Deep learning for optimization of protein expression


Digital Object Identifier (DOI): 10.1016/j.copbio.2023.102941

Link: Link to publication record in Edinburgh Research Explorer

Document Version: Publisher's PDF, also known as Version of record

Published in: Current opinion in biotechnology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Deep learning for optimization of protein expression
Evangelos-Marios Nikolados and Diego A Oyarzún

Recent progress in high-throughput DNA synthesis and sequencing has enabled the development of massively parallel reporter assays for strain characterization. These datasets map a large number of DNA sequences to protein expression levels, sparking increased interest in data-driven methods for sequence-to-expression modeling. Here, we highlight advances in deep learning models of protein expression and their potential for optimizing strains engineered to produce recombinant proteins. We review recent works that built highly accurate models and discuss challenges that hinder adoption by end users. There is a need to better align this technology with the constraints encountered in strain engineering, particularly the cost of acquiring large amounts of data and the requirement for interpretable models that generalize beyond the training data. Overcoming these barriers will help to incentivize academic and industrial laboratories to tap into a new era of data-centric strain engineering.

Addresses
1 School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, UK
2 School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
3 The Alan Turing Institute, London NW1 2DB, UK

Corresponding author: Oyarzún, Diego A (d.oyarzun@ed.ac.uk).

Current Opinion in Biotechnology 2023, 81:102941
This review comes from a themed issue on Systems Biology
Edited by Howard Salis

Available online xxxx
https://doi.org/10.1016/j.copbio.2023.102941
0958-1669/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction
Production of recombinant proteins is a central goal in microbial engineering. A quantitative understanding of the relation between DNA sequence and protein expression is key for designing robust and predictable strains. Although the design of nucleotide sequences with bespoke properties is a long-standing goal of synthetic biology, and a prerequisite for accelerating the design–build–test cycle, predictive design is notoriously challenging because it relies on the ability to predict phenotype from genotype specifications.

Owing to recent progress in batch DNA synthesis and sequencing (Box 1), a number of strategies for massively parallel reporter assays have produced datasets with thousands and even millions of genotype–phenotype associations. There is a growing interest in using such large data for building sequence-to-expression models that predict protein expression directly from nucleotide sequence. Deep learning models, in particular, have found success in a range of applications and are attracting substantial interest from the synthetic biology community. Once trained, such models can be queried in silico to infer relations between sequence and expression levels, which enables the rational design of sequences with improved phenotypes [1,2].

Here, we discuss recent progress in deep learning models for sequence-to-expression prediction. Our focus is on phenotype prediction from libraries of short, non-coding, constructs typically employed for optimizing protein yield via control of transcriptional and translational efficiency. These include regulatory sequences such as promoter operator sites, ribosomal binding sequences, and other genetic elements (Figure 1a). We exclude the large body of work on variant effect prediction [3,4] or whole protein-coding sequences [5], which have been reviewed extensively elsewhere [6–8].

Sequence-to-expression models
The traditional approach to strain engineering relies on phenotyping libraries of sequence variants, and selecting a subset of top producers for further downstream validation, scale-up, or iterative design (Figure 1b). Recent years have witnessed the birth of new model-guided strategies that employ sequence-to-expression predictors that learn the shape of the phenotypic landscape (Figure 1c). Such models can then be queried iteratively within a sequence optimization loop [9,10].

Among various strategies for sequence-to-expression modeling [11], deep learning has rapidly emerged as a promising technology for building highly accurate predictors. At their core, deep learning models are data regressors that predict a continuous variable from a set of inputs. As in any regression problem, key to building accurate sequence-to-expression models is the a priori knowledge of sequence features that correlate with...
expression. Although there is a rich literature on the genetic determinants of gene expression in various contexts [12–15], such relations do not typically have the predictive power required by synthetic biology applications [16]. Purely data-driven models have emerged as an effective alternative for prediction of heterologous expression, as they can detect highly nonlinear correlations between sequence and phenotype. Though such correlations do not have a mechanistic basis, they are useful for prediction, particularly because such regressors can be incorporated into search algorithms to find sequences with increased expression levels [17–19].

Deep learning models made early successes in predicting many phenotypes such as transcription factor binding affinity [20], ribosome loading [21], and RNA splicing [22]. An early work in the field employed deep learning to predict expression from 500,000 untranslated regulatory sequences in BY4741 Saccharomyces cerevisiae grown in synthetic defined media lacking histidine and leucine [23]. A number of more recent works in the synthetic biology community have built models for a variety of design tasks that use short genetic elements to control heterologous expression. For example, a large dataset with over 100,000,000 synthetic yeast promoter sequences using Y8205 cells grown in rich media (YPD) using glucose, glycerol, or galactose as feedstock was employed to train deep learning models and study how transcription factors interpret *cis*-regulatory sequences to control gene expression [24]. The same year, another work built an ensemble of convolutional neural networks (CNNs) to predict expression from approximately 1000 promoter sequences, derived from two natural promoters with varied transcription factor binding sites, in
S. cerevisiae grown in YNB-U medium. Crucially, the work implemented a model-guided sequence optimization strategy (Figure 1c) to generate large and sequence-diverse promoter sets [25].

Recently, the work by Vaishnav et al. explored in detail the relationship between promoter sequence, expression phenotype, and fitness, providing a framework for designing regulatory sequences and addressing a range of questions on the evolution of regulatory sequences [1]. They generated a large genotype–phenotype dataset with more than 20 000 000 randomly generated promoter sequences, and their expression levels in Y8205 S. cerevisiae measured in complex (YPD) and synthetic media lacking uracil (SD-Ura). The dataset was employed to train CNNs that capture fitness landscapes and generalize beyond the sequence space employed for training. This allowed to use model predictions as a surrogate for a fitness function in molecular evolution studies.

Beyond promoter sequences, several studies have focused on prediction of expression from libraries designed to control translational processes. Höllerer and colleagues combined phenotypic recordings with deep studies. Studies relative to the measured variation across biological replicates. Lower requirements on model accuracy may

**Escherichia coli** strain was cultivated in lysogeny broth (LB) [26]. They used a site-specific recombinase to record the effect of gene regulatory elements on DNA, enabling readout of both sequence and translation kinetics of over 300 000 E. coli ribosomal binding sites (RBS). This resulted in a high-resolution dataset to train an ensemble of residual neural networks (Resnet) that predicted RBS activities and quantified the uncertainty of prediction. A different work by Angenent-Mari et al. demonstrated the predictive power of deep learning models trained on RNA toehold switches. Using a dataset with over 100 000 sequences in BL21 E. coli grown in LB medium, they trained deep neural networks that outperformed traditional thermodynamic models. A key innovation was the use of a nucleotide complementarity matrix representation to visualize the learned patterns of RNA secondary structure, which were employed to identify success and failure modes of different designs. A related work [27] built CNNs and language models to re-engineer poorly performing riboregulators.

A recent work explored the relation between model accuracy and data efficiency in a library of ~200 000 UTRs in recA-deleted MDS42 E. coli grown in rich MOPS medium [2]. Unlike previous studies, this library was not randomized but built with a design-of-experiments approach that balanced coverage and depth of the sequence space [14]. The work trained a large panel of machine learning models of varied complexity, including nondeep and deep learning models, revealing that accurate models can be trained on as few as a couple of thousand variants. Moreover, using expression screens in E. coli and S. cerevisiae, they showed that controlled sequence diversity can improve predictive performance across larger regions of the sequence space with important gains in data efficiency.

**Discussion**

The growing interest in deep learning for strain design results from the availability of large screens of genotype–phenotype associations (Box 1). Recent works have demonstrated that accurate prediction together with sequence optimization can substantially accelerate the design of biological circuitry. However, the current literature also reveals a need to better align this technology with the tasks, resources, and constraints encountered in microbial engineering. A number of domain-specific challenges, such as the high cost of data acquisition or the inherent variability of biological data, have not been considered explicitly in the literature and prevent the adoption of deep learning by end users. Next, we explain these gaps in more detail.

**Low-N prediction**

Current literature shows a strong trend toward highly accurate models trained on large datasets (Figure 2a). This trend appears to be inherited from the machine learning field, where high-scoring models are preferred even if they require very large datasets for training. The rationale behind such large screens is to unbias the genotypic space and expand the high-confidence regions of the models, for example, by including weak regulatory elements that would normally be excluded from small-scale datasets. But in microbial engineering, the cost of acquiring such large datasets is beyond reach for most laboratories. While further progress in laboratory automation [33] or reductions in DNA synthesis and sequencing costs may lower this entry barrier in the future, in the short term, there is a need to develop ‘low-N’ predictors that can help incentivize the adoption of machine learning by end users. One route for improvement is the development of low-dimensional sequence representations that correlate well with expression [34], as this would enable training models with smaller datasets.

Moreover, expression measurements display biological and technical variation that place a hard ceiling on the accuracy that should be expected from deep learning models. It is somewhat surprising that sequence-to-expression models are normally scored with absolute metrics of accuracy, as it is common practice in other applications of machine learning, instead of using metrics relative to the measured variation across biological replicates. Lower requirements on model accuracy may
Deep learning for sequence-to-expression modeling. (a) Recent sequence-to-expression models using machine learning. Models trained on datasets of varying sizes and can achieve good prediction accuracy for a variety of constructs. Models shown are random forests [2] (a type of nondeep model), CNN [1,2,25], Resnet [26], multilayer perceptron [35], and transformer architectures [1]. Model accuracy is reported in terms of the coefficient of determination ($R^2$) between predictions and ground truth in test sets in each paper. (b) Typical machine learning pipeline. Sequence and phenotypic data are employed to train a range of regressors, which are benchmarked in terms of their quality of fit cross-validated on suitable datasets to prevent overfitting and aid with model selection. Model accuracies in panel A are presented for qualitative comparison only; each model was trained on different datasets with varied construct length, strains, and growth media. Accuracy for the model in Ref. [1] was computed with the approximation $R^2 = r^2$, where $r$ is the Pearson correlation coefficient reported in the original paper.

Data sources: Fig. 1b in Ref. [1] using S. cerevisiae Y2805 grown in YPD and SD-Ura; Fig. 3c, S5 in Ref. [2] using E. coli MDS42recA grown in rich MOPS; Fig. 2a in Ref. [25] using S. cerevisiae CSY3 (W303 MATa) with the P<sub>GPD</sub> promoter; Fig. 4a in Ref. [26] using E. coli TOP10ΔraaA grown in LB; Fig. 4a in Ref. [36] using E. coli BL21 grown in LB.

Model generalization across laboratories

The data required for training sequence-to-expression models are typically collected in few growth conditions and using lab-specific experimental setups and strains. A key caveat of this approach is that it prevents models to generalize predictions for sequences that have been phenotyped by different laboratories, which likely employ different growth conditions, culture volumes, expression hosts, and varied range of other experimental design decisions. As a result, laboratories must often train their own models using in-house data. This is inefficient and causes models to be poorly interoperable across laboratories. Moreover, such models cannot be employed to predict expression levels in growth conditions different from those employed for training, which can severely limit their scope of applicability, particularly because protein expression can be highly dependent of growth media.

The success of AlphaFold2 highlighted how highly standardized datasets, such as the structural data in the Protein Data Bank, can enable the construction of widely applicable models [38]. In the case of genotyp–phenotype data, the innate biological variation and the current lack of standardization protocols for data acquisition, prevent the construction of such general-purpose models. A potential solution may emerge from recent progress in the field of natural language processing [39], where the use of pretrained models on large datasets has shown an exceptional ability to generalize to new data by fine-tuning and retraining with small, task-
specific, data. Such approaches have already shown some promise for genomic data [40].

**Interpretability**

Purely data-driven models may be sufficient for a specific design task, but in strain engineering, designers often need to identify and interpret the specific molecular processes that contribute to the observed phenotype. This is particularly challenging for prediction of protein expression levels, as these depend on the joint action of transcription, translation, and their multiple interconnected regulatory mechanisms. It is unclear if individual models for each of these processes, all of which require bespoke screening data, can be coupled into larger predictive models. Moreover, deep learning models are notoriously difficult to interpret. Classic methods based on motif detection have been employed in conjunction with sequence-to-expression models [23], and recently the field of Explainable AI [41] has produced many scoring methods that have been employed to quantify the importance of input sequences [2,26]. Other methods borrowed from imaging applications (e.g. saliency maps) have also shown some success in explaining the output of sequence-to-expression models [35]. However, interpretability methods have generally been employed on a case-by-case basis, and their ability to provide insights for design is still limited.

The poor interpretability of machine learning algorithms contrasts with some of the thermodynamic models that have been widely adopted by the community for sequence-based prediction [42]. Such models can incorporate sequence information and contain sufficient mechanistic details for meaningful biological interpretation. Most recently, the newly developed Promoter Calculator demonstrated the benefits of combining such thermodynamic models with machine learning [11]. Thermodynamic models, however, require substantial effort to build and they are difficult to retrain to data from in-house strains or growth conditions employed by end users. A number of whole-cell models have provided mechanistic descriptions of microbial physiology [43,44]. These have a remarkable ability to reproduce wild-type phenotypes in various growth conditions, as well as various modes of failure for synthetic constructs [45], but they are unable to account for sequence information. A notable exception is the whole-cell *E. coli* model [46], but the large efforts required for its construction and simulation make it impractical for design tasks [47]. Overall, there is a strong need for sequence-to-expression models that combine the best of the three approaches: the flexibility of user-trainable machine learning models, the mechanistic interpretability of thermodynamic models, and physiological insights of whole-cell models.

**Conclusions**

Deep learning methods are being adopted by most disciplines in synthetic biology, including, for example, cell-free systems [48], microbial communities [49], and metabolic engineering [50,51]. Protein production is no exception and recent work has led to several sequence-to-expression models that map DNA sequences to protein expression levels. These models promise to revolutionize the design–build–test cycle with phenotypic predictions that can be directly linked to sequence design. Moreover, such models can be trained to in-house data, and provide flexibility to incorporate historical data and refine predictions using iterative design approaches.

Such progress has largely followed the trends in the general AI and machine learning community, with a focus on complex models trained on large datasets. Current sequence-to-expression models have already demonstrated their ability to deliver highly accurate predictions. The next wave of deep learning models should explicitly address the inherent challenges of biological data and the needs of end users in microbial engineering. Budgetary constraints impose hard limits on the size of genotype–phenotype screens, and few academic or industrial laboratories can afford to acquire large data solely for the purpose of model training. One way forward is to prioritize the design of training data over the design of model architectures. This would mark a departure from the general approach in machine learning, and offer exciting opportunities for the development of application-specific tools for predictive modeling. Moreover, given that phenotype data are difficult to transfer across laboratories due to the use of different strains, growth conditions, or measurement hardware, the more widespread adoption of deep learning may ultimately rest on local models that can be trained with in-house data of feasible size. Failure to do so risks an ‘AI winter’ that has pervaded other disciplines in the past, and misses an opportunity for laboratories that could otherwise benefit greatly from this technology. While machine learning models are not a replacement for mechanistic approaches, the ability to acquire high-throughput strain characterization data necessitates fresh approaches to extract actionable insights from such large data, a task for which deep learning is particularly well-suited.

**Data Availability**

No data were used for the research described in the article.

**Conflict of interest statement**

The authors declare no conflict of interests.
Acknowledgements

EMN was supported by a doctoral studentship from the Darwin Trust of Edinburgh. DAO was supported by the United Kingdom Research and Innovation (Grant EP/S02431X/1).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


Produced a large genotype–phenotype map in S. cerevisiae and demonstrated that deep learning models can generalize with excellent predictive performance, enabling sequence design for expression engineering and the study of evolutionary principles in gene regulation.


Systematic mapping of the relation between data size, sequence diversity, and the choice of machine learning models. It demonstrates that nondeep and deep learning models can achieve good prediction accuracy with much smaller datasets than previously assumed, and shows that controlled sequence diversity can provide important gains in data efficiency.


Shows that a generative adversarial network trained on entire regulatory elements of natural sequences from S. cerevisiae can generate valid deco novo regulatory DNA with highly accurate target expression levels.


Combines biophysics and machine learning into an interpretable model of transcription initiation for automated design of regulatory elements. The model was shown to accurately predict transcription rates of thousands of bacterial promoters, and demonstrated the benefits of combining machine learning and thermodynamics.


Comprehensive review of modern sequencing technologies and how they help generating large datasets for a wide array of applications.


Built a CNN-based regressor and a language model-based classifier to uncover design rules in toehold switches. They provide examples of utilizing the sequence-to-function trained models to characterize toeholds that sense pathogenic genomes and to re-engineer poorly-performing designs.


