more targeted and rational design, rather than requiring generating hundreds of thousands of designs to ensure good design recommendations.**  
  
- 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. - 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.

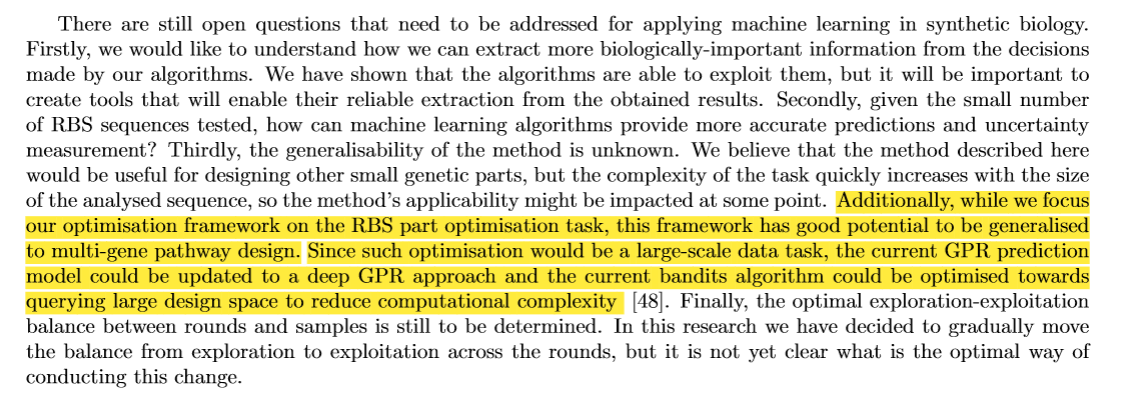
**The authors agree with the fact that in different context the RBS would lose their ability to be seen as “standard”. However**, **standardisation studies should be done in keeping the context as close to the original one as possible - see, for example, the classic work by Canton *et al.*** [13]**.** **This is the reason why we have kept our system simple – so that other researchers could easily replicate the conditions should they decide to** **test their RBSs against ours. We have added the following sentence to the second paragraph of the “The experimental workflow” subsection to underline this point:**

  
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?

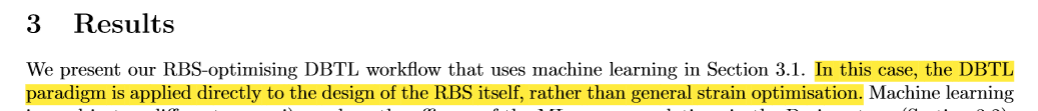
**The authors believe that their approach is appropriate for design of small genetic parts in isolation. It’s important to make a distinction between works focusing on the part itself like Hollerer *et al.*** [3] **and works where the part is optimised with a wider goal of strain optimisation like in Jervis *et al.*** [1]**. The part-centric optimisation, like the one considered in this paper, is an important task in synthetic biology because metabolic pathways are made up of individual parts and understanding their performance in separation allows for better strain design.**

**While we focus our optimisation framework on the RBS part optimisation task, this framework has good potential to be generalised to multi-gene pathway design as reviewed by Lawson *et al.*** [14] **and proposed by HamediRad *et al.*** [15] **. For example, the optimisation goal can be adjusted to address combinatorial optimisation for multi-gene and RBS scenarios; since this would be a large-scale data task, the current GPR prediction model could be updated to a deep GPR approach and the current bandits algorithm could be optimised towards querying large design space to reduce computational complexity** [16]**.**

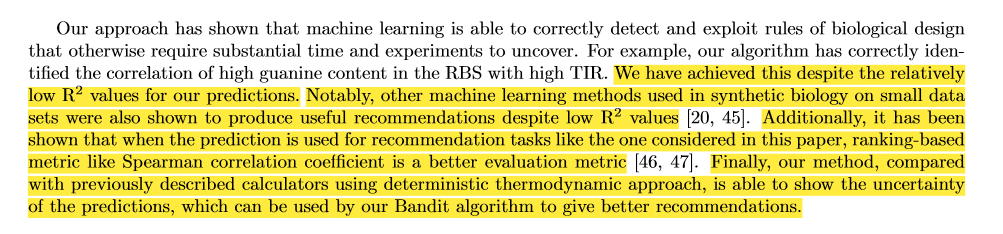
**We have now added the following discussion describing our method’s generalisation ability to the Discussion section:**

  
  
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.

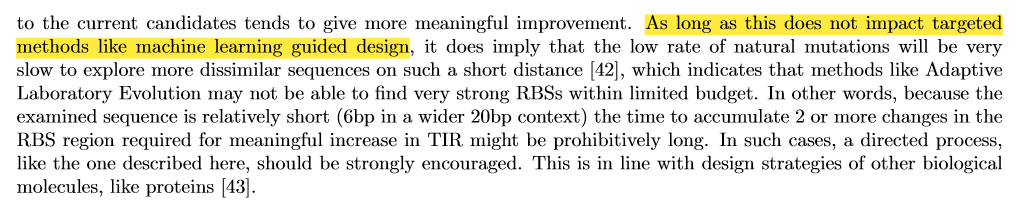
**The authors agree that there is a multitude factors that play into the final performance of a strain. Having said that when we refer to the DBTL cycle in our paper we specifically focus on its use for the design of the RBS as a genetic part rather than wider strain optimisation. We have updated our opening to the results section with the following sentence to better reflect that:**

  
  
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.

**The authors would like to point out that as long as the R^2 value is relatively low, it shouldn’t be viewed separately to other metrics shown in the study. Our ultimate goal is to be able to identify RBS sequences significantly higher than the benchmark sequence within fixed budget and our pipeline has successfully achieved that goal. When the prediction is used for recommendation tasks like the one considered in this paper, ranking-based metric like Spearman correlation coefficient is a better evaluation metric [10], [11]. Intuitively, in each round, as long as we can rank RBSs properly and distinguish RBS sequences with high from the ones with low TIRs, we can recommend RBS sequences exploitatively, even if the true label is not predicted with high precision. Other machine learning methods used in synthetic biology on small data-sets were also shown to produce useful recommendations despite low R^2 values [4], [12]. Finally, our predictor GPR, compared with previously described calculators using deterministic thermodynamic approach, is able to show the uncertainty of the predictions, which can be used by our Bandit algorithm to give better recommendations. We have updated our Discussion section to reflect these additional points.**

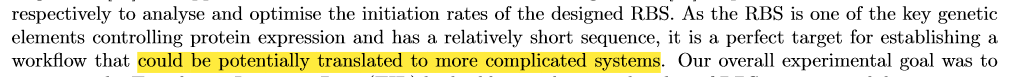
  
  
**TEXT HERE**  
  
LEARN: Not clear why Hamming distance might impact the budget. As the authors are using Machine Learning, predicted sequences that are Hamming distant might be possible, not necessarily just those of single edit.

**The Hamming distance, as mentioned in the text, does not impact Machine Learning. However, it does pose a problem when an untargeted (random) approach like Adaptive Laboratory Evolution would be used as it might not be able to generate sequences distant enough which we show is required for positive change in RBS TIR. We explain this in more detail in third paragraph of subsection “BUILD & TEST: Characteristics of the tested sequences”, which we updated to help prevent future confusion:**

  
  
**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.

**All fixed.** **Line 40 changed to:**

  
  
Page 7:   
- the workflow description is very generic/inconcrete

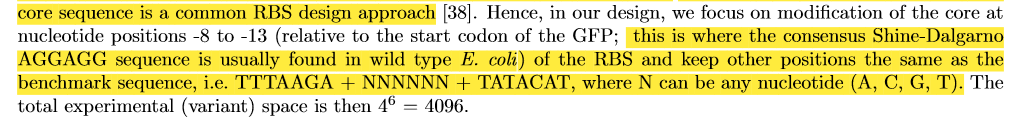
**That is intentional – all the required details are included in the Methods section, following the standard practice.**

- k is mentioned in the Figure caption, but not used in the figure (Fig. 2)

**k is used in the cell in the upper left corner: “n=k”.**

- line 54: the 20 bp sequence is only 19 bp long (T missing); the authors should specify what they mean by “core RBS”

**Added the missing T, added core of the RBS description:**

  
  
Page 8:   
- line 30: what is the “original, consensus sequence”?

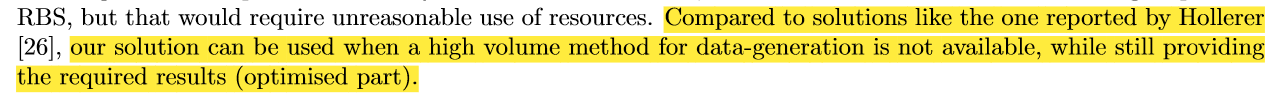
**Corrected from consensus to benchmark.**

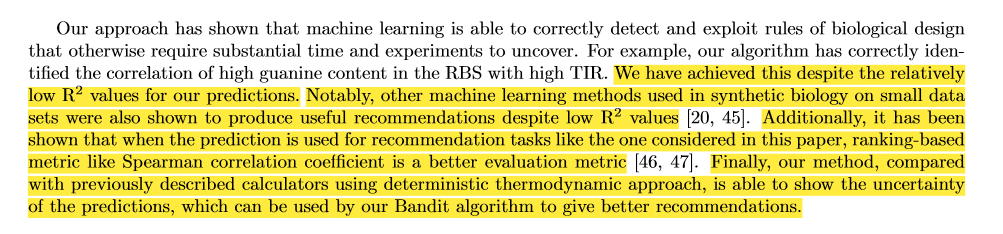
- the authors should explain in more detail what the “Bandit-0” dataset is, how it was generated, and why they selected this

**In subsection 2.2 we give a brief overview of how and why Bandit-0 has been generated with more detailed explanation available in supplementary materials, see section A.4**  
  
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”?

**Description of the Shine-Dalgarno sequence has now been put into text as specified by previous answer.**   
  
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.

**The authors agree that the way that comparison has been put into the discussion section was too simplistic.** **While it is true that the method presented by Hollerer *et al.*** **allows for acquisition of thousands of data points, it does rely on NGS and its analysis, which in itself can be also seen as a complicated method.** **We would also argue that we have not used complicated methods, merely standard laboratory ones scaled up using automation.** **We would still argue that our method compares favourably, as it can be used with much smaller data sets and that is usually the norm when it comes to biological research**. **We have changed the relevant part in the text to better reflect this:**

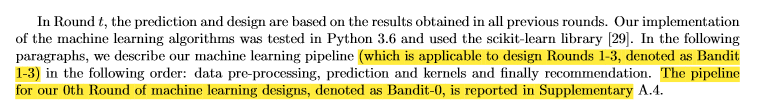
**The authors would like to point out that as long as the R^2 value is relatively low, it shouldn’t be viewed separately to other metrics shown in the study, especially the Spearman correlation coefficient which shows much higher values. When the prediction is used for recommendation tasks like the one considered in this paper, ranking-based metric like Spearman correlation coefficient is a better evaluation metric [10], [11]. Intuitively, in each round, as long as we can rank RBSs properly and distinguish RBS sequences with high from the ones with low TIRs, we can recommend RBS sequences exploitatively, even if the true label is not predicted with high precision. Other machine learning methods used in synthetic biology on small data-sets were also shown to produce useful recommendations despite low R^2 values [4], [12]. We have updated our Discussion section to reflect these additional points.**



Reviewer: 2   
without review: I have some comments/questions (see attached file) which, if addressed, I believe would improve the clarity of the manuscript:

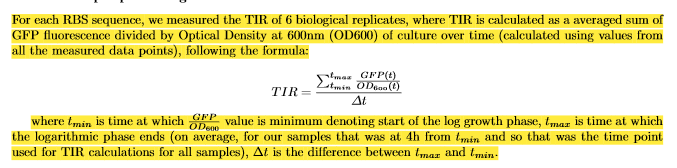
1. It is not completely clear from the beginning what do the authors refer to with the term Bandit and what with Round. It might be helpful to explain it somewhere in the beginning of Section2.2. It would also be advisable to clarify that Round refers to DBTL round and use either “round” or “Round” throughout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

**The authors have clarified the terms in section 2.2 and unified them throughout the text following this suggestion. The second paragraph of section 2.2 now reads:**



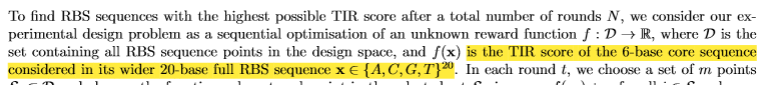
2. Section 2.2.1 – Please clarify how TIR is calculated, in particular the derivative part. An expression in a formula may be helpful.

**The TIR formula has been incorrectly called a derivative in the original text. This has now been corrected and the relevant section now reads:**



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.

**The authors have updated the relevant section following this suggestion and now it reads:**



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).

**We assume that the unknown reward function is sampled from a Gaussian Process (GP), where the GP is specified by its mean function and covariance function. In appendix, we introduced GP regression conditioned on observed samples. The posterior mean and covariance function was introduced with the assumption the mean function is zero, following [27]. The posterior predictions for non-zero mean case can be inferred from the zero mean case trivially. To avoid the confusion, we change the appendix to include the prediction for all cases (including non-zero mean functions).**

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

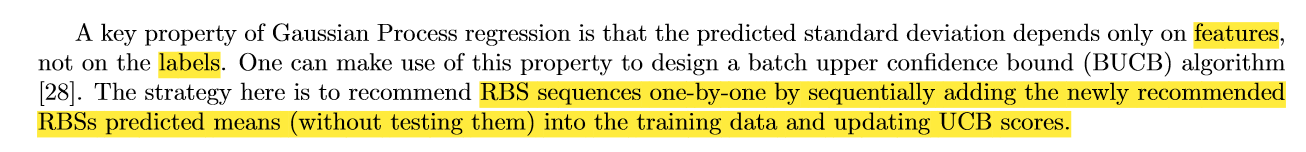
**We have updated our text to reflect this:**

 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”.

**We have revised the text to denote RBS sequences as “features” and TIR as “label” to avoid confusion. The “observed” is now only used for describing the status, i.e. whether we have tested the relevant sequences.**

7. Page 5, last paragraph – Given the batch approach for generating recommendations, it would be useful 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.

**We have updated the Recommendation: Batch Upper Confidence Bound Bandit subsection of the manuscript to read to better explain this point:**



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

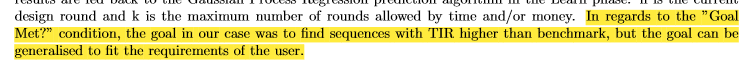
**It was a typo. It should read “...mean ±1.96 standard deviation” and the text has been updated accordingly.**

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

**As suggested, we have increased the font size in all figures.**

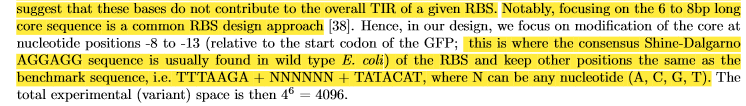
10. Section 3.1, Figure 2 – Please explain TIR(n)=TIR(n-1) in the caption or make a reference in the text. Also, in caption n, k –>n,k.

**Upon additional consideration, the authors have decided that the TIR(n)=TIR(n-1) condition does not represent the intended statement properly. This has now changed to “Goal Met?” in the respective box with the following explanation set in the caption:**



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.

**The authors agree with this observation. However, this is mostly because our benchmark sequence is already very strong and random changes will in most cases result in decreased TIR, as exemplified in our other results. Additionally, our follow-up experiments on bandits groups show that the modification of the core part gives satisfying improvement of TIR. Further, focusing on the 6 to 8bp long core sequence is a common RBS design approach** [7]**. We have added the following sentences to make it clearer in our text:**

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

**Text has been updated following this suggestion.**

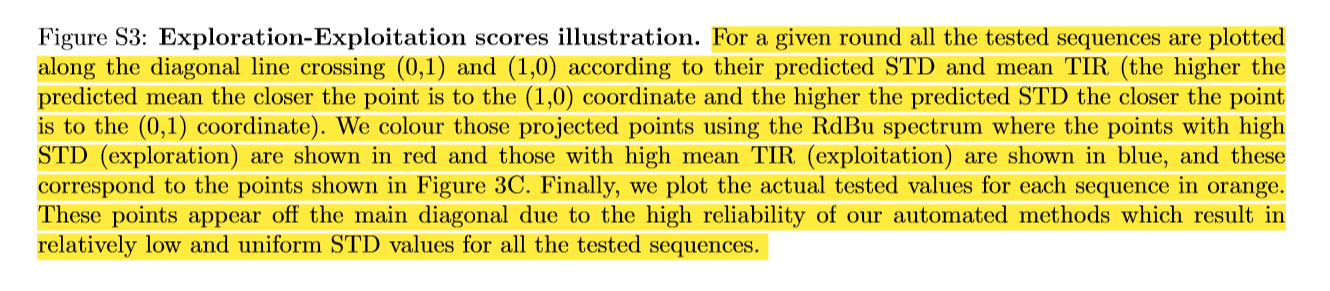
13. Figure 3 caption, line 46 – Two consecutive “the”.

**Second “the” has been removed.**

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

**To better explain this point, we have updated the caption for Figure S3 to read:**

**~~“The orange points show the predicted STD and TIR (with the setting of Figure \ref{fig: Scatterplot}) for test data points (recommendations in Round 1-3) via the Gaussian Process Regressor. We project the orange points to the diagonal line crossing (0,1) and (1,0). We colour those projected points using RdBu palette via the projected values, which are exploration-exploitation colours/scores shown in Figure 3C. The hued red colour corresponds to points that are predicted to have high STD (uncertainty), which are recommended for exploration; the hued blue colour corresponds to points that are predicted to have high TIR, which are recommended for exploitation.”~~**

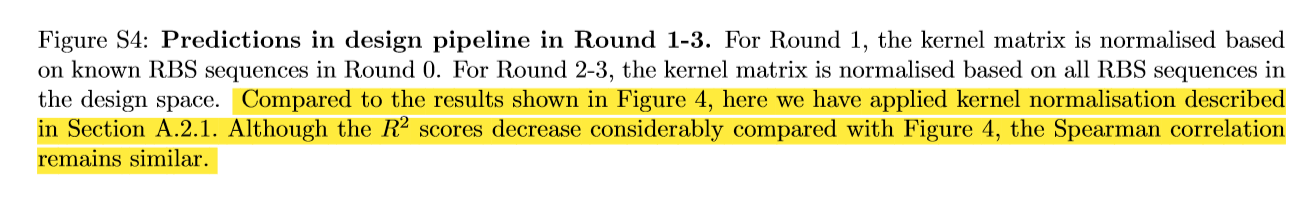


15. Page 10, line 15 – Please give more details on “our embedding function”.

**We have updated this part of the text to specify “our embedding function” which is the weighted degree kernel with shift.**

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.

**In Fig 4 the kernel matrix is not normalised, while in Fig S4, kernel matrix is normalised as shown in A.2.1. We have updated the figure caption to better explain that:**



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

**The sentence has been updated to:**



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

**We have updated the text accordingly.**

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...”

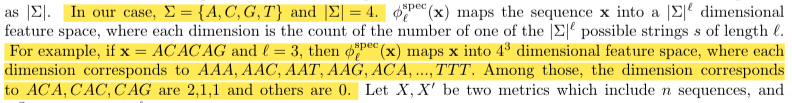
4. The part “diagonal matrixα2Ion the kernel matrix” is not clear, Eq. (5) suggests it is a sum.

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

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).

**We’ve added an example in Sec A.2. i.e.**



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 is d?

**φ is the general feature mapping function used to calculate specific kernel, one example could be φ^spec\_l mentioned in spectrum kernel. d is the dimension of feature.**

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

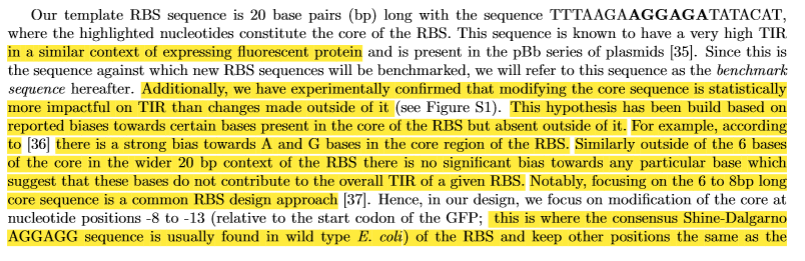
**From Eq. (16) to Eq. (17), we follow the definition of kernel. We have restated it in Sec A.2.1 to avoid confusion in revision.**

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

**We’ve fixed all the typos mentioned in points 1-11 above.**

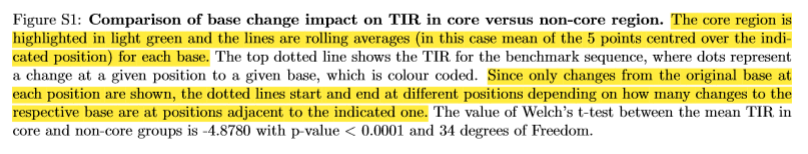
12. Figure S1:– What exactly are the rolling averages? Why they don’t start from the same position for all bases?– There is a self-reference to Fig. S1 within the caption.– What is “set X of sequences”?– It would be useful to mention in the caption that these are single nucleotide changes.

**Part of the main text relating to this plot has been mistakenly left in this caption. This has now been fixed and both main text:**





**and the caption in question:**



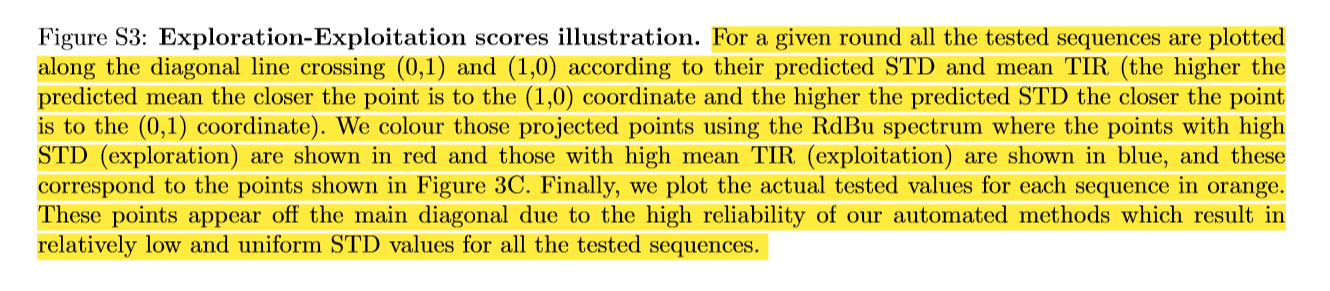
**have been corrected.**

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

**It was a typo, it has been removed for this revision.**

14. Figure S3 – It is not clear what are the projected values

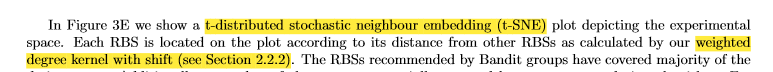
**To better explain this point, we have updated the caption for Figure S3 to read:**

  
  
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.

**Our results for random sampling are in line with other results published in literature. For example, Hollerer *et al*. in their work show relative TIRs for over 300,000 randomly sampled RBSs in Figure 3b** [3]**. There, one can see that most randomly generated RBSs perform worse by at least 25% compared to the strongest sequences. In our case the benchmark sequence is very strong and our randomly sampled sequences also perform at least 25% worse than it does.**   
  
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?

**While focusing on exploitation, the algorithm recommends RBSs that have a high probability of having high TIR, resulting in overall higher TIR of recommended RBSs. On the other hand, in exploration, algorithm focuses on RBSs that have high uncertainty of predictions which results in an overall lower TIR of recommended RBSs since there is a higher chance of finding RBSs with low TIR. However, like we mentioned in text, exploration allows for better understanding of the underlying function and, in result, better predictions in subsequent rounds.**  
  
I believe that t-SNE is not actually defined.

**We have updated the first sentence of the last paragraph of subsection 3.2 to read:**



**References**

[1] A. J. Jervis *et al.*, “Machine Learning of Designed Translational Control Allows Predictive Pathway Optimization in Escherichia coli,” *ACS Synth. Biol.*, vol. 8, no. 1, pp. 127–136, Jan. 2019, doi: 10.1021/acssynbio.8b00398.

[2] Z. Costello and H. G. Martin, “A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data,” *npj Syst. Biol. Appl.*, vol. 4, no. 1, p. 19, 2018, doi: 10.1038/s41540-018-0054-3.

[3] S. Höllerer *et al.*, “Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping,” *Nat. Commun.*, vol. 11, no. 1, p. 3551, 2020, doi: 10.1038/s41467-020-17222-4.

[4] P. Opgenorth *et al.*, “Lessons from Two Design–Build–Test–Learn Cycles of Dodecanol Production in Escherichia coli Aided by Machine Learning,” *ACS Synth. Biol.*, vol. 8, no. 6, pp. 1337–1351, Jun. 2019, doi: 10.1021/acssynbio.9b00020.

[5] J. Beal *et al.*, “Comparative analysis of three studies measuring fluorescence from engineered bacterial genetic constructs,” *PLoS One*, vol. 16, no. 6, p. e0252263, Jun. 2021, [Online]. Available: https://doi.org/10.1371/journal.pone.0252263.

[6] iGEM, “RBS Catalog,” *iGEM Parts Registry*. http://parts.igem.org/Ribosome\_Binding\_Sites/Catalog.

[7] M. Jeschek, D. Gerngross, and S. Panke, “Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort,” *Nature communications*, vol. 7. Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland., p. 11163, 2016, doi: 10.1038/ncomms11163.

[8] A. C. Reis and H. M. Salis, “An Automated Model Test System for Systematic Development and Improvement of Gene Expression Models,” *ACS Synth. Biol.*, vol. 9, no. 11, pp. 3145–3156, Nov. 2020, doi: 10.1021/acssynbio.0c00394.

[9] B. Reeve, T. Hargest, C. Gilbert, and T. Ellis, “Predicting translation initiation rates for designing synthetic biology,” *Front. Bioeng. Biotechnol.*, vol. 2, p. 1, Jan. 2014, doi: 10.3389/fbioe.2014.00001.

[10] J.-S. Kang, D.-H. Shin, J.-W. Baek, and K. Chung, “Activity Recommendation Model Using Rank Correlation for Chronic Stress Management,” *Appl. Sci.*, vol. 9, no. 20, 2019, doi: 10.3390/app9204284.

[11] P. Schober, C. Boer, and L. A. Schwarte, “Correlation Coefficients: Appropriate Use and Interpretation,” *Anesth. Analg.*, vol. 126, no. 5, 2018, [Online]. Available: https://journals.lww.com/anesthesia-analgesia/Fulltext/2018/05000/Correlation\_Coefficients\_\_Appropriate\_Use\_and.50.aspx.

[12] T. Radivojević, Z. Costello, K. Workman, and H. Garcia Martin, “A machine learning Automated Recommendation Tool for synthetic biology,” *Nat. Commun.*, vol. 11, no. 1, p. 4879, 2020, doi: 10.1038/s41467-020-18008-4.

[13] B. Canton, A. Labno, and D. Endy, “Refinement and standardization of synthetic biological parts and devices,” *Nat. Biotechnol.*, vol. 26, no. 7, pp. 787–793, 2008, doi: 10.1038/nbt1413.

[14] C. E. Lawson *et al.*, “Machine learning for metabolic engineering: A review,” *Metab. Eng.*, vol. 63, pp. 34–60, 2021, doi: https://doi.org/10.1016/j.ymben.2020.10.005.

[15] M. HamediRad, R. Chao, S. Weisberg, J. Lian, S. Sinha, and H. Zhao, “Towards a fully automated algorithm driven platform for biosystems design,” *Nat. Commun.*, vol. 10, no. 1, p. 5150, 2019, doi: 10.1038/s41467-019-13189-z.

[16] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, “Taking the Human Out of the Loop: A Review of Bayesian Optimization,” *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, 2016, doi: 10.1109/JPROC.2015.2494218.