Developing Machine Learning Powered Solutions for Cell and Gene Therapy Candidate Validation

Joseph Nipko, PhD, Head of Computational Research, Form Bio, Inc.
joe@formbio.com

Executive Summary

THERE ARE OVER 1,000 ongoing clinical trials evaluating the safety and efficacy of cell and gene therapies in a broad array of therapeutic areas. While only a handful of these new therapies have been FDA-approved, biopharmaceutical investors are banking heavily on cell and gene therapy companies and their potential to unleash powerful cures for rare genetic diseases and many cancers.

However, the infrastructure to support the entire journey from pre-clinical research to approval and commercialization is in its infancy. Most cell and gene therapies being evaluated are in phase I/II trials. They have yet to navigate the scale-up process, and with the current manufacturing capabilities, some may be destined for disappointment. Even therapeutics that have successfully navigated approval face challenges due to excessive pricing and lack of clarity around reimbursement.

New techniques and technologies are needed to make research, development, manufacturing, and commercialization more efficient. Artificial intelligence (AI) algorithms offer approaches to positively impact these processes, improving the quantity and quality of manufactured products and production efficiency. Here, we explore the use of deep learning (DL) to make cell and gene therapies more manufacturable, advancing them into a new era of innovation.

An Introduction to Cell and Gene Therapy

CELL AND GENE THERAPIES use the body's own cells and genetic information to fight diseases but do so using different strategies. Cellular therapies use patient or donor cells as therapy and are introduced into a patient to grow, replace, or repair cells or tissues associated with a specific disease. Several different cell types are used in cell therapies, including hematopoietic stem cells, neural stem cells, lymphocytes, dendritic cells, and many others. Chimeric antigen receptor T-cells (CAR-T cells) have become a well-known form of cellular immunotherapy, and since 2017, six CAR-T cell therapies have been FDA-approved for the treatment of various blood cancers, including multiple myeloma, leukemia, and lymphoma. CAR-T cells are commonly made by isolating T-cells from a patient and introducing the CAR gene using a viral vector.

Gene therapies alter a patient's genome in cells isolated from a patient (e.g., *ex vivo*) or in a patient's body (e.g., *in vivo*). CAR-T cell therapies can be classified as *ex vivo* gene therapy. *In vivo* gene therapies, such as Zolgensma, is an *in vivo* gene therapy that introduces a functional version of the *SMN1* gene in children with spinal muscular atrophy, missing *SMN1,* or having a non-functional version of the gene. The gene is introduced using an adeno-associated virus (AAV) to transduce motor neurons and replace the missing or mutated *SMN1* gene with the functional one.

Power vs. Price: The Double-Edged Sword of Cell and Gene Therapies

THOUGH THEY TAKE different approaches, cell and gene therapies offer immense promise in treating diseases that previously had no treatment options. Today, these therapies lead the way in pre-clinical and clinical successes. In the case of gene therapies for rare monogenic disorders, each manufactured dose has tremendous therapeutic power: The potential to substantially improve a patient's quality of life or cure a disease with a single dose.

Yet, just as any drug development program can face pitfalls and uncertainties, cellular and gene therapy development can face some unique challenges. Though the benefits of successful gene therapies are clear, the price tag to buy such a treatment can be astronomical. Consider Bluebird Bio's gene therapy to treat a rare genetic blood disease called beta-thalassemia, caused by mutations in the beta-globin gene. Bluebird's treatment earned European approval in 2019, but the cost was nearly $2 million per patient. This, in turn, presented reimbursement issues and resulted in market withdrawal from Germany, pausing further development of this treatment for European markets.[1]

Despite this setback, in early 2022, the FDA approved Bluebird's ZYNTEGLO for US patients.[2] Pricing of the therapy, however, remains a point of criticism, earning the unattractive label as "...the most expensive medicines ever sold in the U.S."[3]

The ZYNTEGLO case and others highlight the challenges and risks to cell and gene therapy developers and manufacturers for rare diseases, particularly as the biopharmaceutical industry continues targeting smaller patient populations. These therapies are complex, requiring significant R&D expenditure and clinical testing. There's a considerable focus on what infrastructure or supportive technology is needed to make development and manufacturing processes more efficient and less expensive.

The Need for Improved Viral Vector Manufacturability

THE KEY TO SUCCESSFUL cell and gene therapy is efficient gene delivery to target cells. To achieve this delivery, genes are packaged into vehicles called vectors. Several viral-based vectors have been used for gene delivery, including lentivirus, adenovirus, and adeno-associated virus (AAV).[4,5] AAV has emerged

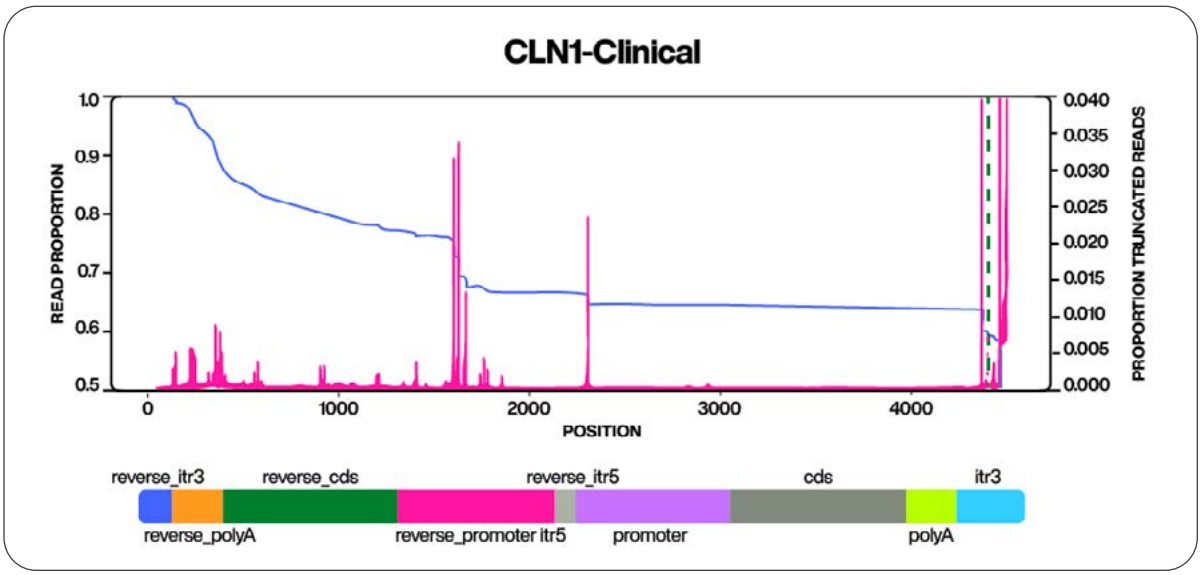


FIGURE 1 The terminal read distribution for CLN1 disease, also known as infantile Batten disease. The construct is a self-complementary DNA strand (bottom panel). The strand is characterized by inverted terminal repeats (ITRs) on the ends, a polyA tail, a promoter (in this case, chicken β-actin (CBh)), and the coding sequence (CDS) region. The left axis (top panel) provides the proportion of reads that terminate at the specific nucleotide (blue line); the right axis (red line) is the truncation propensity derived from the read termination. Following the blue line to the end of the sequence, we observe that the proportion of full-length constructs is approximately 65%.

as a predominant vector due to many desirable attributes, including a lack of pathogenicity, efficient infection of dividing and non-dividing cells, and sustained maintenance of the viral genome.[6]

Despite significant clinical and commercial successes, the cost-effective manufacturing of viral vectors remains a challenge, mainly because of the need for a clear mechanistic understanding of how various processes impact the final AAV product quality.[7] Better manufacturing of smaller batches of AAV vectors for gene therapies would help in the pre-clinical and clinical development of a portfolio of targeted products.[8]

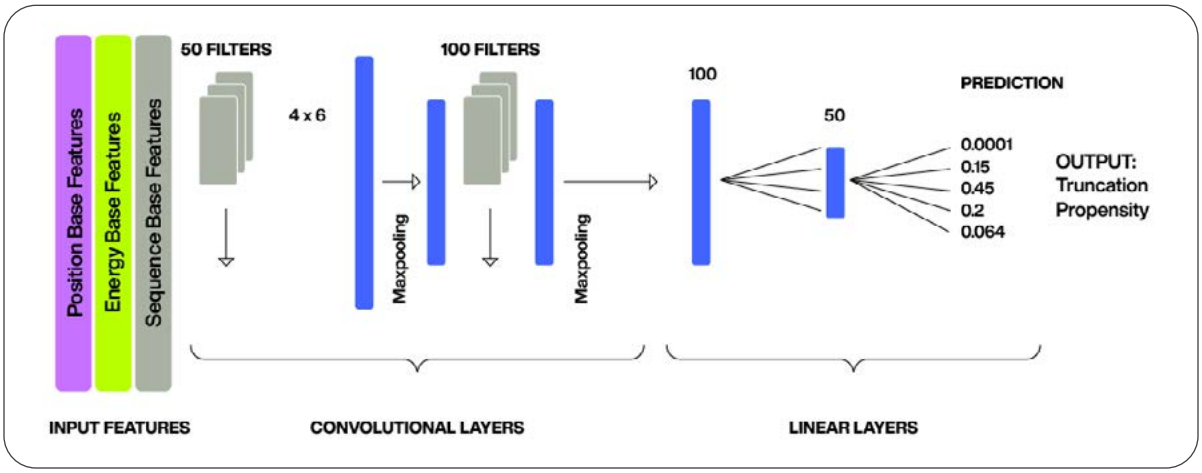
Tackling the Construct Truncation Problem with AI

ONE SIGNIFICANT LIMITING FACTOR in transgene packaging into an AAV delivery vector is the truncation of the construct. Recent research has shown that these truncations can occur due to the secondary structure of the genome. This research has revealed that DNA inverted repeats are hotspots for genome instability because they can form stable hairpin or cruciform structures that interfere with DNA replication.[9] The biochemical and structural mechanism by which these structural artifacts induce truncation during viral packaging is an active area of research necessary for reducing the incidence of truncation and improving the overall quality of manufactured gene therapies.

AI can play a valuable role in modeling and predicting truncation and designing constructs with a reduced propensity for truncation, improving the manufacturability of AAV-based gene therapies. More specifically, deep learning algorithms are an excellent option to model truncation propensity due to their flexible architecture, a variable receptive field for inputs, and ability to self-learn important drivers of signal from the information. Once trained, they provide an *in silico* mechanism for prediction and subsequent optimization of the construct design. As an example, consider the following construct design as a therapeutic for infantile Batten disease caused by mutations to the *CLN1* gene (**FIGURE 1**).

The read distribution shows that the manufacturing process for the constructs leads to approximately 45% truncation of the transgene in the final therapeutic. This degree of truncation makes the manufacture of the therapeutic less viable and can lead to risks of higher toxicity of the final product. By leveraging several different construct read distributions from other therapeutics with the same cell type (central nervous system), sufficient data can be obtained to train a machine learning algorithm to predict these truncations (**FIGURE 2**). In this case, we used a convolutional neural network to provide a framework for the learning algorithm.

Recent research points to structural stability and characteristics of the subsequence energy profiles that lead to observed truncation during transcription in self-complementary DNA.[10] In designing features that can learn representations that target truncation propensity, we leveraged these understandings to arrive at a predictive model. When both supervised and semi-supervised methods are used, along with well-designed input features designed to target truncation, we were able to construct a highly predictive model. The predicted vs. actual truncation propensity distribution for the therapeutic for infantile Batten disease is aligned, validating the accuracy of our CNN (**FIGURE 3**).



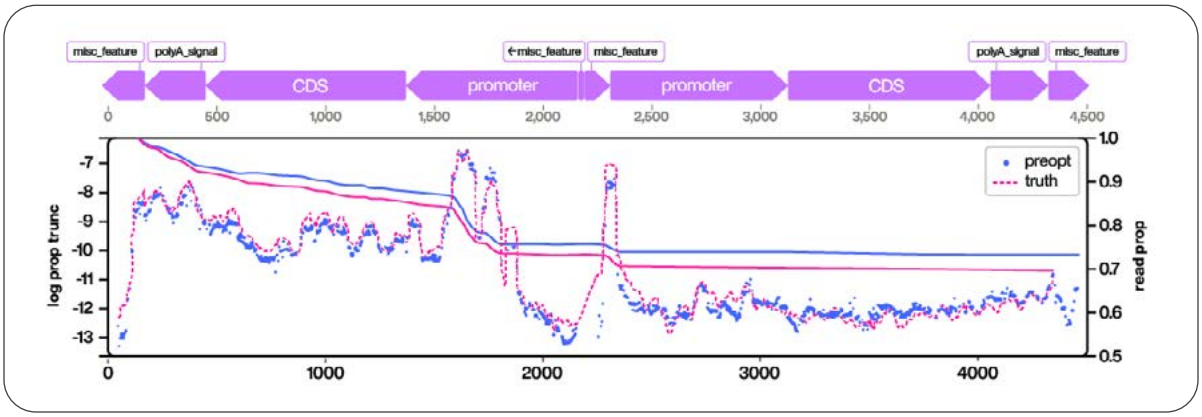


FIGURE 2 The overall structure for a convolutional neural network (CNN) that is architected to predict truncation propensity. The design of the network takes advantage of the nature of the input subsequences such that each nucleotide is encoded into a representation that accounts for the local stability and sequence properties. AAV constructs need to be relatively short (<4.4 kB), and the read length data for the constructs is limited. Consequently, an architecture with limited parameters, such as a CNN, is preferred.

FIGURE 3 The predicted (blue dots) vs. the actual (red line) truncation propensity for the CLN1 construct. The x-axis is the nucleotide position on the strand; the left y-axis is the log of the truncation propensity. While the right y-axis provides the read proportion.

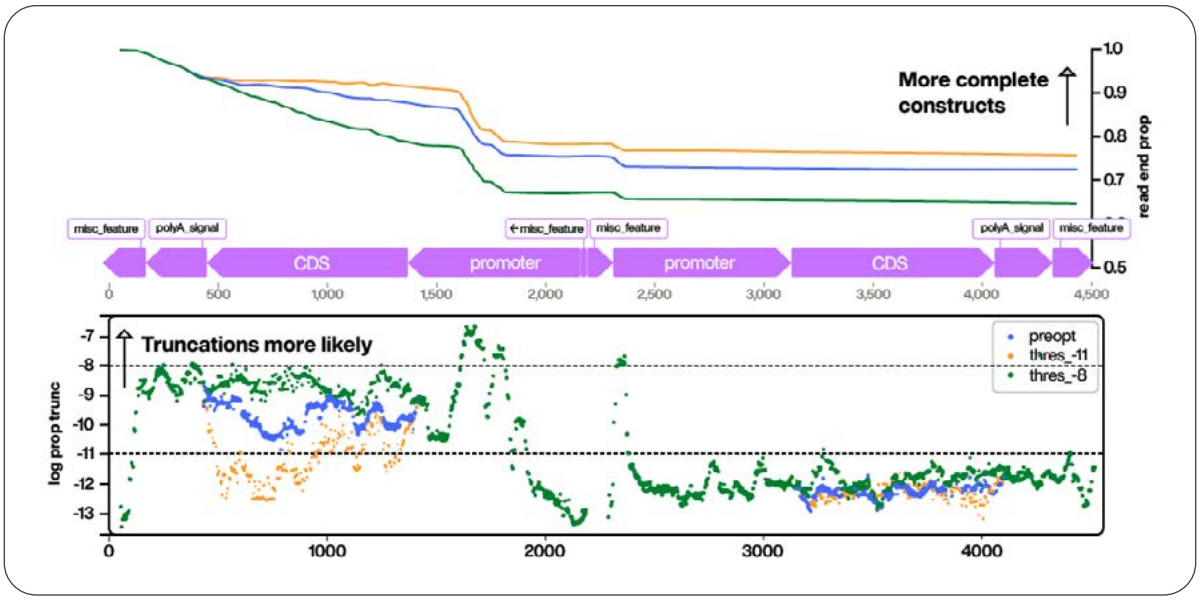


FIGURE 4 A comparison of the optimized construct vs. the original construct. The top panel shows the read terminations proportion for the original construct (blue), construct optimized for minimum truncation (orange), and construct optimized for maximum truncation (green). The bottom panel depicts the truncation propensity for the same three conditions. It demonstrates that clear areas within the CDS are particularly susceptible to structural instability, which can be altered through codon substitution.

With a predictive model in hand, the stability and manufacturability of the construct can now be improved by introducing mutations to the DNA that reduce the truncations while preserving the gene structure of the construct. One approach to achieve this is codon optimization, whereby synonymous codons are changed according to codon bias. This changes the biochemical nature of the DNA coding sequence but not the protein sequence. The objective function for the optimization is the minimization of the function that predicts the propensity for truncation along the construct. When we apply codon optimization or "de-optimization" to constructs in our model, it behaves as expected: Optimized constructs result in a greater proportion of full-length constructs (**FIGURE 4**).

Our work has shown that combining biological knowledge, machine learning, and codon optimization can reduce truncation by as much as 70% in the coding region for some transgenes. This improvement and the continued advancement of predictive DL algorithms should help clear a path for better manufacturability of rare disease gene therapies and create more viable therapeutics in the years to come.

Conclusion

Here, we've presented the development of a DL algorithm that can help cell and gene therapy developers predict the truncation propensity of their AAV constructs. This new tool for cell and gene therapy candidate validation can help improve the pre-clinical development and manufacturability of promising, first-in-class therapeutics. As DL techniques are increasingly applied to the growing pipeline of experimental cell and gene therapies, the cost and time required to bring these novel therapies to market will continue to drop, driving reduced patient costs and broader clinical implementation.

References

1. **Bluebird to withdraw gene therapy from Germany after dispute over price.** Biopharma Dive website: Published April 20, 2021. Accessed November 4th, (2022).
2. **Bluebird bio Announces FDA Approval of ZYNTEGLO®, the First Gene Therapy for People with Beta-Thalassemia Who Require Regular Red Blood Cell Transfusions.** Bluebird bio website: Published August 17th, 2022. Accessed November 4th, (2022).
3. **Bluebird wins U.S. approval for a gene therapy to treat patients with a rare blood disorder**. Stat website: Published August 17th, 2022. Accessed November 4th, (2022).
4. Greenberg B, Yaroshinsky A, Zsebo KM, et al. **Design of a phase 2b trial of intracoronary administration of AAV1/SERCA2a in patients with advanced heart failure: the CUPID 2 trial (calcium up-regulation by percutaneous administration of gene therapy in cardiac disease phase 2b)**. *JACC Heart Fail* 2(1), 84-92 (2014).
5. Rodrigues GA, Shalaev E, Karami TK, Cunningham J, Slater NKH, Rivers HM. **Pharmaceutical Development of AAV-Based Gene Therapy Products for the Eye**. *Pharm Res* 36(2), 29 (2018).
6. **Castle MJ Adeno-Associated Virus Vectors: Design and Delivery. 1st ed. Humana Press, (2019).**
7. Srivastava A, Mallela KMG, Deorkar N, Brophy G. **Manufacturing Challenges and Rational Formulation Development for AAV Viral Vectors**. *J Pharm Sci* 110(7), 2609-2624 (2021).
8. Marks P. **Enhancing gene therapy regulatory interactions**. *Expert Opin Biol Ther* 22(9):1073-1074 (2022).
9. Voineagu I, Narayanan V, Lobachev KS, Mirkin SM. **Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins**. *Proc Natl Acad Sci U S A* 105(29), 9936-9941(2008).
10. Xie J, Mao Q, Tai PWL, et al. **Short DNA Hairpins Compromise Recombinant Adeno-Associated Virus Genome Homogeneity**. *Mol Ther* 25(6), 1363-1374 (2017).