Evaluation and optimization of sequence-based gene regulatory deep learning models

Abdul Muntakim Rafi1, Dmitry Penzar2,3, Daria Nogina3, Dohoon Lee4, Eesht Dhaval Vaishnav5, Danyeong Lee4, Nayeon Kim4, Sangyeup Kim4, Georgy Meshcheryakov6, Andrey Lando7, Payman Yadollahpour5, Arsenii Zinkevich2,3, Dohyeon Kim4, Yeojin Shin4, Il-Youp Kwak8, Byeong-Chan Kim8, Juhyun Lee8, Random Promoter DREAM Challenge Consortium, Sun Kim4, Aviv Regev5, Jake Albrecht9, Wuming Gong10, Ivan V. Kulakovskiy2,6, Pablo Meyer11, Carl de Boer1

1 University of British Columbia, Vancouver, BC, Canada
2 Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
3 Lomonosov Moscow State University, Moscow, Russia
4 Seoul National University, Seoul, South Korea
5 Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia
6 Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia
7 Yandex N. V., Moscow, Russia
8 Chung Ang University, Seoul, Republic of Korea
9 Sage Bionetworks
10 University of Minnesota, Minneapolis, United States
11 Center for Computational Health, IBM Research

*Correspondence: carl.deboer@ubc.ca, abdulmuntakim.rafi@ubc.ca

Abstract

Neural networks have proven to be an immensely powerful tool in predicting functional genomic regions, in particular with many recent successes in deciphering gene regulatory logic. However, how model architecture and training strategy choices affect model performance has not been systematically evaluated for genomics models. To address this gap, we held a DREAM Challenge where competitors trained models on a dataset of millions of random promoter DNA sequences and corresponding experimentally determined expression levels to best capture the relationship between regulatory DNA and gene expression in yeast. To robustly evaluate the models, we designed a comprehensive suite of benchmarks encompassing various sequence types. While some benchmarks produced similar results across all models, others differed substantially. For some sequence types, model performances exhibited correlation scores as high as 0.98, while for others, substantial improvement is still required. The top-performing models were all neural networks, which demonstrated substantial performance gains by customizing model architectures to the nature of the experiment and utilizing novel training...
strategies tailored to genomics sequence data. Overall, our DREAM Challenge highlights the need to benchmark genomics models across different scenarios to uncover their limitations.

Introduction

In eukaryotes, transcription factors (TFs) play a crucial role in regulating gene expression and are critical components of the cis-regulatory mechanism (1). TFs compete with nucleosomes and each other for DNA binding and can enhance each other's binding through biochemical cooperativity and mutual competition with nucleosomes (2). Cis-regulatory complexity grows with the square of the number of TFs involved due to the number of potential pairwise interactions between TFs. Since yeast has about 8.5x fewer TFs than humans (~200 vs. ~1700) (1,3), the cis-regulatory code is theoretically about ~72x less complex in yeast, making it an ideal system to test our understanding of eukaryotic cis-regulation. While the field has made substantial progress in characterizing regulatory mechanisms (4–9), a quantitative understanding of cis-regulation remains a major challenge.

Neural Networks (NNs) have shown immense potential in deciphering gene regulation. While different network architectures, such as CNNs (4,5,7,10), RNNs (11), and transformers (8,12), have been used to create genomics models, there is limited research on how NN architectures and training strategies affect their performance. Standard datasets provide a common benchmark to evaluate and compare algorithms, leading to improved performance and continual progress in the field (13). For instance, the computer vision and natural language processing (NLP) fields have seen a continual improvement of NNs facilitated by standard datasets, such as the ImageNet data (13), MS COCO (14), etc. In contrast, genomics models are often created ad hoc for analyzing a specific dataset, and it remains unclear whether a model's improved performance results from improved model architecture or training data. In many cases, the models created are not directly comparable to previous models due to substantial differences in the underlying data used to train and test the models.

We organized the Random Promoter DREAM Challenge (15) to address the lack of standardized evaluation and continual improvement of genomics models. Here, we asked the participants to design sequence-to-expression models and train them on expression measurements of promoters with random DNA sequences. The models would receive regulatory DNA sequence as input and use it to predict the corresponding gene expression value. We designed a separate set of sequences to test the limits of and provide insight into model performance. Our evaluation across various benchmarks revealed that, for some sequence types, model performances are approaching the previously-estimated inter-replicate experimental reproducibility for this datatype (6), while considerable improvement remains necessary for others. The top-performing solutions in the challenge exceeded performance of all previous state-of-the-art models for similar data and demonstrated that designing NNs inspired by the state-of-the-art models from computer vision and NLP, incorporating the nature of the experiment into NN design, properly tuning hyperparameters, and novel training strategies that are more suited to genomics sequence data can result in considerable performance gains.
Summary of the challenge and the data

<table>
<thead>
<tr>
<th>Promoter sequence</th>
<th>Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>...CAATTGCTGCCA...</td>
<td>![Exp. Icon]</td>
</tr>
<tr>
<td>...CGCGAATTATCA...</td>
<td>![Exp. Icon]</td>
</tr>
<tr>
<td>...TTCAGATGCATC...</td>
<td>![Exp. Icon]</td>
</tr>
<tr>
<td>...TACTATACCGAA...</td>
<td>![Exp. Icon]</td>
</tr>
</tbody>
</table>

6,739,258 sequence expression pairs

Participants develop sequence to expression models

Evaluate models using designed sequences with biological insight

Figure 1: Overview of the challenge. Competitors received a training dataset of random promoters and corresponding expression values (left). They continually refined their models and competed for dominance in a public leaderboard (middle). At the end of the challenge, they submitted a final model for evaluation (right) using a test dataset consisting of eight sequence types: (i) high expression, (ii) low expression, (iii) native, (iv) random, (v) challenging, (vi) SNVs, (vii) motif perturbation, and (viii) motif tiling.

To generate the competition training data, we conducted a high-throughput experiment to measure the regulatory effect of millions of random DNA sequences. Prior research has shown that random DNA can display activity levels akin to genomic regulatory DNA, due to the coincidental occurrence of numerous TF binding sites (6,12,16). As a result, leveraging a vast array of random regulatory sequences enables the learning of intricate regulatory mechanisms from the ground up, without succumbing to biases associated with overrepresented genomic sequences.

Here, we cloned 80 bp random DNA sequences into a promoter-like context upstream of a yellow fluorescent protein (YFP), transformed the resulting library into yeast, grew the yeast in chardonnay grape must, and measured expression by fluorescent activated cell sorting and sequencing, as previously described (6,17,18) (Methods). This resulted in a training dataset of 6,739,258 random promoter sequences and their corresponding mean expression values. We provided these data to the competitors, who could use them however they like to train their model. However, competitors were not allowed to use external datasets in any form to ensure that everyone was training models on the same dataset. Ensemble predictions were also disallowed as they would almost certainly provide a boost in performance but without providing any insight into the best model types and training strategies.

We evaluated the models on a set of “test” sequences designed to probe the predictive ability of the models in different ways. The measured expressions of these sequences were quantified in the same way as the training data but in a separate experiment with more cells sorted per sequence (~100), yielding more accurate estimated expression levels compared to the training data measurements and providing higher confidence in the challenge evaluation. The test set consisted of 71,103 sequences from several promoter sequence types. We included both random sequences and sequences from the yeast genome to get an estimate of performance.
difference between the random sequences in the training domain and the evolved sequences. We also included sequences designed to capture known failure modes of previous models trained on similar data, namely sequences at the high/low extremes of expression and sequences designed to maximize the disagreement between a previously developed CNN and a physics-informed neural network ("biochemical model") (6,12). We previously found that predicting changes in expression between closely related sequences (i.e., nearly identical DNA sequences) is substantially more challenging, and so we included subsets where models had to predict changes that result from single nucleotide variants (SNVs), perturbations of specific TF binding sites, and tiling of TF binding sites across background sequences. Each test subset was given a different weight proportional to the number of sequences in the set and how important we considered it to be (Table 1). For instance, predicting the effects of SNVs on gene expression is a critical challenge for the field due to its relevance to complex trait genetics (19), and so many such sequence pairs were included and given the highest weight in the test set. Within each sequence subset, we determined model performance using Pearson $r^2$ and Spearman $\rho$, which captured the linear correlation and monotonic relationship between the predicted and measured expression levels (or expression differences), respectively. The weighted sum of each performance metric across test subsets yielded our two final performance measurements, which we call *Pearson Score* and *Spearman Score*.

The competition ran for 12 weeks in the summer of 2022 and included two evaluation stages: the public leaderboard phase and the private evaluation phase (Fig. 1). The leaderboard opened six weeks into the competition and allowed teams to submit up to 20 predictions on the test data per week. At this stage, we used 13% of the test data for leaderboard evaluation and displayed only the overall Pearson $r^2$, Spearman $\rho$, *Pearson Score*, and *Spearman Score* to the participants, while keeping the performance on the promoter subsets and the specific sequences used for the evaluation hidden. The participating teams achieved increasing performance each week (Supplementary Fig. 1), showcasing the effectiveness of such challenges in motivating the development of better machine learning models. Over 110 teams from all over the globe competed in this stage. At the end of the challenge, 28 teams submitted their models for final evaluation. We used the rest of the test data (~87%) for the final model evaluation.

**Top-performing submissions in the challenge**

The top-performing submissions were all NNs. Despite recent findings suggesting the prominence of attention-based architectures (12), only one of the top five submissions in the challenge used transformers, placing 3rd. The best-performing submissions were dominated by fully convolutional NNs, with 1st, 4th, and 5th places taken by them. The best-performing solution was based on the EfficientNetV2 architecture (20,21), and the 4th and 5th solutions were based on the ResNet architecture (22). Moreover, all teams used convolutional layers as the starting point in their model design. A recurrent neural network (RNN) with bidirectional long-short-term memory (Bi-LSTM) layers (23,24) placed 2nd. Overall, the teams converged on many similar training strategies (e.g., using Adam (25) or AdamW (26) optimizers), but had substantial differences as well (Table 1).

The competing teams introduced several innovative approaches to solve the expression prediction problem. Autosome.org, the best-performing team, transformed the task into a soft-classification problem by training their network to predict a vector of expression bin probabilities,
which was then averaged to yield an estimated expression level, effectively recreating how the data were generated in the experiment. They also used a distinct data encoding method by adding additional channels to the traditional 4-d one-hot encoding of the DNA sequence used by most teams. The two additional channels indicated (i) whether the sequence provided as input was likely measured in only a single cell (which results in an integer expression value), and (ii) whether the input sequence is being provided in the reverse complement orientation. Furthermore, Autosome.org's model, with only two million parameters, challenged the current trend of designing deep NNs with numerous parameters, demonstrating that efficient design can considerably reduce the necessary number of parameters. Autosome.org and BHI were distinct in training their final model on the entirety of the provided training data (i.e., no sequences withheld for validation) for a prespecified number of epochs (determined previously using cross-validation using validation subsets). Unlock_DNA, the 3rd team, took a novel approach by randomly masking 5% of the input DNA sequence and having the model predict both the masked nucleotides and gene expression. This approach used the masked nucleotide predictions as a regularizer, adding a reconstruction loss to the model loss function, which stabilized the training of their large NN. BUGF, the 9th team, used a somewhat similar strategy where they randomly mutated 15% of the sequence and calculated an additional binary cross-entropy loss predicting whether any bp in the sequence had been mutated. However, BUGF predicted expression using the original unchanged sequence, whereas Unlock_DNA predicted expression using the masked sequence. The 5th team, NAD, employed GloVe (27) to generate embedding vectors for each base position and used these vectors as inputs for their NN, whereas the other teams utilized traditional one-hot encoded DNA sequences. Two teams, SYSU-SAIL-2022 (11th) and Davuluri lab (16th), attempted to train DNA language models (28) on the challenge data, which involved pretraining a BERT language model (29) on the challenge data and subsequently using the BERT embeddings to train an expression predictor. SYSU-SAIL-2022’s pretraining strategy differed from Davuluri lab’s, as they only used the top 20% of sequences in terms of expression. However, the predictors used by these two teams were not identical, and it is not possible to definitively conclude that SYSU-SAIL-2022’s pretraining strategy was superior for training language models on the random sequence space. Due to the varied network architectures and training strategies (Table 1), determining the exact factors that led to the success of the top submissions remains a substantial challenge.

Table 1: Breakdown of the top-performing models into key components

<table>
<thead>
<tr>
<th>Participant team name</th>
<th>NN architecture type</th>
<th>DNA Encoding method</th>
<th>Input flanking region length</th>
<th>Usage of one-hot encoded DNA sequences during training</th>
<th>Train validation split</th>
<th>Number of parameters (in millions)</th>
<th>Optimizer</th>
<th>Loss function</th>
<th>Learning Rate Scheduler</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autosome.org</td>
<td>CNN (EfficientNetV2 (21))</td>
<td>OHE</td>
<td>70</td>
<td>Data aug. (additional channel) + Model (additional channel)</td>
<td>100-0</td>
<td>1.9</td>
<td>AdamW (26)</td>
<td>Kullback-Leibler divergence</td>
<td>One Cycle LR</td>
<td>$r, p^*$</td>
</tr>
<tr>
<td>BHI</td>
<td>CNN + RNN (Bi-LSTM) (23)</td>
<td>OHE</td>
<td>10</td>
<td>Post-hoc conjoined setting (30)</td>
<td>100-0</td>
<td>6.8</td>
<td>AdamW (26)</td>
<td>Huber</td>
<td>Cosine Anneal LR</td>
<td>$r, p^*$</td>
</tr>
<tr>
<td>Participant team name</td>
<td>NN architecture type</td>
<td>Encoding method</td>
<td>Input flank region length</td>
<td>Usage of reverse strand during model training</td>
<td>Train validation split</td>
<td>Number of parameters (in millions)</td>
<td>Optimizer</td>
<td>Loss function</td>
<td>Learning Rate Scheduler</td>
<td>Metric</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------</td>
<td>------------------------------------</td>
<td>----------</td>
<td>----------------</td>
<td>------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Unlock_DNA Transformer</td>
<td>OHE</td>
<td>20</td>
<td>Input to model (concat. with forward strand)</td>
<td>95-5</td>
<td>Adam (25)</td>
<td>MSE + custom</td>
<td>One Cycle LR</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Camformers CNN (ResNet (22))</td>
<td>OHE</td>
<td>30</td>
<td>None</td>
<td>90-10</td>
<td>AdamW (26)</td>
<td>L1</td>
<td>Reduce LR On Plateau</td>
<td>r, p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAD CNN + Transformer</td>
<td>Glove (27)</td>
<td>0</td>
<td>None</td>
<td>90-10</td>
<td>AdamW (26) + GSAM (31)</td>
<td>smooth L1</td>
<td>Linear LR</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wztir CNN (ResNet (22))</td>
<td>OHE</td>
<td>62</td>
<td>Input to model (concat. with forward strand)</td>
<td>99-1</td>
<td>Adam (25)</td>
<td>MSE</td>
<td>Reduce LR On Plateau</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Schoolers Are All You Need (High Schoolers) CNN + Transformer + MLP</td>
<td>OHE</td>
<td>31</td>
<td>Model (RC parameter sharing) (30)</td>
<td>98-2</td>
<td>Adam (25) + SWA (32)</td>
<td>MSE</td>
<td>Multi Step LR</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioNML Vision Transformer (33)</td>
<td>OHE</td>
<td>30</td>
<td>Model (RC parameter sharing) (30)</td>
<td>86-14</td>
<td>Adamax (25) + L2 regularizer</td>
<td>Huber</td>
<td>Multi Step LR</td>
<td>r, CoD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUGF Transformer Embedding</td>
<td>OHE</td>
<td>32</td>
<td>None</td>
<td>94-6</td>
<td>RAdam (34)</td>
<td>Multi-label focal loss (35) + custom</td>
<td>None</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mt GRU (36) + CNN</td>
<td>OHE</td>
<td>62</td>
<td>Model (RC parameter sharing) (30)</td>
<td>99.8-0.2</td>
<td>Adam (25)</td>
<td>binary cross.</td>
<td>None</td>
<td>r, CoD*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*These teams employed the metrics in a cross-validation setting to determine the optimal number of epochs for their models and ultimately saved the model weights after running for the n epochs, without relying on validation metric scores. In contrast, other teams utilized validation metric scores to select the best-performing model.

Top two models robustly outperform others

In order to determine the relative performance of the models, we performed a bootstrapping analysis. Here, we sampled 10% of the test data 10,000 times and, for each sample, calculated the performance of each model and the rankings of the models for both Pearson and Spearman Scores (Methods). We averaged the ranks from both metrics to decide their final ranks. Teams Autosome.org and BHI robustly outperformed the others, coming in 1st and 2nd place, respectively (Fig. 2), with Autosome.org coming second to BHI in only 0.1% and 2.25% of the time for Pearson Score and Spearman Score, respectively. In none of the bootstraps did
another team outperform Autosome.org or BHI. Interestingly, Pearson Score and Spearman Score do not agree after the top two teams, indicating that they capture performance in distinct ways. For instance, Unlock_DNA, the 3rd-place team, substantially outperformed 4th-place Camformers by Pearson Score, but performed slightly worse than Camformers by Spearman Score. We consider 5th-place to be a tie between NAD and wztr as their mean ranks (5.81105 and 5.8152, respectively) were very close.

![Figure 2: Bootstrapping provides a robust comparison of the model predictions. (A, B) Distribution of ranks in n=10,000 samples from the test dataset (y-axes) for the top-performing teams (x-axes), for (A) Pearson Score and (B) Spearman Score.](image)

**Figure 2:** Bootstrapping provides a robust comparison of the model predictions. (A, B) Distribution of ranks in n=10,000 samples from the test dataset (y-axes) for the top-performing teams (x-axes), for (A) Pearson Score and (B) Spearman Score.

**Distinct model performances on test subsets**

Analysis of model performance on the different test subsets revealed distinct and shared challenges for each model. We retrained the transformer model proposed by Vaishnav et al. (12), the previous best performing model for this type of data, on the challenge data and used as a reference in the leaderboard ("reference model"). The top submissions’ overall performance was substantially better than the reference model (Fig. 3). Interestingly, the top two models were ranked 1st and 2nd (sometimes with ties) for each test subset regardless of score metric, showcasing that their superior performance cannot be attributed to any single test subset. Further, the rankings within each test subset sometimes differed between Pearson Score and Spearman Score, reinforcing that these two measures capture performance in distinct ways.

Interestingly, models were highly variable in their ability to accurately predict variation within the extremes of gene expression. The cell sorter has a reduced signal-to-noise ratio for the highest and lowest expression levels, and the sorting bin placement can truncate the tails of the expression distribution (6,12). We included these test subsets (high and low) because we previously found that the Vaishnav et. Al model (12) tends to have limited predictive power within each expression extreme. Overall, model performance was most variable across teams in these subsets (e.g., the median difference in Pearson $r^2$, between highest and lowest performance, was 48% for high and low test subsets and 16% for the others, as shown in Fig. 3), suggesting that the challenge models were able to overcome this issue to varying degrees.
Figure 3: Performance of the top-performing teams in each test data subset. (A,B) Model performance (colour and numerical values) of each team (y-axes) in each test subset (x-axes), for (A) Pearson $r^2$ and (B) Spearman $\rho$. Heatmap colour palettes are min-max normalized column-wise. (C,D) Performance disparities observed between the best and worst models (x-axes) in different test subsets (y-axes) for (C) Pearson $r^2$ and (D) Spearman $\rho$. The calculation of the percentage difference is relative to the best model performance for each test subset.

Importantly, the models differed in their abilities to predict the expression levels of promoter sequences from the yeast genome. While the ranking of models is similar for both random and native sequences, the difference in model performance is more substantial for native sequences. We see 17.6% and 9.6% performance difference in Pearson $r^2$ and Spearman $\rho$ between top vs. bottom model for native sequences, respectively, while there are only 5% and 2.7% differences for random sequences (Fig. 3), suggesting that top models have learned more of regulatory grammar that evolution has produced. Although there is likely more experimental noise present in the native promoter data, owing to the propensity of native DNA to include repetitive sequences, which results in lower sequence coverage and poorer quality data (Supplementary Fig. 2), the substantial discrepancy between performance on native and random sequences suggests that there is yet more regulatory logic to learn.
Expression differences between closely related sequences also revealed differences in model performance (Fig. 3, Supplementary Fig. 3). For instance, perturbing TF binding sites (TFBSs) (e.g., mutating sequences strongly matching the cognate motif for an important TF or variable numbers of binding sites added) is a comparatively large perturbation and could be predicted with simple models that capture the binding of these TFs and can count TFBS instances. However, when TFBSs are tiled across a background sequence, the same TFBS is present in every sequence, and the model must have learned how its position affects its activity, as well as capturing all the secondary TFBSs that are created or destroyed as the motif is tiled. Finally, SNVs are even harder to predict because nearly everything about the sequence is identical but for a single nucleotide that may affect the binding of multiple TFs in potentially subtle ways. Accordingly, model performance was highest for Motif Perturbation, intermediate for Motif Tiling, and lowest for SNVs. However, it is important to note that this partly results from smaller changes in expression being more dominated by experimental noise. Nonetheless, the differences in model performance suggest that the top-performing models have better captured the subtleties of cis-regulation since the differences in model performance are more substantial for subtler changes (% difference between best and worst compared to best in Pearson $r^2$ and Spearman $\rho$: 6.5% and 4% for Motif Perturbation, 17.7% and 7% for Motif Tiling, and 14.6% and 9.6% for SNVs; Fig. 3).

Model ensemble provides marginal performance boost

Figure 4: An ensemble of the top-performing model predictions does not improve performance substantially since the top-performing models make similar predictions. (A,B) Minimal improvement in Pearson Score and Spearman score (y-axes) achieved by linear regressors (x-axes) trained on top of the model predictions. (C) The pairwise Pearson correlation (colour) between different pairs of models and the ground truth (x- and y-axes).

We created ensembles of the top-performing models by training a linear regressor with the individual models’ predictions as input. However, the improvement in performance was only
slight in comparison to the individual top model performance (Fig. 4A,B). A pairwise comparison of the models’ predictions on the test data showed a higher correlation between the models to each other than to the ground truth (Fig. 4C, Supplementary Fig. 4). Furthermore, we observed that the correlation remained high for individual test subsets with the exception of two sequence types (Supplementary Fig. 5). Interestingly, these two sequence types represented the extremes of expression space, and the networks might have learned them differently due to the low signal-to-noise ratio present in the training data at these extremes, which is evident by the differing relationships between predicted and actual expression, particularly at the high end of expression (Supplementary Fig. 6). Overall, the results suggest that the models have a similar understanding of the underlying biology but vary in how much of it they have captured. The models may be approaching the extent of cis-regulation that is learnable from the DREAM Challenge training data.

Discussion

The random promoter DREAM Challenge 2022 presented a unique opportunity for participants to propose novel model architectures and training strategies for modeling regulatory sequences. The top-performing models challenged the trend of using increasingly complex architectures by demonstrating that simpler NN models with fewer parameters can effectively capture much of cis-regulation. In particular, 3/5 of the top submissions did not use transformers, including the best-performing team (which also had the fewest parameters of the top 10), illustrating that attention may not be the most efficient way of capturing cis-regulatory logic in short sequences. This emphasizes the need for further exploration of the architectures and training strategies best suited for genomics sequence data.

The varied results across test subsets illustrate the complexity in evaluating cis-regulatory models effectively. For instance, performance on random sequences, which are in the same domain as the training data (also random sequences), was relatively uniform (Fig. 3). Whereas the domain shift to native sequences, where the relative frequencies of different regulatory mechanisms likely differ since these originated through evolution, highlights the disparities between models (Fig. 3). This indicates that a model that excels in modeling overall cis-regulation may still perform poorly for sequences involving certain regulatory mechanisms that are difficult to learn from the training data, leading to incorrect predictions of biochemical mechanisms and variant effects for sequences that use these mechanisms. This emphasizes the importance of multifaceted evaluation of genomics models (37), and designing specific datasets that test the limits of these models.

Methods

Designing the test sequences

High- and low-expression sequences were designed as described in Vaishnav et al.(12). Native test subset sequences were designed by sectioning native yeast promoters into 80 bp fragments (6). Random sequences were sampled from a previous experiment where the tested DNA was synthesized randomly (as in the training data) and quantified (6). Challenging sequences were designed using a combination of a convolutional neural network model (12)
and a biochemical model (a type of physics-informed neural network) (6). Most of the SNVs represent sequence trajectories from Vaishnav et al. (12), but also include random mutations added to random, designed, and native promoter sequences. Motif Perturbation included Reb1 and Hsf1 perturbations. Sequences with perturbed Reb1 binding sites were created by inserting Reb1 consensus binding sites (strong or medium affinity; sense and reverse complement orientations), and then adding between 1 and 3 SNVs to each possible location of each motif occurrence and inserting canonical and mutated motif occurrence into 10 randomly generated sequences at position 20/80. Sequences with Hsf1 motif occurrence were designed by tiling random background sequences with between 1 and 10 Hsf1 monomeric consensus sites (ATGGAACA), added sequentially from both the right and left of the random starting sequences, or added individually within each of the possible 8 positions, or similarly tiling/inserting between 1-5 trimeric Hsf1 consensus sites (TTCTAGAANNTTCT). The Motif Tiling test subset sequences were designed by embedding a single consensus for each motif (poly-A: AAAAA, Skn7: GTCTGGCCC, Mga1: TTCT, Ume6: AGCCGCC, Mot3: GCAGGCACG, and Azf1: TAAAAGAAA) at every possible position (with the motif contained completely within the 80-bp variable region) and orientation for three background sequences as described in de Boer et al. (6).

Table 2: Summary of the test subsets

<table>
<thead>
<tr>
<th>Subset</th>
<th>no. of sequences</th>
<th>Weight in evaluation metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All sequences</td>
<td>71103</td>
<td>1</td>
<td>All sequences in the test data.</td>
</tr>
<tr>
<td>High</td>
<td>968</td>
<td>0.3</td>
<td>Sequences designed to have high expression.</td>
</tr>
<tr>
<td>Low</td>
<td>997</td>
<td>0.3</td>
<td>Sequences designed to have low expression.</td>
</tr>
<tr>
<td>Native</td>
<td>997</td>
<td>0.3</td>
<td>Sequences that are present in the yeast genome.</td>
</tr>
<tr>
<td>Random</td>
<td>6349</td>
<td>0.3</td>
<td>Random DNA sequences.</td>
</tr>
<tr>
<td>Challenging</td>
<td>1953</td>
<td>0.5</td>
<td>Sequences designed to maximize the differences between a convolutional model and a biochemical model trained on the same data.</td>
</tr>
<tr>
<td>SNVs</td>
<td>44340 pairs</td>
<td>1.25</td>
<td>Two sequences that differ by only a single base.</td>
</tr>
<tr>
<td>Motif Perturbation</td>
<td>3287 pairs</td>
<td>0.3</td>
<td>Two sequences that differ due to perturbations to specific known TF binding site.</td>
</tr>
<tr>
<td>Motif tiling</td>
<td>2624 pairs</td>
<td>0.4</td>
<td>Two sequences that differ due to</td>
</tr>
</tbody>
</table>
Quantifying promoter expression

High complexity random DNA libraries that comprised the training data were created as described previously (6), with the following modifications. The random promoter library in E. coli theoretically contained about 74 million random promoters and was transformed into S288c (delta URA3) yeast yielding 200 million transformants, which were selected in SD-Ura media. 1 L of chardonnay grape must (filtered) was inoculated with the pool to an initiate OD600 of 0.05 and grown at room temperature without continual shaking, with the culture diluted as needed with fresh chardonnay grape must to maintain OD below 0.4, for a total growth time of 48 hours and having undergone >5 generations. Prior to each OD, the culture was gently agitated to decarbonate it, waiting for the resulting foam to die down before agitating again, and continuing until no more bubbles were released. Yeast were then sorted and sequencing libraries prepared as described previously (6), with sequencing libraries pooled and sequenced on an Illumina NextSeq.

Data processing for both N80 (training) and designed (test) libraries were done as described before (6), except that the expression levels of the designed library sequences were estimated using MAUDE (38), using the read abundance in each sorting bin as input, and estimating the initial abundance of each sequence as the average relative abundance of that sequence across all bins.

Competition rules

1. Only the provided training data could be used to train models. Models had to train from scratch without any pre-training on external datasets to avoid overfitting to sequences present in the test data (e.g. some sequences in the test data are derived from extant yeast promoters).

2. Reproducibility was a prerequisite for all submissions. The participants had to provide the code and instructions to reproduce their models. We retrained the top-performing solutions to validate their performance.

3. Augmenting the provided training data was allowed. Pseudo-labeling the provided test data was not allowed. Using the test data for any purpose during training was not allowed.

4. Ensembles were not allowed.

Performance evaluation

We calculated the Pearson $r^2$ and Spearman $\rho$ between predictions and measurements for each test subset to capture performance on each sequence subset. We used Pearson $r$, which captures the linear correlation between predictions and measured expression levels while being robust to the scaling differences that occur between training and test sequences. As the training and the test data are not on the same scale but are linearly related by, $y = mx + b$, where $m$
and $b$ are constants that are affected by sorting bin placement during the experiment, Pearson $r$ was an appropriate metric for us. We also used Spearman $\rho$, which captures the monotonic relation between predictions and measured expression levels while being robust to outliers. The weighted sum of each performance metric across promoter types yielded our two final performance measurements, which we call **Pearson Score** and **Spearman Score**.

\[
\text{Pearson Score} = \frac{\sum_{i=0}^{\text{subsets}} w_i \cdot \text{Pearson}R^2_i}{\sum_{i=0}^{\text{subsets}} w_i} \\
\text{Spearman Score} = \frac{\sum_{i=0}^{\text{subsets}} w_i \cdot \text{Spearman}_i}{\sum_{i=0}^{\text{subsets}} w_i}
\]

Here, $w_i$ is the weight used for the $i$-th test subset. $\text{Pearson}R^2_i$ and $\text{Spearman}_i$ are respectively the square of Pearson coefficient and the Spearman coefficient for sequences in $i$-th subset.

### Description of the approaches used by the participants

In this section, we present an overview of the approaches employed by the participants in the challenge. For the top-performing teams, we provide a detailed description of their methodologies, while for the remaining teams, we offer a concise overview without repeating any details that have already been discussed for other teams. The performance of each approach is illustrated in **Supplementary Fig. 7** for easier comparison.

**Autosome.org**: The team reformulated the initial regression task as a soft-classification problem by replacing the initial target expression value with a vector of 18 probabilities corresponding to individual bins, assuming that they can be deduced from the normal distribution with mean and variance equal to (expression + 0.5) and 0.5, respectively. To obtain a predicted expression value for a sequence during the validation step, the predicted probabilities were multiplied by their bin numbers. They used one-hot encoding for the promoter sequences, where they added a separate binary channel explicitly marking the objects with integer (and thus likely imprecise) expression measurements. They also augmented the dataset with the reverse complementary sequences and added a separate binary channel to denote the supplied strand explicitly (forward or reverse complementary). The proposed model was based on a fully-convolutional network inspired by EfficientNetV2 (21). The following architectural choices contributed to the final performance: (i) grouped convolution (39) instead of the depthwise convolution of the original EfficientNetV2, (ii) the standard residual blocks were substituted with residual channel-wise concatenations, (iii) a bilinear layer was inserted in the middle of the EfficientNetV2 SE-block.

**BHI**: Their approach adopts a “sandwich” architecture consisting of a one-dimensional convolutional layer, a bidirectional long-short-term memory (Bi-LSTM) layer, and another convolutional layer. Each convolutional layer used different kernel sizes. Besides the model architecture, the team found that training details specialized for DNA sequence-based deep learning models were highly important for the overall performance. Among them, the most crucial was to use a ‘post-hoc conjoined’ setting (30), which imposes a reverse-complement equivariance to the model. Test-time augmentation was also effective. Predictions were made for an original sequence, its four shifted variants (generated by -2bp, -1bp, +1bp, and +2bp shifting), and their reverse-complement sequences, then those 10 predictions were averaged to make a final prediction. While training sequences over 110bp were trimmed to the right, sequences shorter than 110bp were randomly padded with the original vector sequences on both sides. This informative padding gave a nonnegligible performance boost. Finally, to be as unbiased as possible for the distribution of the test set, predictions were quantile-transformed using the distribution of expression levels in training data as post-processing.
Unlock_DNA: The team used an end-to-end Macaron-like Transformer encoder architecture with two half-step feed-forward (FFN) layers at the beginning and end of each encoder block. A separable 1D convolution layer was inserted after the first FFN layer and in front of the multi-head attention layer. The sliding k-mers from one-hot encoded sequences were mapped onto a continuous embedding, combined with the learned positional embedding and strand embedding (forward strand vs. reverse complement strand) as the sequence input. Along with the sequence input, several positions (32 in the final model) of "pseudo" expression values were added as the input, where all input expression values were zeros. The model predicted one expression value for each "pseudo" expression position and used the mean of the prediction of all positions as the final predicted expression value.

Camformers: This team used a CNN with residual connections. The model included six convolutional layers with three residual connections allowing the model to bypass every other layer. After the penultimate convolutional layer, a max pooling operation was added to reduce the model size and improve generalization. The output of the final convolutional layer was flattened and fed into a block of two dense layers, followed by a final dense layer outputting the predicted expression level. All layers except the last used a rectified linear unit activation.

NAD: The approach has two stages: (i) generating the embedding vectors for each base position using GloVe (27) and (ii) using the embedding vectors as input of neural networks to predict the gene expression level. The proposed model combines a convolutional neural network for feature extraction and a transformer for prediction.

WZTR: The team used a fully CNN-based architecture. The model begins with two convolutional layers, and six convolution blocks follow these layers. Each convolution block is constructed of 3 convolutional layers and an average pooling at the end. Each convolutional layer consists of a hybrid convolution (40), batch normalization, ReLU activation, and residual connection. Each hybrid convolution takes in a list of dilation values [1,2,4,6], with 4 convolutions processing the input in parallel. Finally, there are three fully connected layers and an output layer. A linear combination of 256 features extracted from all the previous operations on the sequence is used to generate the predicted expression.

High Schoolers Are All You Need (High Schoolers): This team used a mix of CNN and transformer architectures, where the CNN was based on Residualbind's (41) design, with a convolutional layer (with exponential activations) followed by a residual block comprised of a series of dilated convolutional layers with increasing dilation rates. It was followed by attention pooling, a transformer layer with relative positional encodings, and a standard MLP block.

BioNML: The underlying neural network was configured to have a relatively larger set of convolutional kernels and extra dictionaries of short k-mers for spotting potential enriched DNA sequence patterns. Strand-specific streams of these patterns were normalized and consolidated with Swish activation’s (42) fully learnable thresholding. The encoded patterns were fed into a ViT (33) like block but with transformer decoder type of connections and SwiGLU (43) activations for modeling any sequential interdependence. A set of suppressed signals of the encoded sequence-based patterns as queries for the transformer decoder blocks to respond to.

BUGF: A transformer model was used to predict the expression bin classes, as opposed to treating the problem as a regression problem. Random mutations were added to the sequence as an data augmentation strategy and the model was trained to predict where the mutations had been made to the input sequence. An auxiliary loss was calculated based on this prediction, which helped reduce overfitting.
The approach uses GRU and CNNs to regress the strength of the targeted promoters using information encoded in the forward and reverse DNA strands.

SYSU-SAIL-2022: The team first trained a 3-layer BERT (29) using part of the train data. Then, the BERT embedding was used to train an expression predictor.

Wen Group: A deep neural network that adopted concepts of U-Net (44), Transformer, and Squeeze & Excitation blocks (45) was trained from end to end without any data augmentation.

Yuanfang Guan: A neural network that consisted of LSTM layers followed by attention layers was used to predict expression.

Metformin-121: A neural network based on bidirectional GRU was used to predict gene expression.

NGT4: A neural network based on XceptionNet (46) was used to predict gene expression. During training, the expression values were transformed evenly in the range of $i - 0.5 < x < i + 0.5$ (here, $i$ is the integer expression) maintaining the ranking of sequences that was produced by a trained model.

Davuluri lab: The team utilized a transformer-based representation model named DNABERT (28) for predicting gene expression.

*DNABERT was pretrained on human genome, which violated the competition rules. However, we consider this to be an important benchmark that shows the limitation of DNA language models.

Wan&Barton_BBK: The team designed a model based on Temporal Convolutional Networks (47) to predict expression.

Peppa: The team designed a model based on the Enformer (8) that took 110 bp as input (compared to 200 kbp in the original) and included only 2M parameters (compared to ~200M parameters of the original).

The Dream Team: A neural network was used that incorporated convolutional, multihead attention, and LSTM layers. During training, the integer expression values were transformed by replacing them with Normal($i,0.3$) distribution, where $i$ represents the expression.

Noisy-Chardonnay: A model composed of convolutional layers followed by BiLSTM layers was trained without any data augmentation to predict expression levels.

KircherLab: The team trained a simple convolutional neural network with a GC correction step on the training data to help the model focus its decisions on motifs within the sequence rather than the general nucleotide composition.

MadLab: The model is composed of three building blocks, namely a convolutional network, a transformer and a recurrent network.

Auth: A simple hybrid architecture combining a convolutional layer and a BiLSTM layer followed by two fully connected layers was used to predict expression.

UTKbioinformatics: A neural network based on BERT was used to predict expression.

DrAshokAndFriends: An attention based ConvLSTM (48) model was used for prediction.
QUT_seq2exp: A sequence embedding model, dna2vec (49), was applied on the promoter sequences (in a running manner on short k-mers), which are then subsequently used as features for a transformer-based deep neural network model.

Zeta: A transformer model was used for predicting expression.

Data availability

The processed training and test data used in this study are available at Zenodo database under DOI 10.5281/zenodo.7395397.
Supplementary Figure 1: Progression of the top-performing teams’ performance in the DREAM Challenge public leaderboard. (A,B) Performance (y-axes) in (A) Pearson Score and (B) Spearman Score achieved by the participating teams (colours) each week (x-axes).
showcasing the effectiveness of such challenges in motivating the development of better machine learning models.

Supplementary Figure 2: Library coverage differs between sequence supsets and is lowest for native sequences. Cumulative proportion (y-axis) of the number of reads per sequence (x-axis) for different sequence types (colours).
**Supplementary Figure 3**: Expression changes (y-axis) are biggest for motif perturbation, smallest for SNVs, and intermediate for motif tiling.
Supplementary Figure 4: The pairwise Pearson correlation between the ground truth and the predictions made by all models.
Supplementary Figure 5: The pairwise Pearson correlation between predictions made by different models for different test subsets, (A) random, (B) native, (C) high, (D) low, (E) challenging, and (F) SNVs.
Supplementary Figure 6: Models capture expression at high levels in different ways. Scatter plots displaying the relationship between model predictions (y-axes) and ground truth (x-axes) for the top-performing teams in the challenge, for: (A) Autosome.org, (B) BHI, (C) Unlock_DNA, (D) Camformers, (E) NAD, (F) wztr, (G) High Schoolers, (H) BioNML, (I) BUGF, (J) mt, and (K) reference model.

Supplementary Figure 7: Performance of the teams in each test data subset. (A,B) Model performance (colour and numerical values) of each team (y-axes) in each test subset (x-axes), for (A) Pearson $r^2$ and (B) Spearman $\rho$. Heatmap colour palettes are min-max normalized column-wise.

Acknowledgement

We extend our sincere gratitude to Jonathan Caton from Google Brain for his invaluable technical assistance in setting up the cloud resources for the challenge participants. Without his expert guidance and support, the successful organization of the challenge would not have been possible. We are also deeply grateful to TPU Research Cloud for providing the necessary cloud resources, which allowed us to ensure a level playing field for all challenge participants. We also thank Deep Genomics for providing travel grants to the top-performing teams in the challenge.
Consortia:
Random Promoter Dream Challenge Consortium members:


Affiliations:

1 University of Cambridge
2 University College London
3 University of Nevada Reno
4 University of Washington
5 Syosset High School
6 Cold Spring Harbor Laboratory
7 Friends Academy
8 Half Hollow Hills High School
9 Jericho High Schoo
10 Cincinnati Children's Hospital Medical Center
11 The Wellcome Centre for Anti-Infectives Research, Dundee University
12 University of Michigan
13 Australia National University
14 Northwestern University
15 Stony Brook University
16 Niigata University School of Medicine
17 National Taiwan University
18 Academia Sinica
19 ANIWARE
20 Development Center for Biotechnology, Taipei
21 National Chengchi University
22 School of Medicine, China Medical University
23 University of London
24 German Cancer Research Center, Heidelberg
25 St. Jude Children’s Research Hospital
26 Berlin Institute of Health at Charité – Universitätsmedizin
27 Global Blood Therapeutics
28 University of Southern Denmark
29 University of Tennessee at Knoxville
30 Aristotle University of Thessaloniki
31 SASTRA University
32 Queensland University of Technology
33 Beijing Normal University-Hong Kong Baptist University United International College

References:


