Regression-Based Modeling of Antisense Oligonucleotide Efficacy Using Sequence, Structural, and Off-Target Features

Ava George
Computer Science
Fullerton College
Fullerton, USA
02044636@fullcoll.edu

Dr. Yu Bai
Computer Engineering
California State University, Fullerton
Fullerton, USA
ybai@fullerton.edu

Jialin Tang
Computer Engineering
California State University, Fullerton
Fullerton, USA
tjl_0516@csu.fullerton.edu

2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC) | 979-8-3315-9397-1/26/$31.00 ©2026 IEEE | DOI: 10.1109/CCWC67433.2026.11393870

***Abstract*—Antisense oligonucleotides (ASOs) are a promising class of nucleic acid–based therapeutics that regulate gene expression by binding target mRNAs, with applications in genetic and rare diseases. However, designing effective ASOs remains difficult due to the vast combinatorial space of sequences, secondary structures, and chemical modifications. Recent work has leveraged deep learning and graph neural networks to address these challenges. Building on this foundation, the present project explores a complementary pipeline using classical machine learning and statistical methods for ASO design and evaluation. The workflow integrates multiple computational stages: retrieval of target mRNA sequences from NCBI, interaction prediction using the miRanda algorithm, and structural analysis via ViennaRNA. Off-target interactions were systematically assessed, and custom Python scripts were developed to merge outputs into a unified dataset. Feature engineering incorporated both numeric and categorical predictors, such as cell line and density, enabling model testing of inhibitory efficiency. Features of sequence, structure and off-target interactions trained multiple regressors including Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting. Models were evaluated using nested cross-validation with group-aware splits to prevent leakage. Random Forest achieved the highest predictive performance to predict inhibition outcomes ($R^2 \approx 0.627$, MAE $\approx$ 9.47). These results highlight both the feasibility and the challenges of applying interpretable machine learning techniques to ASO design, particularly in the presence of substantial missing data. Future directions include exploring a normalized hybridization energy gradient with relative energy per nucleotide. This work demonstrates the potential for combining bioinformatics tools, structural modeling, and machine learning to advance the rational design of therapeutic ASOs.**

Keywords*—, *Artificial intelligence, Antisense oligonucleotides, Machine Learning, Off-Target Effects, Permutation Importance, RNA Secondary Structure

I. Introduction

Antisense oligonucleotides (ASOs) are a promising class of drugs that can bind to mRNA sequences at specific locations of base pairs. ASOs are typically 15-30 bases long and are chemically modified to increase stability, improve target affinity and bioavailability.

Nucleic acid-based technologies have grown in popularity for therapeutic and diagnostic treatments. Synthetic oligonucleotides are typically 8–50 nucleotides in length. Antisense oligonucleotides (ASOs) are synthetic oligonucleotides that can precisely target an mRNA transcript. ASOs prevent mRNA from being translated into proteins and can therefore inhibit gene expression [1]. Steric oligonucleotides are ASOs that regulate the expression of key proteins and can also be used to repair defective RNA and eliminate disease-associated splice variants by modulating pre-mRNA splicing [2].

Rare diseases are estimated to affect millions worldwide despite being individually uncommon. Most rare diseases are genetic in origin [3]. Rare diseases are challenging to diagnose and treat due to their low prevalence and the pathogenic genetic alterations in single genes. Artificial intelligence and advanced analytics are advancing rare disease therapy development by identifying novel therapeutics for screening and evaluation [4].

ASOs have the ability to target and modulate specific genes, offering potential therapeutic applications in the treatment of rare diseases. Since 1998, 16 antisense treatments have received FDA/EMA approvals for genetic disorders [5]. Chemical modifications to the backbone and sugars have improved the stability, large-scale synthesis, and cellular uptake of ASOs. For example, ASO degradation by nucleases can be prevented by creating phosphorothioate linkages. Further modifications have increased target affinity and biostability of ASOs. Risks

979-8-3315-9397-1/26/$31.00 ©2026 IEEE

from off-target binding are associated with hybridization-dependent toxicities in the liver, kidney, and immune system. Shorter ASOs reduce the risk of off-target effects and allow for improved cellular uptake [6].

Current ASO therapies and gapmer design approaches continue to be refined [5][7]. However, designing high-efficacy ASOs poses a challenge due to the vast chemical space of RNA sequences occupied by combinations of the four nucleotide bases (A, U/T, G, or C). Dataset features can be created from sequence, structural, and off-target binding properties. Regression-based models can provide predictive solutions, and chemical modifications are often represented using the Simplified Molecular Input Line Entry System (SMILES) [8].

II. METHODLOGY

A. Data Collection and Preprocessing

Antisense oligonucleotide experimental data used to build the regression model was obtained from the publicly available *experiments_with_smiles.*csv file hosted on the Spidercores ASOptimizer repository. The dataset provides experimentally validated in vitro ASO results annotated with key fields such as target gene, ASO sequences, SMILES representation, experimental conditions and measured outcomes. Example experimental conditions include ASO concentration, cell line, transfection reagent.

This study focused primarily on ASOs targeting the Hepatitis B Virus (HBV) transcript, which are explicitly denoted in the dataset. The ASO entries included inhibition measurements across multiple concentrations and replicate conditions in HepG2 cells, transfected using LIpofectAMINE2000. A subset of the dataset containing ASOs against other transcripts like SNHG14 were used for comparative analysis of the model.

Using a Python script, the data entries were filtered to extract ASO sequences relevant to HBV. Duplicate ISIS identifiers were removed and sequences were then exported in a FASTA format for analysis.

To prepare the target transcripts, the full-length HBV and SNGH14 RNA sequences from NCBI's nucleotide database in FASTA format.

B. Feature Engineering

I. On-Target Binding Energy (x1)

The miRanda algorithm was employed to assess binding interactions between each ASO and the HBV RNA transcript. The program computes an alignment score and thermodynamic free energy (△G) of hybridization and returns the best-scoring hit per ASO. The free energy values were then extracted and stored as feature x1 representing the predicted hybridization strength between the ASO and its intended target.

II. Off-Target Binding Energy (x2)

To evaluate the risk of unintended interactions, the miRanda scans were repeated using non-target RNA sequences. The resulting alignment energies were parsed and calculated the mean of the top five most stable off-target interactions per ASO to reflect the off-target binding propensity.

III. Structural Accessibility (x3)

RNA structural accessibility was quantified by folding HBV transcripts into secondary structures using ViennaRNA. Using the resulting .ct files and the binding coordinates provided by miRanda, we computed the proportion of unpaired nucleotides within each ASO binding site, normalized by ASO length. For each ASO, multiple structures were averaged to account for folding variability.

C. Data Integration

The response variable (y) was inhibition percentage to indicate the degree of gene knockdown. All features were merged using the ISIS identifier as a primary key and merged into a single dataset. The dataset omitted missing values and anomalous inhibition scores. Duplicate ISIS identifiers were removed, and replicate measurements were merged by averaging inhibition percentages to reduce bias. Sequences for other transcripts, such as SNHG14, were retained for comparative analysis.

Data splits for machine learning followed an 80/20 train/test split, with group-aware cross-validation applied during training to prevent data leakage between replicate sequences. All numeric features were normalized using StandardScaler within the ML pipelines. All scripts for data preprocessing and feature extraction are publicly released for reproducibility.

D. Machine Learning Models

A variety of regression models were created to evaluate the relative contribution of sequence, structural, and off-target features on ASO efficacy. Linear Regression was used as the baseline, followed with Ridge, Lasso, Random Forest and Gradient Boosting. Experimental categorical variables were encoded using one-hot representations to allow their inclusion in regression frameworks. By comparing the varying model fits across different feature subsets, the predictive power of each model, feature class and variables were assessed.

Models were evaluated using cross-validation with GroupKFold in the outer CV to prevent leakage across replicates and KFold in the inner CV for hyperparamter

tuning via GridSearch CV. Ablation studies were conducted by retraining models using only specific feature groups (sequence, structural, off-target) to quantify their individual contributions.

III. Results

Two sequences of HBV-targeting ASOs were identified in the dataset. They were tested in HepG2 cells under LipofectAMINE 2000 transfection, with inhibition data recorded at varying concentrations.

1) Nested CV Performance Table

The regression framework created successfully incorporated on-target binding energy and off-target interactions. Preliminary analyses suggested that structural accessibility (x2) exhibited a strong correlation with inhibition efficacy than sequence binding energy alone, consistent with prior studies. Additionally, inclusion of off-target binding energy (x2) reduced unexplained variability and highlighted the necessity of specificity in ASO design.

TABLE I.

Model	R^2 mean ± std	RMSE mean ± std	MAE mean ± std
Linear Regression	0.0644 ± 0.0292	22.2317 ± 1.3462	17.1050 ± 0.9289
Ridge	0.0706 ± 0.0251	22.1551 ± 1.2750	17.0753 ± 0.9160
Lasso	0.0684 ± 0.0248	22.1793 ± 1.2454	17.1428 ± 0.8938
Random Forest	0.4882 ± 0.0398	16.4154 ± 0.8994	12.1974 ± 0.8796
Gradient Boosting	0.4835 ± 0.0387	16.4882 ± 0.7724	12.2943 ± 0.7181

The methodology created successfully demonstrates the feasibility of using regression-based models to integrate experimental factors into ASO efficacy predictions. Applied specifically to HBV, this approach provides a computational pipeline for screening and prioritizing ASO candidates before costly in vitro and in vivo experiments.

2) Permutation Importance

Permutation importance highlighted the top predictive features, showing that off-target binding energy (x2) and structural accessibility (x3) contributed most to the model's predictive power. Sequence-based features (x1) were also informative but less influential.

IV. Conclusion

A regression machine learning pipeline was created to successfully evaluate antisense oligonucleotide inhibition profiles using a curated dataset of experimental results. Future work should incorporate non-linear models such as Random Forests or gradient boosting frameworks which are better suited for capturing feature engineering relationships. Moreover, improving such pipelines and nucleic acid base technologies are essential to accelerate the design of antisense therapeutics.

V. References

[1] P. Crooke et al., "Antisense drug discovery and development," Nat. Rev. Drug Discov., vol. 16, pp. 419–437, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667416/

[2] S. Crooke, "The therapeutic potential of steric blocking oligonucleotides," Nat. Rev. Drug Discov., vol. 12, pp. 829–844, 2013. [Online]. Available: https://www.nature.com/articles/nrd3625

[3] S. N. Wakap et al., "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database," Eur. J. Hum. Genet., vol. 28, pp. 165–173, 2020. [Online]. Available: https://www.nature.com/articles/s41431-019-0508-0

[4] R. F. Bishop et al., "Artificial intelligence for rare disease therapy development," Paediatr. Perinat. Epidemiol., vol. 35, e15974, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/apa.15974

[5] J. Khvorova and A. Watts, "The chemical evolution of antisense therapeutics," Mol. Ther. Nucleic Acids, vol. 7, pp. 1–13, 2017. [Online]. Available: https://www.cell.com/molecular-therapy-family/molecular-therapy/retrieve/pii/S1525001617301223

[6] S. McDowall, M. Aung-Htut, S. Wilton, and D. Li, "Antisense oligonucleotides and their applications in rare neurological diseases," Front. Neurosci., vol. 18, 1414658, 2024. [Online]. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11456401/

[7] S. Lauffer et al., “Possibilities and limitations of antisense oligonucleotide therapies for monogenic disorders,” Commun. Med., vol. 4, Art. 6, 2024. [Online]. Available: https://www.nature.com/articles/s41431-019-0508-0

[8] Spidercores, “ASOptimizer dataset repository,” 2024. [Online]. Available: https://github.com/Spidercores/ASOptimizer

[9] G. Hwang et al., “ASOptimizer: Optimizing antisense oligonucleotides through deep learning for IDO1 gene regulation,” Mol. Ther. Nucleic Acids, 2024 Jun 11. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11066473/