# Opportunities and Challenges in Designing Genomic Sequences

**Mengyan Zhang** [1 2] **Cheng Soon Ong** [2 1]

## Abstract

Experimental design based on black-box optimization and batch recommendation have been increasingly used for the design of genomic sequences. We briefly outline our recent results on bacteria gene expression maximisation with Bayesian optimisation, where machine learning enabled us to discover a strong regulatory element. Using the Design-Build-Test-Learn (DBTL) workflow as a case study of how to effectively use machine learning in genomic sequence design, we argue that machine learning has tremendous potential in this area. Based on our experience, we discuss several opportunities and challenges that we have identified, and conclude with a call to action for more collaborations.

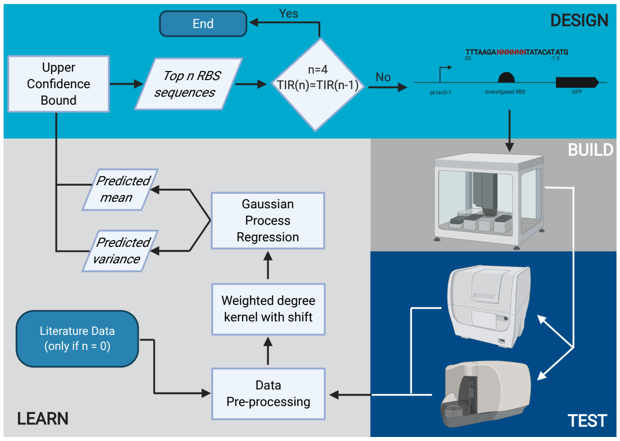

*Figure 1.* Our Design-Build-Test-Learn (DBTL) workflow: Bayesian Optimisation based experimental design.

## 1. Introduction

Recent years have witnessed increasing needs for designing genomic sequences with the aid of the machine learning (ML) algorithms, where the task is to maximise the protein expression level by designing biological sequences in batches. The Design-Build-Test-Learn (DBTL) cycle has been increasingly adopted in genomic sequences design framework (Opgenorth et al., 2019) and recommendation tools have been proposed to address DBTL cycle (Radivojević et al., 2020). However, the use of machine learning algorithms is still in an immature stage. For example, predictions usually lack strong correlations to the measured labels due to the noisy measurement and a small number of data points; the uncertainty quantification and batch recommendation is not well-addressed (Lawson et al., 2021).

LEARN and DESIGN parts can be addressed by two ingredients of the Bayesian optimisation framework. The first ingredient is a probabilistic prediction model (LEARN) such as Gaussian Process Regression (GPR) (Rasmussen, 2004), which captures the updated belief over objective functions with observed data coming sequentially (or in batches). The second ingredient is a decision making policy (DESIGN), which gives an acquisition function guiding the exploration and lies across the bandits algorithms (Lattimore & Szepesvári, 2020). One key point is to balance the exploration-exploitation: the *exploration* of the high-uncertainty parts of the design space where ribosome binding site with a high label can be hidden, and *exploitation* whose goal is querying areas that are predicted to give relatively high labels.

We show a case study (Section 2) on designing of the Ribosome Binding Site (RBS), which controls the recruitment of a ribosome during the initiation of translation and thus influences the protein expression level. Our workflow can be described as the Design-Build-Test-Learn (DBTL) cycle (Figure 1), where the given genetic part or organism are continually improved in batches. We use the Gaussian Process Batch Upper Confidence Bound (GP-BUCB) algorithm (Desautels et al., 2014) for LEARN and DESIGN in our DBTL workflow. Combining Bayesian optimisation and DBTL cycle helps us find RBSs with high protein expression level with a small budget. Based on our case study and the recent literature, we discuss the opportunities and challenges in the interdisciplinary of machine learning and genomic sequence design in Section 3.

[1] Department of Computer Science, Australian National University [2] Data61, CSIRO. Correspondence to: Mengyan Zhang <mengyan.zhang@anu.edu.au>, Cheng Soon Ong <cheng-soon.ong@anu.edu.au>.

The 2021 ICML Workshop on Computational Biology.

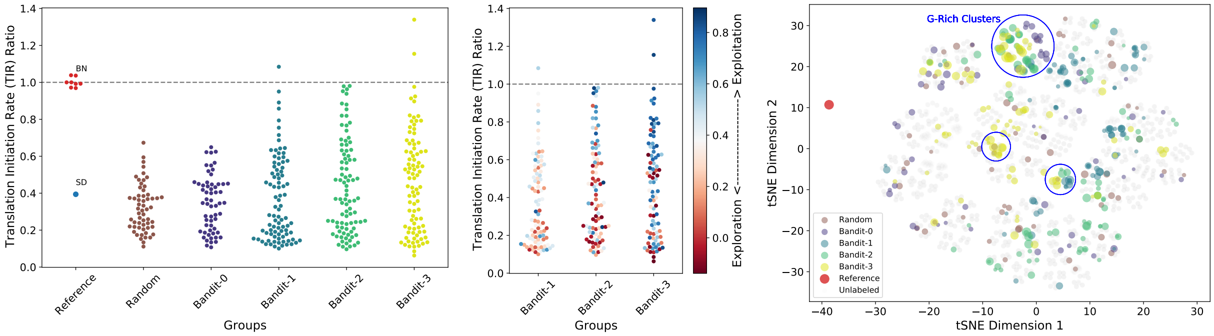

*Figure 2.* Experimental Result of case study. The TIR results in all subplots are shown normalised to the respective benchmark sequence sample which acts as an internal standard. From left to right, we show the swarm plot of all measured groups, the exploitation v.s. exploration for Bandit 1-3, and the tSNE plot over the measured groups.

## 2. Case Study: RBS design

To provide a concrete example to illustrate the opportunities and challenges in Section 3, we briefly summarise a recent project (Zhang et al., 2021). Our recent work uses Bayesian optimisation as part of the DBTL cycle to predict (LEARN) and recommend (DESIGN) variants of *E. coli* RBS. Our overall experimental goal is to biologically construct *E. coli* with large Translation Initiation Rate (TIR, a measure of protein expression level), for a protein of interest.

**Methods:** Our workflow is shown in Figure 1. In the zeroth round, randomised RBS sequences and preliminary machine learning recommendations based on the literature data (Jervis et al., 2018) are designed to explore the experimental space. In the subsequent rounds, based on the data obtained in the previous rounds, designs (RBS sequences) are recommended by *Gaussian Process Batch Upper Confidence Bound (GP-BUCB)* algorithm (Desautels et al., 2014). We used the *weighted degree kernel with shift (WDS)* as the Gaussian process covariance function to capture sequence similarities. The designs are then physically constructed in batches of 90, to fit the number of wells available in our high throughput automated laboratory system. For each recommended RBS sequence, we tested 6 biological replicates, and measured the corresponding TIR. After construction, the plasmids harbouring the new genetic devices are measured using spectrophotometry of the green fluorescent protein (GFP). The resulting TIR values are used as labels for GP-BUCB to recommend the next round of designs.

**Experimental Setup:** In *E. coli*, the RBS is usually located in the 20 bases upstream of the start codon. In our design, we focus on randomising the core of the binding site, at positions -8 to -13 relative to the start codon, and fix the other bases to be the same as the benchmark sequence, i.e. TTTAAGA+NNNNNN+TATACAT, where N in the core part can be any choice of A, C, G, T. The total experimental (variant) space to search is $4^6 = 4096$. We have the following groups tested on the core part: **BN**: known benchmark sequences *AGGAGA* (Lee et al., 2011); **SD**: consensus RBS core sequence called the *Shine-Dalgarno sequence*, which in *E. coli* is *AGGAGG*; **Random**: 60 randomly generated RBS sequences; **Bandit-0**: 60 RBS sequences recommended by our algorithm based on literature data (Jervis et al., 2018); In the subsequent 3 rounds, 90 designs were generated using our algorithm based on the data obtained from the previous rounds (**Bandit 1-3**).

**Results:** Our results are shown in Figure 2. We obtain the *TIR ratio* by taking the ratio between the raw TIR and the average TIR of the benchmark sequence (which are run in triplicate) in each round. The left plot shows the swarm plot of the TIR ratio for all the examined groups. Bandit 1-3 shows better performance than other groups and notably we have identified sequences that were 34%, 15% and 8% stronger than the benchmark sequence. At the same time, we build an extensive, reliable library of novel RBSs with diverse sequences.

In the middle plot, we coloured the data points for Bandit 1-3 groups according to their relative exploration-exploitation affinity. Those with high predicted mean are coloured blue and represents exploitation, those hued red are with high predicted uncertainty and represent exploration. We can see the RBSs with high TIRs tend to come from exploiting the design space whereas the explorative points give relatively low TIR but expand our knowledge about unknown parts of the design space.

The right plot shows a tSNE plot, where the relative distances between sequences in our design space are calculated based on the WDS kernel. The area of each dot represents the experimentally obtained TIR for measured groups. The measured RBSs have covered the majority of the design space. A number of clusters with high TIRs (e.g. *G-Rich Clusters*) are increasingly targeted by our recommendation algorithm over rounds.

## 3. Opportunities and Challenges

In this section, we will start by reviewing the observations from our case study and related literature. Then we will discuss opportunities and challenges in genomic sequence design.

### 3.1. Generalisation of Our Workflow

The DBTL workflow based on Bayesian optimisation approaches has the potential to be generalised.

**Opportunities:** Most of the biological experiments only provide one or two sample labels for each queried genomic sequence (Jervis et al., 2018; Opgenorth et al., 2019). In our case study, the high throughput automated laboratory workflow enabled us to measure the TIR of six biological replicates for each RBS sequence, where we observed a high signal to noise ratio. Our case study focused on designing RBS sequences on the core part (6-bps). A future direction is to extend our DBTL workflow to promoters to a larger design space. Furthermore, a general framework or tool which can be used to accelerate the design of the genomic sequences for various organisms is needed. Other biological sequences, such as designing peptide sequences for a particular surface protein in a vaccine, are also amenable to similar techniques.

**Challenges:** One particular challenge is how to deal with the large design space in such an interactive recommendation system. While exploring a large design space, the size of labelled RBS sequences is quite small, especially among the first few cycles. In such a case, generating high accuracy prediction with high confidence is challenging. One strategy is to make suitable assumptions, such as smoothness or sparsity, over the design space and predictors. Another strategy is to use *transfer learning* (Weiss et al., 2016), where one can transfer knowledge gained from other related datasets or tasks to the current task.

### 3.2. Measurement Noise and Data Normalisation

The measurement process in biology is stochastic, with poorly understood measurement noise. As mentioned in Section 3.1, automated workflows could generate more data, which may allow us to empirically estimate a noise model. Classical independent identically distributed (iid) statistical assumptions may no longer hold in the adaptive experimental design settings.

**Opportunities:** Our automated lab workflow returned 79% construct with less than 40% coefficient of variation (STD/AVERAGE, over 6 biological replicates) in each batch, and we successfully got 99% of constructs. High throughput biological experiments allow results to have more statistical power. A high signal to noise ratio can be achieved with automated laboratory workflows, which are also cheap and efficient. This provides an opportunity for reproducible experiments in biology. Better noise properties also allow machine learning to generate more reliable predictions. For example, different levels of noise can be modelled in ML algorithms (Mchutchon & Rasmussen, 2011).

**Challenges:** In the case study, there was variation per sequence (6 replicates) in each batch, and variation between batches. Proper normalisation methods over both the biological replicates and different batches are needed for non-independent sampling. A question to ask is that how many biological v.s. technical replicates are needed, considering the balance of measurement noise and time/money. Additionally, biological experimental designs are usually time-consuming and conducted over several months or years. The measured label for the same sequence in different batches could be significantly different. A better noise model also allows better downstream algorithms, such as using the uncertainty in predictions for bandit algorithms.

### 3.3. Exploitation-Exploration Tradeoff

Guided by the UCB algorithm, our case study results in Figure 2 show both good coverage of the design space (exploration) and the clear pattern of increasing querying of RBS sequences with high TIR over rounds (exploitation).

**Opportunities:** The growth of machine learning allows increasing power of generating high accuracy prediction and uncertainty quantification, which enable efficient search of sequence candidates. The recommendation tool *ART* recently proposed by Radivojević et al. (2020) provides a good example considering both prediction and uncertainty in the DBTL cycle. More advanced techniques can be applied in such workflow, including conformalized quantile regression (Romano et al., 2019), batch recommendations (Desautels et al., 2014; Wilson et al., 2018), and practical bandit algorithms with impact (Bouneffouf et al., 2020). More generally, we could consider reinforcement learning (RL) algorithms for experimental design.

**Challenges:** Although bandits algorithms are well-studied in theoretical view of points for a large number of iterations (Desautels et al., 2014; Zhang & Ong, 2021), there are gaps between theory and practice. For example in practice, how to choose a hyperparameter that controls the exploration-exploitation tradeoff in the interactive DBTL cycles remains to be an open question. It is uncommon for bandit algorithms to consider structured discrete design spaces such as required in genomics.

### 3.4. Representation of Biological Sequences

Representing biological sequences into numerical vectors plays an important role in capturing the biological structures and similarities between sequences. Most of the previous studies (Jervis et al., 2018; Radivojević et al., 2020; Opgenorth et al., 2019) used one-hot embedding to convert biological sequences into vectors. String kernels (Ben-Hur et al., 2008) are also adopted and show an improvement on downstream prediction tasks.

**Opportunities:** Natural language processing (NLP) has progressed in recent years, which provides opportunities for understanding the language of biological sequences. Recent work show successes in applying NLP techniques such as BERT (Devlin et al., 2019) to DNA (Ji et al., 2020) and proteins (Rao et al., 2019), where pre-trained models are trained on large-scale unlabelled data in a self-supervised way and fine-tuned on relatively small datasets.

**Challenges:** It is widely accepted that one of the key difficulties with genomic data is the issue of high dimensionality coupled with low sample size. Furthermore, it remains a question that what would be the evolutionary distance of the organisms, in terms of transfer learning? e.g. when the model pre-trained on *E. coli* can be transferred to yeast. An empirical study showing the transferability of the pre-trained model over different organisms and tasks would be interesting and useful for future studies. Additionally, *kmers* are treated as *words* in NLP-based models. Learning how to split biological sequences into meaningful *words* might help us understand more about biological language.

### 3.5. Evaluation and Interpretablity

How to measure success in experiments where we recommend new designs? In our case study (Section 2), we compared with random designs. There is no unified evaluation method for the DBTL design in the literature.

**Opportunities:** Appropriate evaluation methods need to be designed into the DBTL cycle and provide feedback on the design in future batches (Walsh et al., 2021). Beyond numerical evaluation, visualization can have the benefits of increasing the interpretability of the recommendation approach (Hu et al., 2019). The interpretability of the models and results are important in genomic sequence design, which helps us to understand the approach and generate new biological knowledge based on the observations.

**Challenges:** The evaluation metric in bandit literature is the *expected regret* (Lattimore & Szepesvári, 2020), which is the expected difference between the reward obtained and the best rewards one can obtain. The expectation is taken with respect to the randomness of the environment, while we only have one chance to experiment in practice. The best TIR remains to be unknown unless we label the whole design space, which is ususally intractable in practice. Another choice is to combine domain knowledge and evaluation. For example, how to design the reward for increasing biological knowledge? Are there downstream tasks (e.g. drug efficacy) we can use to help measure success? Whether the biological prior knowledge (assuming it can be encoded) would bias the exploration and lead to miss potential useful design areas.

### 3.6. Interdisciplinary Collaboration

DBTL cycles take inputs from both machine learning (LEARN and DESIGN) and synthetic biology (BUILD and TEST). Interdisciplinary collaboration is needed to drive the development and progress in this field.

**Opportunities:** In our case, collaboration creates benefits for both ML and synthetic biology researches: the ML algorithm enables a more efficient search of the design space; the experimental apparatus provides opportunities to design ML algorithms to better suit real-world tasks. More generally, for other genomic sequence design tasks, recommendation algorithms such as bandits enable us to efficiently explore the exponentially large design space. The high throughput experiments currently undertaken by genomic projects provide an excellent place for ML researchers to identify new problems and new learning tasks. Furthermore, tools and platforms (e.g. open-source software, open access) can help to communicate and disseminate progress in both fields.

**Challenges:** As this community knows, there are hard problems to be addressed. Communication between experts in different areas is challenging, as the same word (such as model) can mean different things in each field. The level of abstraction in discussions is often confusing (for example ML algorithm could mean training or prediction). Cultural norms such as publication venues, as well as traditional organisational structures in university departments do not provide many viable career options for early-career researchers. Ultimately the motivation of collaborators may not be the same, e.g. publish a paper v.s. developing a more efficient drug pipeline. Cross-disciplinary collaboration has occurred because of social networks, and the risks associated with research is further amplified in cross-disciplinary settings.

## 4. Call to action

We conclude by summarising the opportunities and challenges. As identified half a century ago by Thomas Kuhn, scientific revolutions occur when there is cross pollination of ideas. In designing genomic sequences, the human expert and the ML algorithm needs to work together, implying the need for better evaluation metrics, and interpretable results that allow biological knowledge to be obtained. One key aspect is to capture our knowledge about DNA and protein sequences in an appropriate vectorial representation suitable for ML. With accurate and efficient ML predictors, we can empirically study the exploration-exploitation tradeoff, providing motivation to novel theoretical analysis. The resulting experimental data from adaptive designs will provide new statistical settings that do not follow classical iid assumptions. We hope that our case study, and generalisations to our workflow, will stimulate research in both biology and machine learning.

## Acknowledgments

The authors would like to acknowledge CSIRO's Machine Learning and Artificial Intelligence, and Synthetic Biology Future Science Platforms for providing funding for this research. The authors would also like to thank CSIRO BioFoundry for help with performing the experiments.

## References

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. Support vector machines and kernels for computational biology. *PLoS computational biology*, 4 (10), 2008.

Bouneffouf, D., Rish, I., and Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2020. doi: 10.1109/CEC48606.2020.9185782.

Desautels, T., Krause, A., and Burdick, J. W. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923, 2014.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4171–4186, 2019.

Hu, K., Bakker, M. A., Li, S., Kraska, T., and Hidalgo, C. *VizML: A Machine Learning Approach to Visualization Recommendation*, pp. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450359702.

Jervis, A. J., Carbonell, P., Vinaixa, M., Dunstan, M. S., Hollywood, K. A., Robinson, C. J., Rattray, N. J., Yan, C., Swainston, N., Currin, A., et al. Machine learning of designed translational control allows predictive pathway optimization in escherichia coli. *ACS synthetic biology*, 8 (1):127–136, 2018.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *bioRxiv*, 2020. doi: 10.1101/2020.09.17.301879.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lawson, C. E., Martí, J. M., Radivojevic, T., Jonnalagadda, S. V. R., Gentz, R., Hillson, N. J., Peisert, S., Kim, J., Simmons, B. A., Petzold, C. J., Singer, S. W., Mukhopadhyay, A., Tanjore, D., Dunn, J. G., and Garcia Martin, H. Machine learning for metabolic engineering: A review. *Metabolic Engineering*, 63:34–60, 2021. ISSN 1096-7176. Tools and Strategies of Metabolic Engineering.

Lee, T. S., Krupa, R. A., Zhang, F., Hajimorad, M., Holtz, W. J., Prasad, N., Lee, S. K., and Keasling, J. D. BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *Journal of biological engineering*, 5:12, 9 2011. ISSN 1754-1611 (Electronic).

Mchutchon, A. and Rasmussen, C. Gaussian process training with input noise. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Opgenorth, P., Costello, Z., Okada, T., Goyal, G., Chen, Y., Gin, J., Benites, V., de Raad, M., Northen, T. R., Deng, K., et al. Lessons from two design–build–test–learn cycles of dodecanol production in escherichia coli aided by machine learning. *ACS synthetic biology*, 8(6): 1337–1351, 2019.

Radivojević, T., Costello, Z., Workman, K., and Martin, H. G. A machine learning automated recommendation tool for synthetic biology. *Nature communications*, 11(1): 1–14, 2020.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Rasmussen, C. E. *Gaussian Processes in Machine Learning*, pp. 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., focus group, T. E. M. L., Harrow, J., Psomopoulos, F. E., and Tosatto, S. C. E. Dome: Recommendations for supervised machine learning validation in biology, 2021.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Wilson, J., Hutter, F., and Deisenroth, M. Maximizing acquisition functions for bayesian optimization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Zhang, M. and Ong, C. S. Quantile bandits for best arms identification with concentration inequalities. *International Conference on Machine Learning*, 2021.

Zhang, M., Holowko, M. B., Hayman Zumpe, H., and Ong, C. S. Machine learning guided design for ribosome binding site. *Under Review*, 2021.