Leveraging Model Interpretability Methods to Predict Gene Therapy Manufacturing Failures

Nick Ketz, PhD, Computational Scientist, Form Bio, Inc.
nick@formbio.com

MODEL INTERPRETABILITY METHODS provide an understanding of complex model decisions and verify that a meaningful difference in the data has been identified. We have applied model interpretability methods to our predictive model of genome truncation within adeno-associated virus (AAV) manufacturing and revealed that the model uses a set of DNA secondary structures predictive of truncation. These secondary structures provide a simple mechanism for understanding AAV truncations and a strong basis for independently validating our model's predictions. Moreover, these structures have been well studied and are shown to be related to DNA replication errors; however, only one of these structures (i.e., hairpins) has been previously implicated in AAV manufacturing failures. Future research will concentrate on additional contributors to our truncation model to gain a deeper understanding of AAV manufacturing failures more generally.

An Introduction to Viral Vector Manufacturing

Viral vector manufacturing is a crucial step in developing gene therapies, as it involves producing a large number of viral particles that will be used to deliver a therapeutic gene to a patient. However, this process has its challenges. Failures in gene therapy manufacturing can occur for various reasons, including incomplete transgene replication and packaging into the viral shell, colloquially referred to as 'truncations.' Here, the viral genome packaged into the AAV capsid is a truncated form of the intended complete genome.

Analysis of these truncated viral genomes reveal key sequence locations where replication fails. DNA inverted repeats, or hairpins, are a specific secondary structure related to these manufacturing failures as shown in *Xie at al.*[1] This research indicated that replication failures leading to truncation can occur at the site of the secondary structure. Current theoretical models suggest that a strand or template switching phenomena is driving these failures.

Despite the documented impact of hairpins on viral vector production, their effect on manufacturing success can vary greatly across transgene designs. In some cases, hairpins may be present but not lead to any replication failures, while in other contexts, they can cause a significant number of truncations. Conversely, in the absence of hairpins, there can still be a high likelihood of viral genome truncation. Our data shows variation in truncation rates from 5 to 45% of samples exhibiting incomplete AAV genomes across transgene designs, yet all of these designs have hairpins present to some degree. This variability makes it difficult to predict which vectors will be successful and which will not, ultimately leading to delays in manufacturing and slower time-to-market for new therapies.

To address this problem, we've interrogated our convolutional neural network designed for AAV truncation prediction using popular model interpretability techniques for deep learning models.[2] By understanding the model and what it has learned, we will be able to confidently apply our model more broadly and further the understanding of AAV manufacturing failures to drive better solutions.

Data and Truncation Prediction Model Used

Our data is derived from long-read sequencing of viral products from a total of ten manufacturing runs spread over four unique AAV transgene designs. All sequences are self-complementary AAV-9 designs with reverse followed by forward strands as the intended order in the target genome.

Briefly, the version of our truncation prediction model used is a neural network composed of 2 convolutional layers (50 then 100 filters, respectively) followed by two linear layers (100 then 50 units, respectively), as shown in Figure 1A. Each convolution layer is followed by ReLU activations, max-pooling operations, and a drop-out rate of 0.35. Input features are 100 nucleotide long sequences of one-hot encoded DNA concatenated with a reverse/forward-strand categorical variable of the same length. The corresponding output labels for each input are the log of the average truncation propensity for that 100 nucleotide window. Truncation propensity is calculated for each nucleotide as the percentage of reads terminating at that nucleotide divided by the total number of unterminated reads remaining. For a 100 nucleotide input window, truncation propensity is averaged over the middle 50 nucleotides to provide a single learning target.

Data were split in a 3 train to 1 test ratio, using a grouped and stratified method to ensure equal distribution between train and test sets. Grouping was across positional index within a given AAV transgene design. Stratification was over three quantiles of truncation propensity – low, medium, and high – with each having an equal number of samples. Trained model performance is shown in Figure 1B for a specific transgene design. Performance across all data was assessed using the Pearson correlation between the predicted and true log truncation propensity, with train data yielding 0.89 and test data yielding 0.86 r-values.

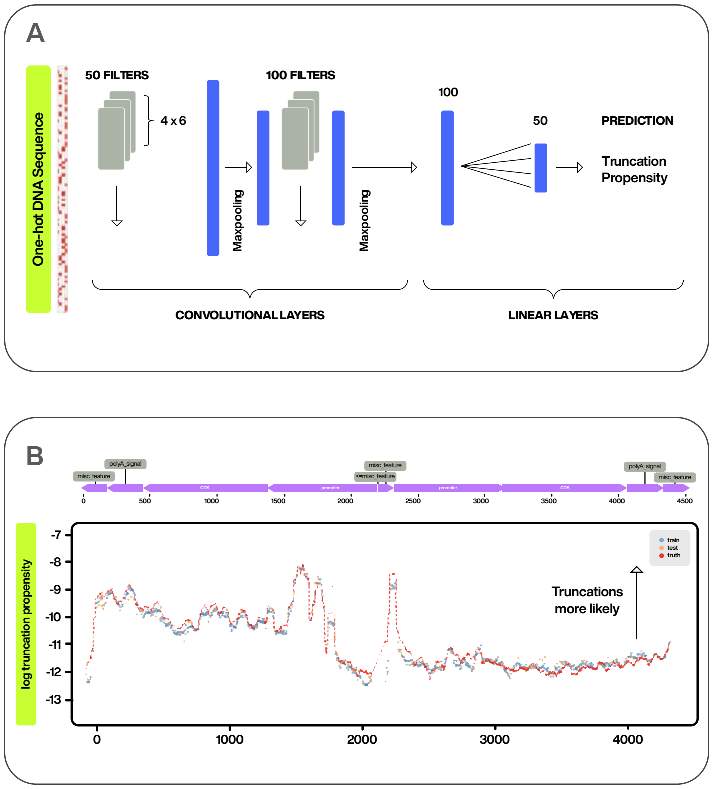

FIGURE 1. Our truncation prediction model **A)** Sketch of network architecture. **B)** Example model performance for a given transgene design. The log of truncation propensity is plotted as a function of the sequence index, with true truncation propensity shown with the red dashed line and trained model predictions for train and test data points shown in blue and orange dots, respectively.

Truncation Prediction Model Interpretation

One validated approach for interpreting convolution models of DNA sequences is to derive motifs, or representative sub-sequences, that co-occur with high or low output values. Once identified, these motifs can be analyzed for unique properties within the problem domain. A good example of this is illustrated in BPNet, a transcription factor binding model that reveals motifs of binding locations, which can then be matched against known transcription factors for validation.[3] Our intention was similar in that we sought to find motifs for subsequent analysis of secondary structures (or other features) that lead to truncation.

The analysis path we used to achieve this is as follows: Derive importance scores as a contrast between the gradients of our input patterns and their corresponding null patterns using Integrated Gradients, and then extract and cluster sub-sequences that co-occur with high and low truncation propensity using TF-MoDisco.[4,5]

The analysis to derive importance scores was done using Integrated Gradients as implemented in Captum.[6] Here, each input pattern from our dataset was paired with a 'null' pattern derived from ten different dinucleotide shufflings of the given input pattern. The contrast of the gradients in each pair is averaged together to arrive at a single importance value for each nucleotide in each input pattern. Positive values indicate that the nucleotide contributes to higher levels of truncation and negative values indicate that a nucleotide contributes to lower levels of truncation.

Once importance scores were calculated, motifs were discovered through sophisticated filtering and cluster techniques within the TF-MoDisco methods. One somewhat subjective component in this process is providing hypothetical/counterfactual importance scores to contrast with the true importance scores. For our hypothetical inputs, we calculated importance scores for all permutations of each nucleotides in each input pattern in a similar manner above (i.e., each permutation compared with 10 dinucleotide shufflings). Otherwise, default values were used throughout the motif discovery process.

Truncation Motifs Are Associated with the Formation of DNA Secondary Structures

Examples of two motifs related to increased truncations and their information content are shown in Figure 2. Information content was calculated as the KL-Divergence between the probability distribution for each nucleotide in the motif importance scores normalized within each nucleotide and the probability distribution of each nucleotide normalized across all input patterns.

Post-hoc inspection and literature review of the discovered motifs revealed a class of DNA folding patterns, or secondary structures, that have been implicated in various

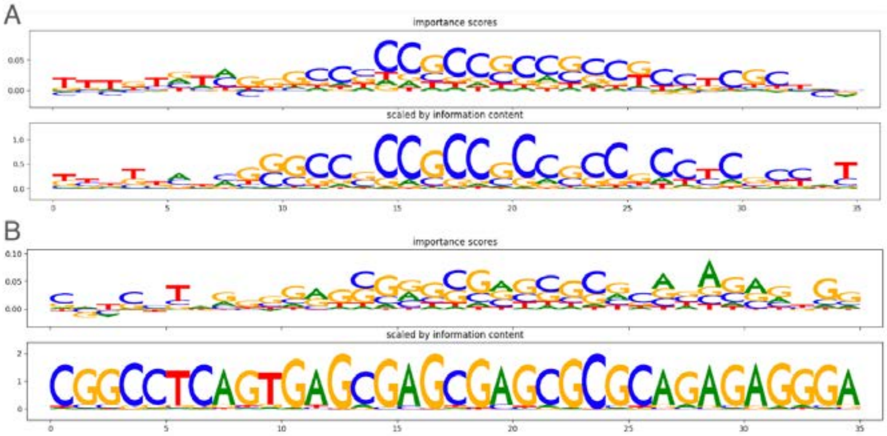

FIGURE 2. Two example motifs are associated with increased truncation propensity. Motif importance scores are shown on top, and the normalized importance scaled by information content is shown above. **A)** This motif shows a repeating pattern of 'CCG' within the center of the motif, most visible when scaled by the information content. **B)** This motif shows a repeating pattern of 'AGCGA' within the center of the motif, again most visible when scaled by the information content.

studies of replication-related mutagenesis and human diseases.[7,8] For example, the motif shown in Figure 2A illustrates a repeating pattern of 'CCG,' which can form a secondary structure involved in Huntington's disease, myotonic dystrophy type 1, and Fragile XE syndrome.[9] Similarly, Figure 2B shows a motif with repeating 'AGCGA' patterns, which forms secondary structures linked to various neurodevelopment and neurological disorders .[10]

This family of secondary structures, as illustrated in Figure 3A, is generally referred to as non-B or non-canonical DNA (we will use these terms interchangeably throughout this document). Non-B DNA can be considered any confirmation of DNA outside the standard Watson-Crick double helix pattern which often rely on nucleotide bindings outside the canonical A-T and G-C pairings. These anomalous structures are not new, and in fact have a long history in genomic research. Recently, non-B DNA has been studied in the context of replication errors leading to increased mutation rates.[7,8] As shown in Figure 3B, the proposed mechanism suggests that during DNA replication, polymerase speed slows when it encounters these problematic folding patterns, leading to increased errors in the replication process.

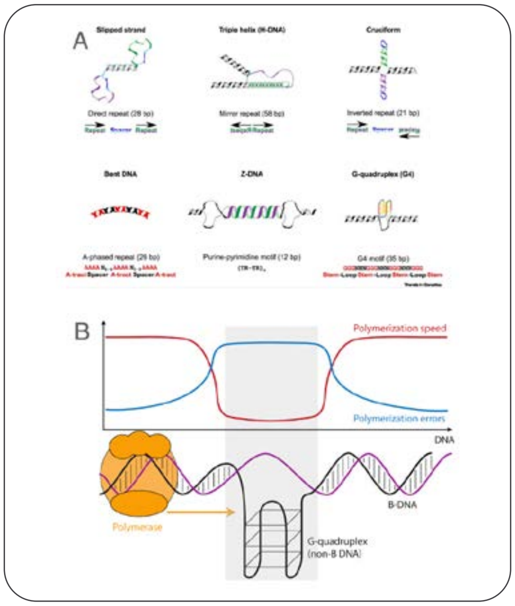

FIGURE 3. Set of secondary structures **A)** Examples of non-B DNA secondary structures and their characteristic sequences[11] **B)** Proposed model for the relationship between non-B DNA and replication errors. Image credit to Guiblet et al.[8] The polymerase slows when it encounters non-canonical DNA folding during replication, leading to an increased error rate in the copied DNA.

This mechanism provides a compelling basis for understanding why truncations occur in AAV manufacturing; however, no existing literature supports this connection. Is our model truly using these secondary structures to predict truncations?

Truncation Propensity is Highest When Motifs and Non-Canonical Secondary Structures are Present

Based on the results from the motif analysis, we sought to validate the presence of the secondary structures implicated by the discovered motifs. We intended to answer this key question: Are non-canonical secondary structures related to truncations? And if so, is the model solely using them to predict truncation propensity?

To achieve this, we developed an analysis framework based on existing work to identify the set of secondary structures, independent of the motifs themselves, and indicate their presence within our dataset.[12] Similarly, we marked input patterns that contain discovered motifs using existing methods from TF-MoDisco. Once input patterns are categorized, we then quantified their relationship to truncation propensity, which can be seen in Figure 4. Here, input patterns fall into 1 of 4 categories: Inputs with motifs and secondary structures (motif hits: True, non-canonical: True), inputs with motifs and no secondary structures (motif hits: True, non-canonical: False), inputs with no motifs that do have secondary structures (motif hits: False, non-canonical: True), and finally inputs with no motifs and no secondary structures (motif hits: False, non-canonical: False). An example of a given input pattern with both a motif and secondary structure is shown in Figure 4C.

Each input pattern has a corresponding truncation propensity, and the interaction plot in Figure 4A shows the average truncation propensity for the four categories with error bars showing 95% confidence intervals.
This plot shows that the average truncation propensity is highest when both motifs and non-canonical secondary structures are present in the input. This shows that, the model is in fact using motifs to identify secondary structures that lead to truncations.

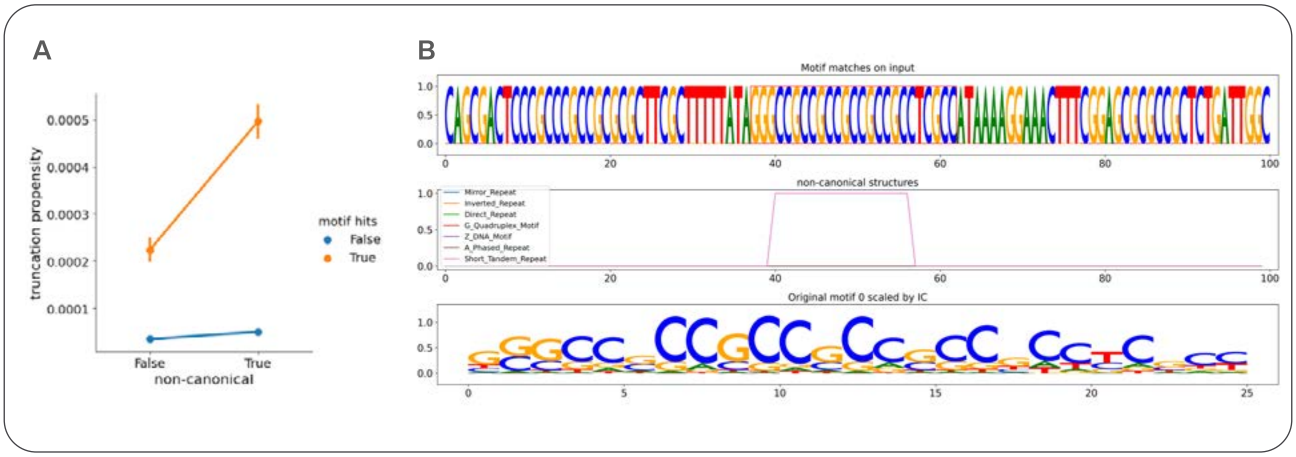

FIGURE 4. A) Average truncation propensity as an interaction between the presence of discovered motifs and non-canonical secondary structures. Error bars show 95% confidence intervals. **B)** Example input pattern showing the presence of a motif and a secondary structure. The top panel shows the one-hot encoded input pattern with motif match highlighted in black. The middle panel shows the presence and type of non-canonical secondary structure as a binary indicator. The bottom panel shows the matching motif scaled by its information content.

This plot also shows that other features, beyond just secondary structure, are being used to predict truncations: Input patterns with a motif, but no secondary structure present, have a higher truncation propensity than input patterns with no motif present (compare left-most orange data point to left-most blue data point in Figure 4A). Although these inputs show a reduced truncation propensity compared to those with non-canonical secondary structure, they also account for a significant portion of the truncation variance.

Conclusions

The implications of these findings are still being explored; however, a few summary points can be made from the results shown thus far.

First, we find evidence from our motif analysis that our model identifies a set of repeating patterns (i.e., motifs) that are predictive of failure points (i.e., truncations) in AAV manufacturing (Figure 2). Previous publications have explored these patterns surrounding non-B, or non-canonical, DNA folding (Figure 3).
Second, as shown in Figure 4, our secondary structure analysis revealed that these repeating patterns are related to truncation propensity independent of the model's learned representations. More specifically, the truncation propensity increases in the presence of non-canonical secondary structures. However, a significant proportion of predicted truncation propensity does not co-occur with the presence of non-canonical secondary structures, demonstrating that our current understanding of non-canonical secondary structure alone is insufficient for understanding AAV manufacturing failures.

In a subsequent white paper, we will explore this data further to understand the relationship between truncations and non-canonical secondary structures more precisely. Specifically, how much of this variation in truncations can be attributed to hairpins vs other types of non-canonical structures? Similarly, we want to understand how our codon optimization process (reported in a previous white paper) interacts with these secondary structures to achieve a lower predicted truncation propensity.[2] Finally, we will distill our findings into new research directions that advance the efficiency of viral vector manufacturing.

References

1. Xie, J. et al. Short DNA Hairpins Compromise Recombinant Adeno-Associated Virus Genome Homogeneity. Mol. Ther. 25, 1363–1374 (2017).

2. Developing Machine Learning Powered Solutions for Cell and Gene Therapy Candidate Validation. Form Bio. Published Dec 2022. Accessed Jan 20, 2023.

3. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. 53, 354–366 (2021).

4. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks.

5. Shrikumar, A. et al. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5.

6. Captum. Published 2023. Accessed Jan 23, 2023.

7. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res. 28, 1264–1271 (2018).

8. Guiblet, W. M. et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 28, 1767–1778 (2018).

9. Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. Crystallographic characterization of CCG repeats. Nucleic Acids Res. 40, 8155–8162 (2012).

10. Kocman, V. & Plavec, J. Tetrahelical structural family adopted by AGCGA-rich regulatory DNA regions. Nat. Commun. 8, 15355 (2017).

11. Makova, K. D. & Weissensteiner, M. H. Noncanonical DNA structures are drivers of genome evolution. Trends Genet. 0, (2023).

12. Cer, R. Z. et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res. 41, D94–D100 (2013).