**Project Title:**

**Neural Transformers and State-of-Art LLMs for Therapeutic Antisense Oligonucleotides**

**Literature Search**

**Abstract** Antisense oligonucleotides, or ASOs, are a promising class of drugs, and can target and modulate genes associated with various diseases. Large Language Models, or LLMs, had high utility for ASOs with drug discovery applications in healthcare instead of the traditional approaches with linear models and Edge Graph Transformer Neural Network. ASOs target ribonucleic acid (RNA) and form ASO-RNA heteroduplexes, which act as substrates for RNase enzymes in the cytoplasm. RNases degrade RNA in the heteroduplex with an advantageous gapmer design that has a central region of unmodified nucleotides and aid in RNase H1 activity. This consequently increases the binding affinity and resistance properties of enzymes using a phenomenon known as flanking. Two major stages of ASO design were used: sequence engineering and chemical engineering with SMILES initially used for molecular embeddings across large language models for optimizing the ranking algorithm with ASOptimizer. However, SMILES was not advantageous to produce efficacy score predictions for ASO binding. ChemDraw and HELM RNA strings were used as well, however, those on its own could not produce the efficacy score predictions due to not having a target gene or transcript with an accessionMap. Sequence alignment helps the model pinpoint how or where the ASO binds, which was completely missing from the SMILES approach. Hence, DNA sequences with target genes using an accessionMap were used to capture better sequence-function relationships with specific DNA sequences for each row of the datasets instead of chemical similarity with SMILES.

PFRED (Pfizer Functional RNAi Enumeration and Design Tool) is an open-source desktop platform for designing efficient siRNA and antisense oligonucleotides, allowing researchers to input a target gene with an Ensembl ID, generate sequence of libraries of ASOs and siRNAs, and apply specific bioinformatics filters for stability and off-target binding effects. Prominent features of this desktop platform are that GUI (Java-based) is used for users to interact locally with a Docker-based server for running computationally intensive on the cloud or AWS. It is fully compatible with REST APIs and can be locally deployed via Docker. Design workflows start with a gene ID and retrieve ortholog sequences with k-mer oligos enumerated, thereby computing SNP filtering, GC content, intron/exon alignment, thermodynamics stability using OligoWalk, off-target hits via Bowtie, and sequence toxicity motifs like poly-G and poly-T. In order to calculate Efficacy Predictions, a Support Vector Machine was utilized for multiple descriptors. Sequence motifs and thermodynamic properties were incorporated and trained on datasets like AOBase with more than 500 ASOs across 46 gene targets, 522 to be precise. Advanced visualization tools were used with HELM notation integrated for modified nucleotides like LNA and 2’-OMe. Multiple oligo views were provided with sequence-based, block-based, and monomer-based. Users potentially can export selected oligos via CSV, apply mathematical formulas for scoring, and customizing filters to effectively handle constraints. Challenges in therapeutic oligo design like unintended off-target binding, cross-species activity, and design of chemically modified oligos were speculated with improvements integrated from prior tools like siDirect, BLOCK-iT, and OptiRNAi due to a more modular and extensible environment.

The LLMs used after the prior work with MolEval and models like ChemBERTa, RoBERTa, and Molformer were Galactica-6.7B, LLaMA2-7b-hf, and GPT-3.5-Turbo. The GPT models produced significantly better results than Galactica-6.7B and LLaMA2-7b-hf. GPT models are more suitable for these numerical regression tasks and prompting methods with zero-shot, few-shot, and chain of thought reasoning. Galactica-6.7B is not as suitable for these types of tasks and is more a scientific model rather than a numerical one. The model is only given the task instructions with the DNA tags and accessionMap for gene mapping with a single test input for zero-shot prompting. There were not any examples given for zero-shot prompting with the goal of predicting a efficacy score as a reference for the respective model and dataset. With few-shot prompting, the model is given a few example DNA strings and efficacy score as input to predict an output efficacy score based on what was given.

Several other models like MolLM, Llasmol, T0pp, and Tx-LLM were all used to determine which would be the top contenders for numerical regression tasks. MolLM is more of a multimodal pre-trained model with encoding architecture that handles texts and molecular graphs with 2D and 3D provisions. The overarching goal of utilizing MolLM as a literary work is to show that 3D information shows better performance on molecular tasks. LlaSMol helped close the gap with domain-specific models in chemistry, achieving near SOTA architectures with only 0.58% of its model parameters finetuned. T0pp is very advantageous for Named Entity Recognition and achieved first place and second place in relation extraction and novelty detection. The main goal of using T0pp is to show that task-specific fine-tuning can beat generalist LLMs like LLaMA2 and GPT for general-purpose tasks. Tx-LLM was fined-tuned with Google’s PALM to work across 66 different therapeutic tasks with an instruction-tuning framework and achieved or exceeded SOTA on 43 out of 66 therapeutic tasks.

**Introduction**

ASOs have demonstrated a unique ability to utilize sequence-binding on the target DNA and use the gene mapping with accession to create a more optimal efficacy. As ASO binding strength increases, diseases can be targeted at DNA level itself since the flow of transcription and translation is from DNA to RNA to protein. Several therapeutic strategies for targeting downstream processes can be identified and the limitations with lack of proper sequence-function relationships and sequence alignment can easily be fixed. Earlier chemical transformer models like ChemBERTa and RoBERTa attempted to harness the chemical structure with ASOs. Models like LLaMA2-7b-hf and Galactica-6.7B aimed to target a downstream regression task and improve better prompting methods with zero-shot, few-shot, and chain of thought reasoning. Baseline results gathered from the PFRED.csv, openASO.csv, and ASOptimizer.csv were used to determine if there was an improvement in the efficacy binding results compared to the original model. Prior to incorporating prompting, embeddings were utilized with the SMILES strings for possibly obtaining an efficacy score, which was found to not be possible with the utilization of the MolEval dataset. As mentioned earlier, Gibbs Free Energy and MIRANDA were used with a ranking algorithm for the ASOptimizer framework with sequence and chemical engineering. Models like GPT-4o and GPT-3.5 are more suitable for numerical regression tasks and the specific prompting methods that are used.

Researchers in the past relied more on their individual expertise with molecular biology and physical observations for generating potential ASO candidates, however, this approach quickly became obsolete with the chemical space exponentially increased with a multitude of RNA sequences. Each position occupied by one of four nucleotide bases with ASOs of length l led to 4l potential combinations. All ASO sequences that were examined were less than 20 in the past with this primitive approach. Thus, computational approaches that researchers subsequently developed significantly lowered the vast number of ASO candidates for experimentation.

The ASOptimizer framework evaluated therapeutic ASOs at the sequence and molecular level in two stages: linear factor with miRNA binding sites and a deep graph neural network accepting the molecular graph of ASOs as input and predicting the efficacy as output. The IDO1 expression levels and cytotoxicity were compared to the standard gapmer approach where ASOs less than 20 base pairs and longer targets are used for calculating Gibbs Free Energy with MIRANDA. Thermodynamic stability and structural accessibility partially explain ASO mediated gene regulation.

From the basic framework, LLMs: both chemical transformer and general-purpose models were used with the MolEval and ChemLLMBench frameworks to understand the SMILES strings and embeddings application to find enhanced ASO binding patterns. With respect to MolEval, ChemBERTa, RoBERTa, LLaMA2, Mol2Vec, and Morgan were used with libraries like RDKit used for molecular fingerprints. This approach focused more on chemical structures and similarities on that front and was evaluated in comparison to the Graph Edge Transformer and MIRANDA approach with Gibbs Free Energy for sequence and chemical engineering.

**Previous work by others**

ASOptimizer integrated several modules and predictors to achieve the goals of thermodynamic, structural, and sequence-based models. Gibbs Free Energy was used to measure thermodynamic stability of ASO-mRNA duplex, with a negative Gibbs Free Energy indicating a more stronger and stable hybridization, correlating to a higher inhibition. ASOptimizer calculated the hybridization energy between the ASO and target mRNA with tools like MIRANDA – miRNA target prediction algorithm. The goal with MIRANDA was to identify microRNA target binding sites with sequence complementarity and thermodynamic binding energy. MIRANDA helped offer fast and modular prediction of target binding while handling lack of perfection with base pairs. ASOptimizer’s sequence engineering involves the following: selecting and modifying base compositions, motifs, and backbone patterns like GC content. ASOptimizer penalizes sequences prone to off-target binding and ranks candidate ASOs using these following features: Duplex stability, RNA accessibility, and location on intron/exon. Chemical engineering refers to chemical modifications that stability and binding affinity with LNAs (Locked Nucleic Acids), PS linkages (Phosphorothioate). ASOptimizer uses the HELM RNA notations to help integrate the final inhibition efficacy pattern with the following inputs: DNA sequence, target site accessibility, change in Gibbs Free Energy from MIRANDA, sequence motifs, and modification patterns with models like SVM or Random Forest creating a list of ranked set of ASOs with predicted inhibition efficacy scores.

openASO contains ASO sequences, efficacy labels, and metadata across multiple targets without computing change in Gibbs Free Energy explicity unlike ASOptimizer’s approach. It uses the similar approach as ASOptimizer for sequence and chemical engineering with favoring specific GC content, optimizing binding site location, and enabling positional analysis with motifs for more effective ASOs. The ASOeffective score shows the biological dependencies with the gene sequences and targets with and without chemical modifications. PS backbones and 2’-O methyl in-silico were included before testing on RNA structure tools. Current openASO datasets do not include full HELM annotated structures, which is a possibility with future versions. Sequence engineering is a core component of openASO dataset with the ASOseq column.   
   
 PFRED also employed the MIRANDA and sequence/chemical engineering phenomenon like openASO and ASOptimizer. However, PFRED avoids self-complementary regions and CpG dinucleotides, using GC% range constraints and improved location bias towards exons and coding regions. Orthologs are conserved and aligned properly. Each candidate oligo is scored by multiple sequence rules and MIRANDA’s purpose of filtering candidates by binding specificity and off-target avoidance is ingrained with PFRED’s approach. Without explicitly embedding MIRANDA, alignment based off-target scanning is performed with Bowtie, SNP filtering, and accessibility filtering. PFRED’s SVM efficacy based model take in the change in Gibbs Free Energy with only favorable ASOs that are strengthened to bind to a target but not excessively strong to cause off-target issues included in the SVM’s selection process. This SVM directly calculates the Gibbs Free Energy using OligoWalk and computes the Gibbs Free Energy change for ASO self-folding and target-accessibility. PFRED’s outputs can be visualized with block-style and monomer level annotations. HELM support and scoring is entirely integrated for PFRED unlike openASO where it is not structurally annotated.

MolEval and ChemLLMBench were the two major frameworks that inspired these findings for R squared and RMSE scores with the three datasets as well as the prompting strategies. MolEval was highly useful to understand the applicability of SMILES strings and embeddings to create enhanced ASO binding efficacy predictions. Several prior models like ChemBERTa, RoBERTa, LLaMA2, Mol2Vec, and Morgan were evaluated to determine the most promising model for our study prior to proceeding with prompting strategies and DNA sequences with genes as a more feasible approach. Evaluating other models like MolLM, Llasmol, and T0pp as a viable tool for prompting and instruction-based approach were crucial to realize that Galactica, LLaMA, and GPT models were the top three contenders to include in this research study. Other models like MPT-7B and Mistral were also evaluated, however, none of these produced any results and could not be included in this report for that reason.

MolLM is a multimodal pre-trained language model that combines biomedical data with 2D and 3D structures. It encapsulates both chemical and natural language into one model for a broad set of tasks with two encoders: a text transformer encoder for biomedical literature and a molecular graph transformer for specific 2D/3D molecular graphs. Contrastive learning was used for 160K molecule-text pairs for training purposes with SMILES strings, molecular graphs (2D), 3D coordinates, and pertinent biomedical texts as input. MolLM was the first model to learn efficiently from text, 2D, and 3D molecular data. It performs well on molecular property prediction, captioning molecules, cross-modal retrieval, and molecule editing with text prompts. Utilizing 3D information will boost performance on molecular tasks.

Tx-LLM is a general purpose LLM model for drug discovery, which was built by fine-tuning PaLM-2. It works across 66 therapeutic tasks with a unified instruction-tuning framework. PaLM2 is Google’s large scale LLM that serves as a base model for a training dataset with 709 datasets from the Therapeutic Data Commons (TDC) with instruction-answer pairs. Tx-LLM is capable of handling small molecules, proteins, DNA/RNA, diseases, and cell lines and supports classification, regression, and generalization tasks. The instruction tuning used for Tx-LLM is the following: TxT prompt format with instruction, context, question, and answer. Tx-LLM exceeds the SOTA architecture on 43 of 66 therapeutic tasks and can successfully combine SMILES and free text. It can also handle positive transfer across drug states.

T0pp is an advantageous NLP system with ensemble learning and a fine-tuned T0pp model – based on the T5 architecture series – for extracting biomedical data to build knowledge graphs. Two phases exist – Named Entity Recognition with BioBERT, PubMedBERT, and BioM-ELECTRA and Relation Extraction with Novelty Detection with a finetuned T0pp model (11B parameters) with rule based methods and Random Forest. T0pp achieved first place in Named Entity Recognition, second place in Relation Extraction and novelty detection and showed that task-specific fine-tuning beats general purpose LLMs like ChatGPT-3.5 and GPT-4. However, due to time constraints, this could not be experimented on to contribute to the Results section: only three models were used for zero-shot and few-shot prompting: Galactica-6.7B, LLaAM2, and GPT-3.5-Turbo.

LlaSMol is a type of instruction-tuned large language model for chemistry tasks, trained on SMolInstruct, and this has the potential to outperform GPT-4 and Claude-3 for chemistry related benchmarks. SMolInstruct Dataset has 3.3 million examples across 14 chemistry tasks with name conversion, property prediction, chemical reactions, and natural language covered. Fine-tuned LLMs with LoRA were used on SMolInstruct with base models including Mistral, Galactica, LLaMA2, and Code LLaMA. LlaSMol was found to beat GPT-4 and Claude-3-Opus on almost all tasks like property prediction and SMILES translation, almost achieving state of the art results with fine-tuning occurring only for 0.58% of the model’s parameters. Canonical SMILES and task diversity are very important.

ChemLLMBench serves as a benchmark for an open-source evaluation framework to evaluate the performance of Large Language Models on real-world chemistry tasks. It was introduced in the NeurIPS 2023 paper and assessed LLMs like GPT-4, GPT-3.5, DaVinci, LLaMA2, and Galactica on realistic chemistry applications – of which GPT-3.5-Turbo, LLaMA2, and Galactica were the three models utilized to prepared results for this report. The benchmark LLM performance on eight diverse tasks tested the following metrics: understanding, reasoning, and explaining. Naming molecules, predicting chemical reactions, and generating molecule captions are all examples of these metrics. The eight tasks that are evaluated are name prediction, property prediction, yield prediction, reaction prediction, retrosynthesis, reagents selection, molecule captioning, and molecule generation. Datasets used for these were the following: PubChem, BBBP, HIV, Tox21, (HTE-BH, Suzuki), USPTO, USPTO-50K, Suzuki-Miyaura, and ChEBI-20. The evaluation metrics used were Accuracy, F1, Validity, BLEU, human evaluation, Levenshtein, and Exact Match. Zero-shot prompting where a model gets a task description and input, Few Shot In Context Learning with random sampling and scaffold based retrieval, and LLM model benchmarks tested across GPT-4, GPT-3.5, DaVinci-003, LLaMA2-13B-Chat, and Galactica-30B were all evaluated with this ChemLLMBench framework. GPT-4 consistently ranked the highest but failed with SMILES related tasks and name predictions. LLMs inherently struggle with raw SMILES strings due to poor tokenization approaches and lack of chemical grammar encoded within the LLM architecture. Few-shot prompting typically improves performance. LLMs are not as reliable with critical tasks like synthesis due to its tendency to hallucinate chemical outputs regularly. It performed phenomenal with molecule captioning and property classification.   
  
**Results**

Baseline MSE, RMSE, and R2 for Test and Validation for the PFRED Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test RMSE | Test MSE | Test R2 | Validation RMSE | Validation MSE | Validation R2 |
| **0.2422** | 0.06 | 0.28 | 0.2828 | 0.08 | 0.21 |

Transcript-Level MSE, RMSE, and R2 for the openASO.csv

|  |  |  |
| --- | --- | --- |
| Transcript-Level MSE | Transcript-Level RMSE | Transcript-Level R2 |
| 0.0521 | **0.2283** | 0.3028 |

Normalized Per-Gene Baseline for MSE, RMSE, and R2 for the ASOptimizer.csv

|  |  |  |
| --- | --- | --- |
| Normalized Per-Gene Baseline for MSE | Normalized Per-Gene Baseline for RMSE | Normalized Per-Gene Baseline for R2 |
| 0.0437 | **0.2091** | 0.4020 |

**PFRED Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Zero-Shot RMSE | Zero-Shot R2 | Few-Shot RMSE | Few-Shot R2 |
| Galactica-6.7B | 0.4745 | -1.4528 | 0.4302 | -1.3388 |
| LLaMA2-7b-hf | 0.4534 | -1.6451 | 0.4212 | -1.2588 |
| GPT-3.5-Turbo | 0.3275 | 0.3637 | 0.3591 | 0.6381 |

**openASO Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Zero-Shot RMSE | Zero-Shot R2 | Few-Shot RMSE | Few-Shot R2 |
| Galactica-6.7B | 0.5925 | -3.6965 | 0.3542 | -0.6784 |
| LLaMA2-7b-hf | 0.3635 | -0.7860 | 0.3431 | -0.5859 |
| GPT-3.5-Turbo | 0.2822 | -0.0651 |  |  |

**ASOptimizer Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Zero-Shot RMSE | Zero-Shot R2 | Few-Shot RMSE | Few-Shot R2 |
| Galactica-6.7B | 0.4478 | -1.7443 |  |  |
| LLaMA2-7b-hf |  |  |  |  |
| GPT-3.5-Turbo | 0.3185 | 0.3471 | 0.3506 | 0.6340 |

**Methods**

**1. Evaluating all Possible Models**The following three models – Galactica-6.7B, LLaMA2-7b-hf, and GPT-3.5-Turbo – were used on the three datasets– PFRED.csv, openASO.csv, and ASOptimizer.csv – as a basis to evaluate ASO binding efficacy strength with DNA sequences. Different tags were used before feeding the DNA sequences from these three datasets for each respective model – Galactica-6.7B used [START\_DNA] and [END\_DNA], LLaMA2-7b-hf used [DNA\_START] and [DNA\_END], and GPT-3.5-Turbo used a plain-text DNA sequence with DNA Sequence as the appropriate tag before inserting the DNA sequence. In order to arrive at these models as the top three contenders, it was necessary to evaluate several other models and examine their reasoning capabilities and prompt handling effectiveness.

Methods – talk about how the datasets were constructed, how the dataset were edited by me, go through and highlight strategies and methods

**2. Used SMILES Strings with Embeddings with DNA Sequences**

The MolEval dataset was utilized with its classification and regression metrics for SMILES (Simplified Molecular Input Line Entry System) strings with tokenized embeddings. SMILES, in essence, are textual representations of chemical structures, which are converted into vector representations, or embeddings, with two types of commonly used tokenization strategies: character-level tokenization and atom-level tokenization. With character level tokenization, every character is treated as a separate token and with atom-level tokenization, and atomic symbols and structural annotations are grouped. Subsequently, the tokenized characters or atoms in the SMILES strings are mapped to numerical indices based on a vocabulary. The encoded indices are then passed through the embedding layer, with each token index mapped to a dense vector.

This formula is used to represent an embedding matrix, E:  
  
E∈R∣V∣×dwhere |V| is the vocabulary size and d is the embedding dimension.  
  
With this approach, the input DNA sequences from the three datasets – PFRED.csv, openASO.csv, and ASOptimizer.csv - were used to create SMILES strings with ChemDraw for Excel. PFRED.csv has 522 DNA sequences, while openASO.csv and ASOptimizer.csv have 1708 and 1267, respectively.   
  
  
**3. Used SMILES Strings as Embedding with HELM RNA Sequences**

HELM, or Hierarchical Editing Language for Macromolecules, is designed exclusively for biological macromolecules with DNA, RNA, and peptides, with structural modifications denoted clearly. SMILES strings, unlike HELM, are used for small molecules and standard chemical structures.

To convert the RNA sequences into SMILES, nucleotides needed to be translated into explicit chemical structures that SMILES notation could encode.   
  
HELM RNA Sequences were converted to SMILES strings with the following steps:

**a. Parse HELM string**  
  
A sample parsed HELM string without any modification would be the following:  
  
RNA1{r(A)p.r(C)p.r(C)p.r(G)p.r(G)p.r(A)p.r(U)p.r(G)p.r(G)p.r(C)p.r(G)p.r(G)p.r(U)p.r(G)p.r(G)p.r(C)p.r(U)p.r(G)p.r(C)p.r(U)p}$$$$V2.0  
  
This denotes an RNA strand with four nucleotides – Adenine, Guanine, Cytosine, and Uracil – connected by phosphate bonds.   
  
A sample parsed HELM string with modifications would look like the following:  
  
RNA1{r(U)[sp].r(U)[sp].r(C)[sp].r(A)[sp].r(G)[sp].r(G)[sp].r(C)[sp].r(A)[sp].r(G)[sp].r(U)[sp].r(U)[sp].r(G)[sp].r(U)[sp].r(U)[sp].r(A)[sp].r(A)[sp].r(A)[sp].r(A)[sp].r(U)[sp].r(A)[sp]}$$$$V2.0

This denotes an RNA strand with four nucleotides – Adenine, Guanine, Cytosine, and Uracil – with the substitution of a phosphate group with a monomer group to denote a modified oligonucleotide.  
  
**b. Identify nucleotides and backbones**

The nucleotides r(A), r(C), r(G), and r(U) are translated into a canonical SMILES with a backbone P for phosphate linkage.

**c. Nucleotides need to be connected**  
  
In order to form the complete RNA strand in SMILES notation, connect each nucleotide with phosphate linkages, via phosphodiester bonds.   
  
  
  
A sample SMILES string looks like this:

C1=NC2=C(N1)N=CN=C2C=OC1=NC(=O)C(C1=O)=C2N=C(N2)CC1=NC(=O)C(C1=O)=C2N=C(N2)CC1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC2=C(N1)N=CN=C2C=OC1(C(=O)N(C1=O)C2=C(N=C(N2)C)C=O)C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC(=O)C(C1=O)=C2N=C(N2)CC1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1(C(=O)N(C1=O)C2=C(N=C(N2)C)C=O)C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC(=O)C(C1=O)=C2N=C(N2)CC1(C(=O)N(C1=O)C2=C(N=C(N2)C)C=O)C1=NC2=C(N1)N=CN=C2C(=O)N1C=CN=C1C1=NC(=O)C(C1=O)=C2N=C(N2)CC1(C(=O)N(C1=O)C2=C(N=C(N2)C)C=O)

However, this approach led to very long and complex SMILES strings which are by itself not capable of generating an efficacy score for the three datasets utilized. Also, secondary structure information was lost since SMILES cannot encode secondary structure like RNA folding. Modified nucleotides in HELM with the substitution of a phosphate group need to be very tediously defined to ensure accurate SMILES representation.

**3. Optimized Efficacy Predictions with DNA Sequences and Target Genes**Due to the challenges listed above with SMILES strings, a more effective approach was utilized where DNA sequences and its respective target genes were used for the datasets for the following reasons:

**a. Biological Relevance and Specificity**  
  
DNA sequences directly capture the biological context for ASOs with respect to base-pairing. Target-gene information with accession mapping shows directly how ASOs function in a biological setting, thereby creating more accurate and meaningful predictions.

**b. Sequence-Structure Function Relationship**

ASO efficacy depends on sequence-structure function relationship with sequence complementarity to RNA or DNA targets, secondary structures, and interaction specificity.

DNA sequences accurately pinpoint the nucleotide-level interactions which SMILES is incapable of doing as it can only represent the chemical structure.

**c. Predictive Model Capacity**  
  
Galactica-6.7B and GPT-3.5-Turbo, which are trained on biological sequences and genomic contexts, understand RNA and DNA sequences much better than chemical strings with SMILES. DNA sequences with target genes leverage pretrained biological knowledge in these language models, thereby increasing the accuracy with predictions.

**d. Reduction of Chemical Representation Complexity**  
  
Converting from HELM to SMILES produced very large and structurally complex relationships, creating confusing biological interactions and an unreliable prediction. Thus, using a straightforward and explicit DNA sequence and a clear biological context makes predictions much more reliable and accurate due to less ambiguity.

**e. Facilitating Few-Shot and Zero-Shot Learning**  
  
Utilizing DNA sequence and clear target-gene annotations help with zero-shot and few-shot prompting and strengthen the model’s performance with ASO efficacy binding. SMILES strings, on the contrary, lack proper biological context and clarity about the target gene, thereby creating a suboptimal prompting and poorer ASO effectiveness RMSE and R squared scores as a result.  
  
**4. Zero-Shot Prompting**

With zero-shot prompting, three models were used with the datasets – Galactica-6.7B, LLaMA2-7b, and GPT-3.5-Turbo – for ASO efficacy prediction and to create the respective RMSE and R squared scores as a metric for this numerical regression task. The ChemLLMBench framework with its respective GitHub was used to model the prompting for an instruction-based style that the model could understand. Zero-shot prompting is advantageous for a quick prototype or testing general model capabilities when limited example data is provided. Zero-shot is also useful when using models with strong domain-specific pretraining. The prompts used across all these three models were similar, with just the respective DNA start and end tags different based on the model type – Galactica-6.7B had [START\_DNA] and [END\_DNA], LLaMA2-7b had [DNA\_START] and [DNA\_END], and GPT-3.5-Turbo used a plain-text DNA Sequence as the tag. HuggingFace was utilized for Galactica-6.7B and LLaMA2-7b-hf and GPT-3.5-Turbo was utilized with OpenAI. The only difference between Galactica-6.7B and LLaMA2-7b-hf was that the former was open-source and not a gated model, while LLaMA2-7b-hf needed a huggingface token to run. GPT-3.5-Turbo required an OpenAI API Key.

Galactica integrated scientific context and reasoning decently, enabling specific numerical predictions for DNA sequences with its pretraining on biological literature. Galactica’s strength is that it has a proven record of specialized scientific training with papers, chemical data, and bioinformatics knowledge. Galactica’s prompt format was clear and domain-specific phrasing as follows:   
  
You are a molecular biology expert predicting inhibition efficacy of antisense DNA sequences.\n"

"Only respond with a numeric score between 0 and 1 (e.g., 0.42). Do not provide explanation or units.\n\n"

f"[START\_DNA]{generatedSequence}[END\_DNA]\nPredicted inhibition efficacy:"  
  
LLaMA2-7b-hf is a general-purpose model with a proven ability to handle structured prompts efficiently. LLaMA2’s prompt format is structured with explicit instructions as follows:  
  
You are a molecular biology expert predicting inhibition efficacy of antisense DNA sequences.\n"

"Only respond with a numeric score between 0 and 1 (e.g., 0.42). Do not provide explanation or units.\n\n"

f"Target gene: {generatedGeneSequence}\n[DNA\_START]{generatedSequence}[DNA\_END]\nPredicted inhibition efficacy:"

GPT-3.5-Turbo is a superior general reasoning model with advantageous natural language understanding. The prompt format for GPT-3.5-Turbo is with natural language instructions with explicit guidance as follows:

"You are a senior researcher in molecular biology, with deep expertise in antisense oligonucleotide (ASO) design.\n"

"Given the DNA sequence and its gene target, estimate the inhibition efficacy of gene expression.\n"

"Consider thermodynamics, off-target effects, hybridization potential, and accessibility.\n"

"Respond with only a number between 0 and 1 (e.g., 0.13, 0.57, 0.89).\n"

"Avoid repeating values. Avoid values like 0.75 for every entry. Use biologically diverse reasoning.\n\n"

f"DNA sequence: {generatedSequence}\n"

f"Target gene: {generatedGeneSequence}\n"

f"Predicted inhibition efficacy:"

Separate output CSV files were generated for each model and respective dataset: Galactica-6.7B, LLaMA2-7b, and GPT-3.5-Turbo with the prompt and the efficacy predictions for each DNA sequence displayed in entirety within the CSV files.

**5. Few-Shot Prompting**

With few-shot prompting, three models were used with the datasets – Galactica-6.7B, LLaMA2-7b, and GPT-3.5-Turbo – for ASO efficacy prediction and to create the respective RMSE and R squared scores as a metric for this numerical regression task. The ChemLLMBench framework with its respective GitHub was used to model the prompting for an instruction-based style that the model could understand. Few-shot prompting is more practical compared to zero-shot when there is greater access to representative labeled examples and there is scope for high accuracy and stable predictions.

Generalist models like LLaMA2-7B and GPT-3.5-Turbo are much more suitable for few-shot prompting due to their strengthened ability to handle explicit examples compared to scientific reasoning models like Galactica-6.7B. The prompts used across all these three models were similar, with just the respective DNA start and end tags different based on the model type. Unlike zero-shot prompting, few-shot prompting had a value for the number of input few-shot examples that had to be used for the models to output an efficacy score. Two major types of few-shot prompting used in the research community are random and scaffolding.   
  
Galactica’s prompt format was clear and domain-specific phrasing as follows with the few-shot appropriate specifications added into the prompt:

generatedPrompt = (

"You are a molecular biology expert predicting inhibition efficacy of antisense DNA sequences.\n"

"Only respond with a numeric score between 0 and 1 (e.g., 0.42). Do not provide explanation or units.\n\n"  
  
generatedPrompt += f"Target gene: {generatedGeneDescription}\n[START\_DNA]{generatedSequence}[END\_DNA]\nPredicted inhibition efficacy: {generatedLabel:.2f}\n\n"  
LLaMA2’s fully explicit-instruction based prompting for few shot looks like the following:  
  
generatedPromptValue = (

"You are a molecular biology expert predicting inhibition efficacy of antisense DNA sequences.\n"

"Only respond with a numeric score between 0 and 1 (e.g., 0.42). Do not provide explanation or units.\n\n"

)  
  
generatedPromptValue += f"Target gene: {generatedGeneDescription}\n[DNA\_START]{generatedSequence}[DNA\_END]\nPredicted inhibition efficacy: {generatedLabel:.2f}\n\n"  
  
GPT as a general reasoning model would follow this kind of prompting style for few-shot for a natural language based understanding:  
  
 generatedPromptValue = (

"You are a senior researcher in molecular biology, with deep expertise in antisense oligonucleotide (ASO) design.\n"

"Given the DNA sequence and its gene target, estimate the inhibition efficacy of gene expression.\n"

"Consider thermodynamics, off-target effects, hybridization potential, and accessibility.\n"

"Respond with only a number between 0 and 1 (e.g., 0.13, 0.57, 0.89).\n"

"Avoid repeating values. Avoid values like 0.75 for every entry. Use biologically diverse reasoning.\n\n"

)

generatedPromptValue += f"DNA sequence: {generatedSequence}\nTarget gene: {generatedGeneValue}\nPredicted inhibition efficacy:”

With random few-shot prompting, examples that are used for prompting are randomly selected from the dataset with no direct chemical or structural similarity evaluated with the selection of the input examples from the datasets. k examples are randomly chosen from the training set for every prediction, with the prompt for each consistent across all DNA sequences. Advantages of random few-shot approach are increased diversity with wide coverage of different kinds of data points, decreased bias due to a lowered risk to repeat similar DNA sequences, and more promising generalization for predicting ASO efficacy strength across varied sequences and target genes. Disadvantages of random few-shot approach include, but are not limited to, higher variance in predictions due to randomness and potentially creating lower accuracy when structural similarity matters.

Scaffold few-shot prompting examples are selected based on structural similarity – chemical scaffold, similar sequences, and structural motifs – to name a few. “Scaffolding” signifies only selecting examples that closely match molecular structures of the few-shot example that needs to be generated as output. Examples most structurally like the query sequence like similar nucleotide patterns are chosen for a given prediction. Scaffolding is very useful when chemical or structural similarity correlates highly with biological function and when this structural correlation can strongly predict ASO efficacy. Disadvantages of scaffolding include higher risk for bias due to the model’s tendency to select similar examples repeatedly, which limits generalization and reinforcing incorrect biases, especially when scaffold similarity is not the overarching prediction for efficacy.

In essence, random few-shot is superior to scaffold few-shot for this task since the datasets utilized for this report have varied gene and DNA sequences. Random few shot would capture the diversity of genes and DNA sequences significantly better than scaffolding and allow the model to generalize effectively across different genes and nucleotide sequences. Scaffolding bias was avoided too in my research since scaffold-based methods often repeatedly choose similar sequences and lock the model into narrower reasoning patterns. ASO binding strength is not influenced by just structural similarity and is driven by multiple complex underlying parameters. Scaffold few-shot is not capable of handling the multitude of parameters that govern ASO efficacy and random few-shot is an optimal approach for this reason. Also, random examples would provide broader contextual clues to the model and enable it to perform more optimally than scaffolding as it is very narrow with its context-based processing capabilities.

Models like Galactica-6.7B, GPT-3.5-Turbo, and LLaMA2-7b-hf require contextual cues with respect to prompting to generalize across the DNA sequences. Random few-shot selection exposes the model to different functional scenarios and makes it more robust across new sequences or genes. Scaffolding approaches for few-shot prompting led to very suboptimal predictions with the same efficacy score repeating itself and poor/negative R squared scores.

There is empirical evidence from the Results section that highlights how predictive accuracy, lower RMSE, and higher R2 resulted after utilizing random few-shot approach in my observations.

Despite utilizing random few-shot approach for all the datasets, GPT-3.5-Turbo was the only model to generate positive R2 scores for both zero-shot and few-shot. The other models – Galactica-6.7B and LLaMA2-7B – generated negative R squared scores consistently for zero-shot and few-shot for the following reasons: differences in model architecture compared to GPT-3.5-Turbo, different training methods, suboptimal ability to follow prompts compared to GPT models, lack of numerical prediction stability compared to GPT-3.5-Turbo and worse domain generalization. GPT-3.5-Turbo is heavily fine-tuned for instruction-following, strong general reasoning, and robust numerical predictions with its exemplary ability to interpret numerical instructions clearly. Galactica-6.7B is not fine-tuned specifically for numerical efficacy predictions for prompts and has very specific scientific applications. LLaMA-2-7B is more general-purpose with linguistic abilities, however, has much fewer instruction-following adaptations with lower domain-specific knowledge than GPT models. This is less optimal explicitly for numerical efficacy tasks.

**Discussion**

**PFRED.csv  
A B**

**A graph of blue dots

Description automatically generated A graph of blue dots

Description automatically generated**

**A B  
  
A graph of blue dots

Description automatically generated A graph of blue dots

Description automatically generated**

**A B  
A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated  
openASO.csv  
  
A B**A graph of blue dots

Description automatically generated **A graph of blue dots

Description automatically generated**

**A B  
A graph with blue dots

Description automatically generated** A graph of blue dots

Description automatically generated

**A B  
A graph with blue dots and a line

Description automatically generated**

**ASOptimizer.csv  
  
A**A graph of blue dots

Description automatically generated  
  
**A B**  
  
A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

Zero-shot prompting served as a basis for evaluating how Galactica-6.7B, LLaMA2-7b, and GPT-3.5-Turbo performed with respect to prompting to generate the R squared and RMSE scores. The zero-shot RMSE with respect to Galactica-6.7B was the highest for the openASO Dataset with a value of 0.5925. Few-shot RMSE improved considerably with the strict prompting methods and avoiding hallucinating responses to 0.1206. The zero-shot RMSE for openASO Dataset – Galactica-6.7B was significantly worse than the baseline transcript-level RMSE of 0.2283. Few-shot RMSE, on the contrary, was much more improved compared to the baseline-transcript level RMSE – with a value of 0.1206 instead of 0.2283. LLaMA2-7b-hf was more numerical based and and had an improved RMSE zero-shot score for openASO.csv compared to Galactica-6.7B. The zero-shot RMSE score was 0.3635 for LLaMA2-7b-hf instead of 0.5925 like Galactica-6.7B, so the starting zero-shot RMSE score was more improved due to better capability with prompting. With respect to PFRED.csv, LLaMA2-7b-hf had a very similar zero-shot RMSE compared to Galactica-6.7B, with a value of 0.4534 instead of 0.4745. Galactica-6.7B is more scientific reasoning and not as suited for numerical regression tasks. That contributed as an underlying factor for the negative R squared scores. LLaMA2-7b-hf also generated negative R squared scores due to its lack of ability to generalize efficiently with the training data points. LLaMA2-7b-hf, although powerful, lacks sufficient reasoning ability or specialized domain knowledge with respect to ASO binding and efficacy compared to larger domain-specific models like GPT-3.5-Turbo.

For this reason, GPT-3.5-Turbo never displayed a negative R squared value. Other possible factors for negative R squared score include parsing and numerical extraction errors, unsuitable or inconsistent data with outliers causing the predictions to significantly deviate from the true values, ineffective prompts that can trigger the model to misunderstand the task and create higher residuals, and hallucinated predictions. LLaMA2-7b is a generative model and many of these types of models often produce hallucinated results that are not closely aligned with the ground truth. Since this will result in more potential errors, the sum of squared residuals will be greater than the total sum of squares, hence creating a negative R2 value.

Certain models like Galactica-6.7B, LLaMA2-7B, and GPT-3.5-Turbo were used for zero-shot and few-shot prompting in antisense oligonucleotide efficacy prediction due to its ability to follow prompting instructions and its strong reasoning capabilities. LLaMA2-7b is often chosen for its strong linguistic background and its proven track record of handling diverse reasoning tasks. Galactica-6.7B is trained explicitly on scientific literature and is advantageous for biological and chemical prediction tasks. Galactica provided more scientific predictions compared to LLaMA2-7b, which was more numerically driven. GPT-3.5-Turbo was the most promising with its instruction-following approach, superior reasoning ability, and reliability in interpreting natural language contexts across various contexts.

Models like Mistral-7B, T0pp, and MPT-7B are less suited for precise, domain-specific precision tasks like ASO efficacy predictions with zero-shot and few-shot prompting. Mistral-7B is capable of general language understanding, however, does not generate as reliable an output for scientific regression tasks, creating lower reliability and accuracy in prediction-intensive scenarios. T0pp is more optimized for general multitasking prompting scenarios and knowledge based queries rather than scientific reasoning like the ASO binding efficacy strength criteria we have here. MPT-7B has greater utility with conversational and general-purpose tasks and lacks the specific scientific knowledge capable of producing accurate chemical or biological inference. MPT-7B is not as likely to produce results with efficacy binding like the three models used for this study. For this reason, Galactica-6.7B, LLaMA2-7b-hf, and GPT-3.5-Turbo were selected due to their proven track record of scientific domain expertise, versatile structured-prompt handling capacities, and general reasoning strengths. As discussed earlier, even with all these advantages, R squared repeatedly came negative for Galactica-6.7B and LLaMA2-7b-hf unlike GPT-3.5-Turbo which produced reliable and better R squared scores compared to the baseline results for the three datasets.  **References:**Cao, Zhonglin, Simone Sciabola, and Ye Wang. Large-Scale Pretraining Improves Sample Efficiency of Active Learning Based Molecule Virtual Screening. arXiv, 20 Sept. 2023, arXiv:2309.11687. https://doi.org/10.48550/arXiv.2309.11687.  
Crooke, Stanley T., et al. Antisense Technology: A Review. Journal of Biological Chemistry, vol. 296, 2021, article no. 100416, https://doi.org/10.1016/j.jbc.2021.100416.  
  
Guo, Taicheng, et al. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Track on Datasets and Benchmarks, 2023. <https://github.com/ChemFoundationModels/ChemLLMBench>.

Hwang, Gyeongjo, et al. ASOptimizer: Optimizing Antisense Oligonucleotides Through Deep Learning for IDO1 Gene Regulation. Molecular Therapy: Nucleic Acids, vol. 35, June 2024, <https://doi.org/10.1016/j.omtn.2024.102186>.  
  
Li, Zhao, et al. "Ensemble Pretrained Language Models to Extract Biomedical Knowledge from Literature." Journal of the American Medical Informatics Association, vol. 31, no. 9, 2024, pp. 1904–1911, <https://doi.org/10.1093/jamia/ocae061>.  
  
Sadeghi, Shaghayegh, et al. Can Large Language Models Understand Molecules? BMC Bioinformatics, vol. 25, no. 225, 2024, https://doi.org/10.1186/s12859-024-05847-x.  
  
Sciabola, Simone, et al. PFRED: A Computational Platform for siRNA and Antisense Oligonucleotides Design. PLoS ONE, vol. 16, no. 1, 2021, e0238753. <https://doi.org/10.1371/journal.pone.0238753>.  
  
Tang, Xiangru, et al. "MolLM: A Unified Language Model for Integrating Biomedical Text with 2D and 3D Molecular Representations." Bioinformatics, vol. 40, suppl. 1, 2024, pp. i357–i368, <https://doi.org/10.1093/bioinformatics/btae260>.  
  
Zambrano Chaves, Juan Manuel, et al. "Tx-LLM: A Large Language Model for Therapeutics." arXiv, 10 June 2024, <https://arxiv.org/abs/2406.06316>.  
  
https://osu-nlp-group.github.io/LLM4Chem/ - reusable pipeline and open-access dataset with LlaSMol