P101

AI-driven promoter optimization at MeiraGTx

Enrico Mossotto, Dustin Lee, Josefa M. Sullivan, Matthew J. During, Alexandria Forbes, & Ce Feng Liu

MeiraGTx New York

MEIRAGTx

1. Deep Learning as a new frontier for promoter development

Promoters are crucial elements for regulating potency and specificity of transgene expression. Whilst many promoter sequences have been characterized/optimized through conventional low-throughput analyses (rational design) or newer high-throughput methodologies (e.g., MPRA), these approaches still require an expensive and time-consuming in-vitro optimization. With better data and more advanced artificial intelligence models, we have repurposed and optimized a convolutional neural network (CNN) to predict promoter potency and optimize in-house promoter elements (Figure 1).

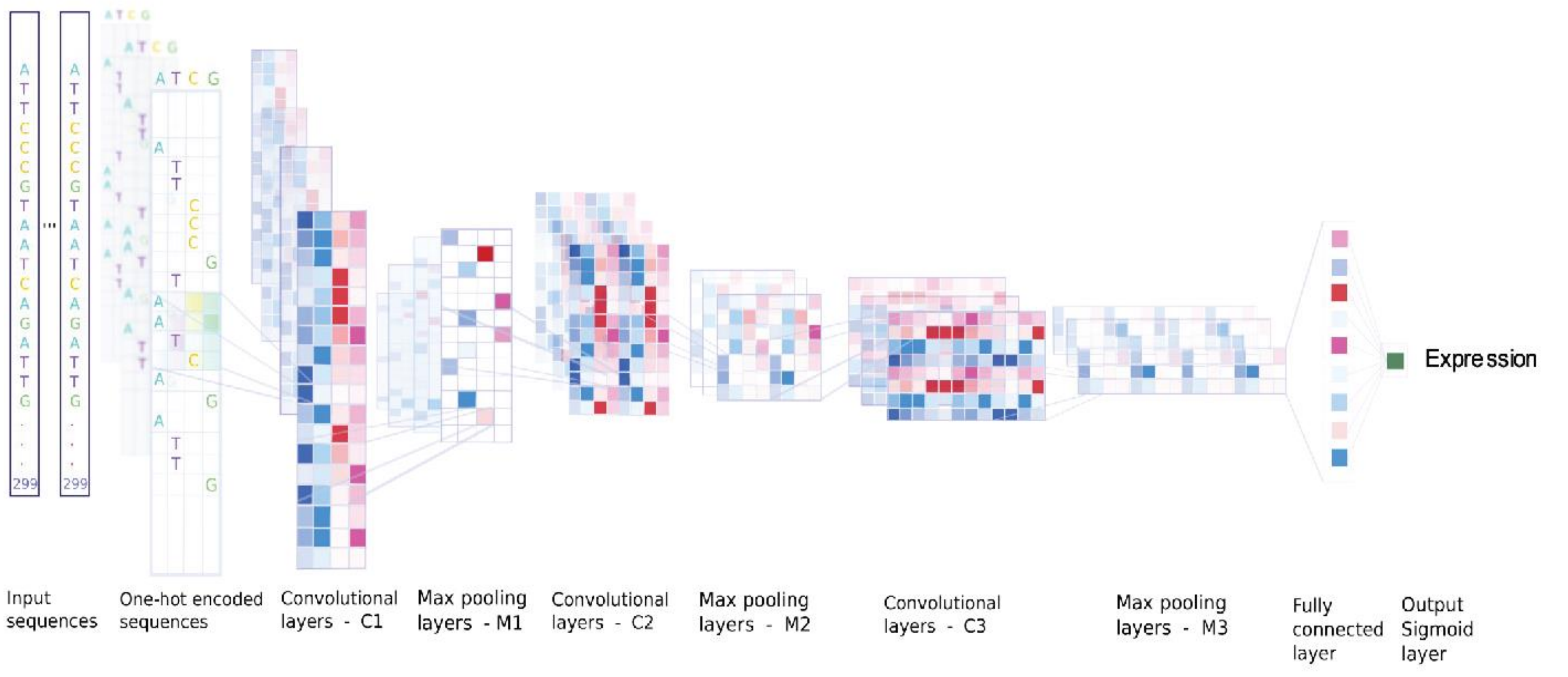

Figure 1. Example of a convolutional neural network model applied to genomic data. The CNN takes an input sequence and process it through a series of filters (convolutional and pooling layers) to extract quantitative and spatial signals (e.g. motifs and their location). A last fully connected (dense) layer then further process such information and returns the predicted expression (potency) level. Image adapted from doi:10.7717/peerjcs.278/fig-1

2. Optimization strategies to mimic large in-vitro screens

i) Promoter Enhancement: all known enhancer fragments from the ENCODE database were inserted upstream of the promoter sequence. Subsequently, our CNN was used to select those with the strongest predicted potency and to optimize the spacing in between elements;
ii) Saturated Mutagenesis: we simulated all possible point mutations (no indels) for each base pair in the C6 promoter. The CNN was used to select those with the largest predicted positive impact and to calculate their combinatorial effect (greedy approach).
From both approaches, we obtained a set of candidate promoters that were then synthesized and tested in-vitro. Best performers of each approach were combined in a rational design and further tested in a second in-vitro round of testing.

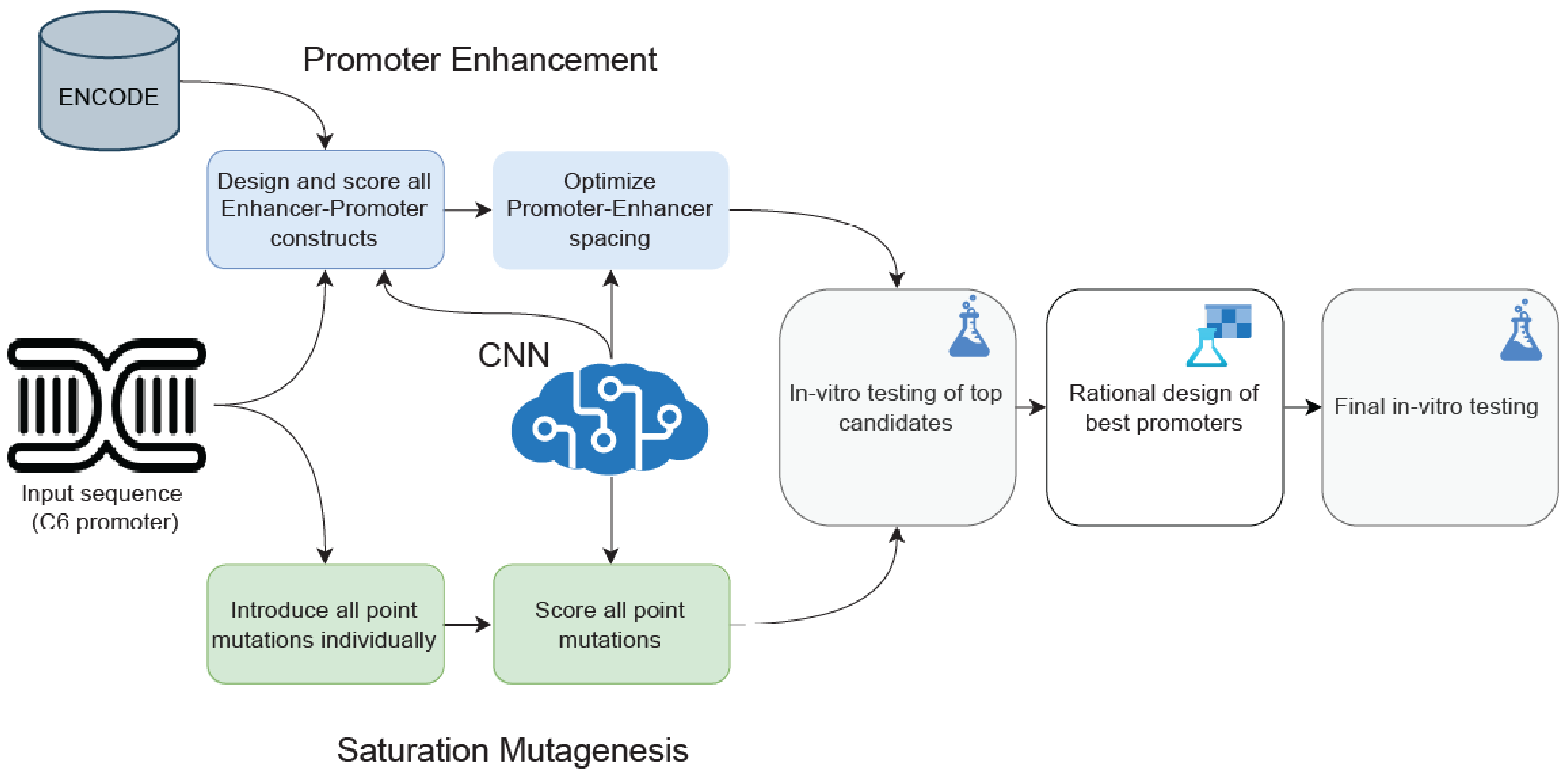
Promoter Enhancement
ENCODE
Design and score all Enhancer-Promoter constructs
Optimize Promoter-Enhancer spacing
CNN
Input sequence (C6 promoter)
In-vitro testing of top candidates
Rational design of best promoters
Final in-vitro testing
Introduce all point mutations individually
Score all point mutations
Saturation Mutagenesis

Figure 2. Bioinformatics pipeline for the AI-driven design and optimization of a given promoter sequence.

3. AI screens and predicts optimal enhancer elements

Through our CNN model we predicted the effect of 708,157 enhancer fragments in-silico cloned upstream the C6 promoter sequence (Figure 3A).
The majority of enhancers are predicted to have a positive effect on C6 potency with ~2,000 constructs producing a ~2x improvement. This subset was further analysed to identify the optimal spacing between the enhancer element and the promoter sequence (Figure 3B). Introducing a spacing sequence significantly improved the potency of the construct with the largest improvement observed with a spacer of 400bp. The top 10 constructs were synthesized and tested in vitro (C118 to C127).

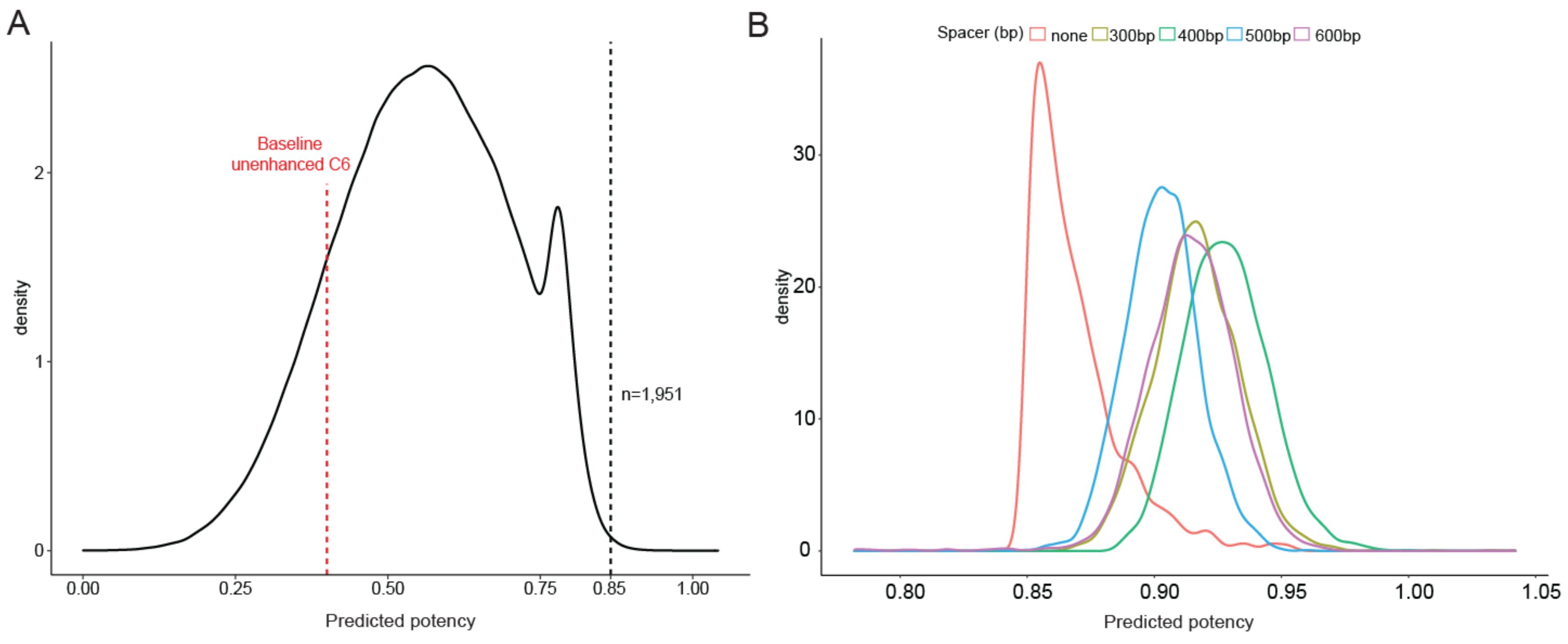

Figure 3. A Predicted effect of all ENCODE enhancers in combination with a C6 promoter. The red line indicates the predicted expression level of the C6 promoter without any enhancer sequence upstream. The black dotted line shows the threshold for selecting constructs for the spacing optimization step. B. Enhancer-Promoter spacing optimization. Each distribution shows the effect of four spacing sequences differing in size from a minimum of 300bp to a maximum of 600bp.

4. AI screens and predicts optimal point mutations

Figure 4A shows the predictions of all C6 mutants after introducing each possible point mutation in isolation. Each dot represents a point mutation and its effect (y-axis) normalized against the non-mutated C6 promoter. Few mutations are predicted to provide a 20% or more increase in potency (predicted potency >1.2). We then expanded the concept of saturation mutagenesis to identify a "hyper mutated" construct which maximizes potency. This approach introduces multiple point mutations within the same C6 promoter (Figure 4B). E.g., at the 1.13 threshold (x-axis), there are 44 mutations (blue line) individually boosting the C6 activity by at least 13%. By combining all those 44 mutations in the same C6 promoter, we predict a final improvement of the C6 promoter by 3 folds.

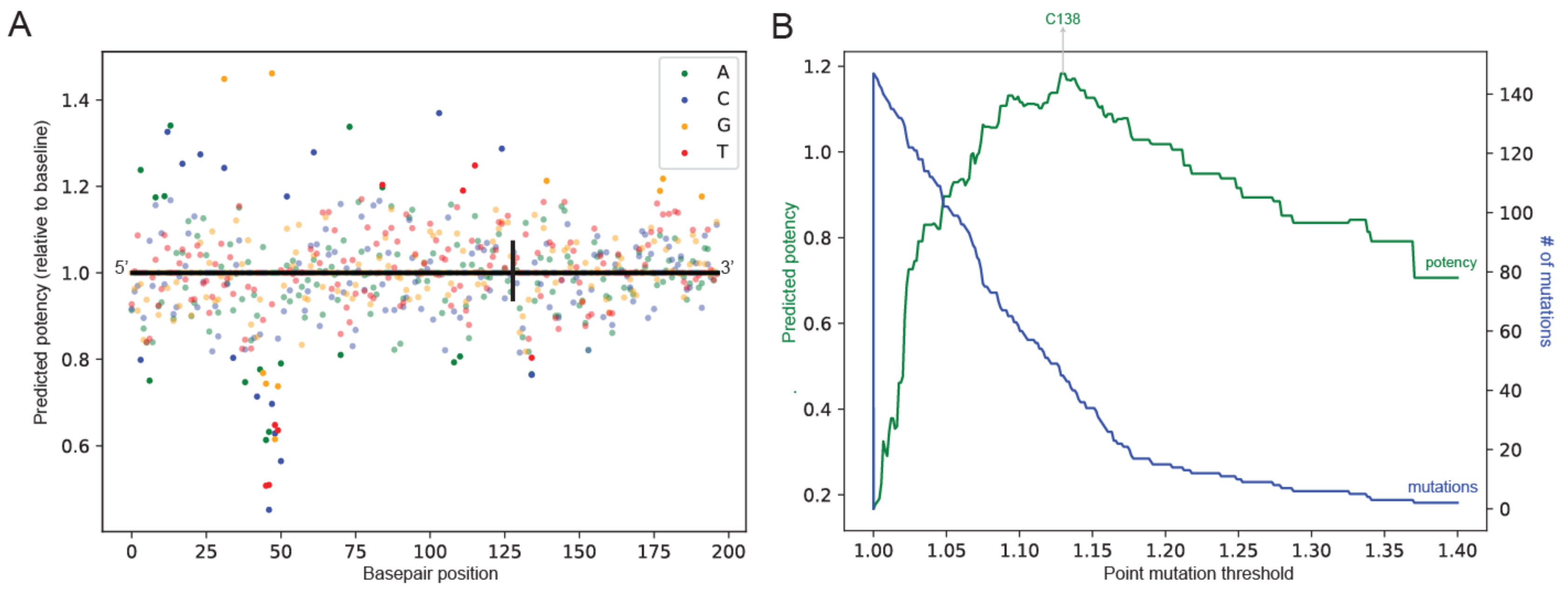

Figure 4. A. Saturation mutagenesis of the C6 promoter. The effect of each point mutation was normalized against the non-mutated construct. The black bar indicates the location of the transcription start site (TSS). B. Identification of the optimal mutations to be simultaneously introduced in the C6 promoter to improve performance. The peak construct was labelled as C138.

5. In-vitro validation of AI-designed constructs

Constructs with the highest predicted potency were synthesized and tested in HEK293T cells. Results from the first generation show most of the Enhancer+Promoter constructs perform equally well or better than the parent C6 promoter. In particular, constructs C120 and C124 exhibit an improved potency of ~30% and ~40% respectively. Similarly, some point mutations resulted in a boost in performance between 10% and 30%. Hyper-mutated constructs (with either a iterative or greedy approach) disrupted promoter activity in HEK293T cells.
Based on the results from the first generation, we rationally designed a new set of construct based on a combination of those with the highest in-vitro expression. The second generation was tested in HEK293T cells and all constructs exhibited higher expression than the parent C6. In particular, construct C187 reached ~2 fold improvement when compared to the original C6 construct.

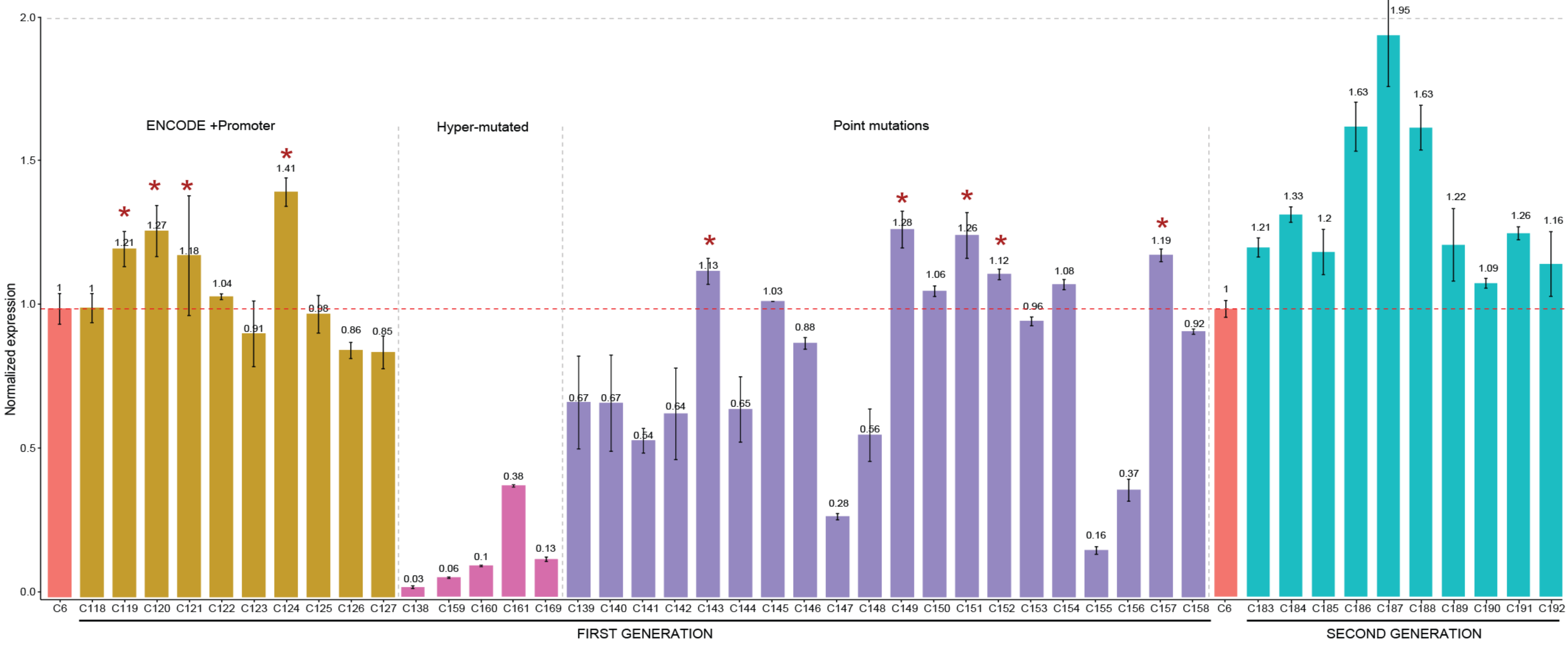

Figure 5. In-vitro results of constructs designed by the AI model (first generation). Different colors represent different design approaches. Red asterisks indicate those top performers which were selected for the subsequent rational design (second generation). Expression levels were normalized against the parent promoter C6 (red bar).

6. Conclusions

At MeiraGTx we are actively pursuing cutting-edge approaches to develop novel gene therapies. One of our main focus is on developing strong promoters that are ubiquitously active in humans.

While conventional rational design is a powerful tool to improve construct potency, it cannot be scaled to a large number of promoters and its throughput is limited. Here we showed how at MeiraGTx we are employing deep learning models (AI) to screen large libraries of regulatory elements and select those with the greatest predicted potency. Such approach is a cornerstone for gene therapy as it drastically reduces the search space and consequently accelerate the discovery of novel therapies.

In this presented work, we demonstrated how, with just two generations of promoter optimization, our AI model was capable of directing us in choosing regulatory elements and mutations that significantly boosted the activity of a parent promoter.