Molecular Cues to Smart Sequences: Optimizing Early Round SELEX Sequences

Soniya [1] **Runjhun Saran Narayan** [2 3 4] **Nived Ambadipudi** [2] **Apurva Narayan** [1]

Abstract

Rapid discovery of high-affinity aptamers is often hindered by low enrichment and high sequence diversity in early SELEX rounds, leading to prolonged experimental cycles. To overcome this challenges, we present a ligand- and structure-aware refinement framework designed to enhance the quality of aptamer candidates. Our approach leverages *TxGemma*, a generative model conditioned on the ligand's (small molecule target) molecular descriptors and the aptamer's predicted secondary structure. *TxGemma* selectively edits its hairpin loops and stems of early-round candidates to enhance the chance of molecular interactions with the small molecule ligand. Using Theophylline as a test case, we show that candidate sequences from early round (Round 10), when modified, align closely with the enriched sequences in later SELEX rounds (Round 16–20). This demonstrates the model's ability to emulate evolutionary convergence and accelerate aptamer discovery with reduced experimental effort. This work provides a robust platform for in-depth exploration and generation of high-affinity aptamers, potentially eliminating the need for further SELEX rounds, and accelerate aptamer discovery with reduced experimental effort.

1. Introduction

Aptamers are short, single-stranded DNA or RNA sequences capable of binding diverse molecular targets- including small molecules, peptides, proteins, and whole cells with high specificity and affinity. Their structural flexibility, low immunogenicity, thermal stability, and ease of chemical synthesis have positioned them as promising alternatives to antibodies for a broad range of applications, including therapeutics, diagnostics, and biosensing platforms (Zhang et al., 2019; Keefe et al., 2010).

The primary method for aptamer discovery remains Systematic Evolution of Ligands by EXponential enrichment (SELEX). SELEX is an iterative process involving incubation of candidate DNA sequences with the target, partitioning of bound from unbound sequences, and amplification of the bound sequences, until the populaion of high affinity aptamers enriches(Tuerk & Gold, 1990; Ellington & Szostak, 1990a). While SELEX has yielded numerous functional DNA aptamers, it is inherently labor-intensive, time-consuming, and poorly scalable. A typical experiment may require 8–20 rounds of selection to achieve sufficient enrichment of aptamer sequences, with performance often sensitive to PCR bias, nonspecific amplification, and target complexity. Moreover, SELEX provides little insight into the biochemical mechanisms underlying aptamer-target binding and offers limited control over sequence–structure–function relationships.

2. Research Gap

In recent years, a variety of computational methods have been developed to accelerate aptamer discovery, including motif identification from enriched SELEX pools (Hoon et al., 2011), machine learning models for affinity prediction (Chen et al., 2022), and inverse folding algorithms that generate sequences compatible with target secondary structures (Reuter & Mathews, 2010). Although these approaches have shown success, they typically rely on data from late-stage SELEX rounds, where aptamer pools are already highly enriched. This narrow focus overlooks the vast and diverse sequence space in early-round SELEX libraries, where many promising binders remain weakly enriched and structurally heterogeneous. The low signal-to-noise ratio in these early datasets presents a significant challenge for conventional models, which often fail to generalize in such noisy regimes (Gioacchino et al., 18).

A critical challenge lies in identifying high-affinity binders

[1] University of Western Ontario, London, ON, Canada [2] MOLwise Biosciences Inc, Kitchener, ON, Canada [3] University of Waterloo, Waterloo, ON, Canada [4] The University of British Columbia, Kelowna, BC, Canada. Correspondence to: apurva.narayan <apurva.narayan@uwo.ca>.

Proceedings of the Workshop on Generative AI for Biology at the 42^{nd} *International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

during early selection, which leads to prolonged experimental cycles and missed opportunities to exploit potentially functional candidates. Existing computational approaches often rely solely on sequence-level features and remain agnostic to the molecular properties of the target, making them poorly suited for guiding early-stage selection or enabling cross-target design (Wang et al., 2024; Gioacchino et al., 18; Bashir et al., 2021). Therefore, a computational framework that can integrate the existing knowledge about aptamer design, early-round SELEX sequence data, and target molecule properties can greatly advance this field by guiding aptamer optimization from early SELEX data.

To address this challenge, we introduce a ligand- and structure-aware generative framework that directly refines early-round aptamer sequences using the exisiting knowledge about various molecular interactions known of DNA aptamers with small molecule ligands, ligand properties, and aptamer secondary strcutures. Central to this framework is TxGemma (Wang et al., 2025), a generative model that attempts to overcome key limitations of SELEX by proposing high-affinity candidate aptamer sequences. By suggesting modifications based on secondary-structure features, TxGemma attempts to eliminates the need for late-round SELEX experimentation. This molecular-driven approach provides a basic framework that can serve as a platform which in future can accelerates aptamer discovery and offer a scalable, generalizable alternative to traditional sequence-based models. Future work may include improvization of model generalization, expansion of the molecular diversity of training data, and refinement of sequence optimization strategies.

3. The Proposed Approach

In this work, we propose a ligand- and structure-aware framework for aptamer sequence refinement using the TxGemma model. TxGemma operates on early-round SELEX candidate sequences which are often weakly enriched and structurally diverse. It suggests modifications in the DNA sequence of of the early-round DNA sequence in order to increase their probability of being a high-affinity aptamer. The model also provides a rationale for the modifications it has suggested. These rationales are inspired by the existing knowledge regarding a few known aptamer-ligand pairs, generally observed sequence design and secondary structure characteristics of known aptamers, generally observed molecular interactions between known aptamers and their ligands, and small molecule ligand chemical structures.

TxGemma incorporates functional group information of the ligand (deciphered from the SMILES codes of the ligand) alongside the predicted secondary structure of the sequence, represented in dot-bracket notation. This structural representation highlights regions such as hairpin loops and stems that are typically involved in target recognition by the aptamers. To navigate the sequence optimization process, the model employs a *few-shot learning approach*, using examples of known sequence–ligand pairs to identify effective design patterns. In addition, the model applies explicit *chain-of-thought* reasoning to justify each proposed sequence modification, ensuring that a precise, stepwise analysis of sequence features and structure supports every design decision. By combining functional group information, secondary structure modeling, and reasoning-based learning, the model attempts to enable the rational design of sequences that are both structurally informed and generalizable, moving beyond traditional sequence-based design methods.

By operating directly on early-round sequences and conditioning on chemically relevant information, the proposed approach attempts to minimize the need for multiple rounds of SELEX enrichment, thereby intending to reduce time and experimental overhead. Its structure-aware design allows for targeted optimization based on molecular properties and secondary structure, improving the likelihood of identifying high-affinity binders. This framework offers a platform to build a more scalable, systematic, and biochemically guided strategy for next-generation aptamer discovery as compared to conventional models which depend on late-stage data and sequence-only features.

4. Simulation

To systematically guide optimization of DNA sequences for enhanced small-molecule binding, we employed a structured large language model (LLM) prompt designed to simulate expert-level reasoning in aptamer design. The prompt instructed the model to analyze a given DNA sequence and propose targeted sequence modifications aimed at improving binding affinity toward the specified ligand. It provided essential background on sequence structural features, including sequence motifs, secondary structure, loop and stem characteristics, and ligand-related properties such as molecular structure in the form of SMILES notation and related chemical attributes.

The prompt also explicitly outlined the molecular interactions known to influence aptamer–ligand binding, including:

- *π–π stacking*: Interactions between aromatic nucleobases (adenine, guanine) and ligand aromatic or planar regions (Meyer et al., 2003).
- *Hydrogen bonding*: Interactions involving hydrogen bond donors and acceptors from nucleobases and ligand functional groups, such as hydroxyl, amine, carbonyl, or heterocyclic groups (Anslyn & Dougherty, 2005; Alberts et al., 2015).

- *Electrostatic complementarity*: Interactions between the negatively charged DNA backbone and positively charged ligand groups, including amines and metal cations (Dougherty, 1997).
- *Hydrophobic interactions*: Nonpolar contacts between nucleobase rings or methyl groups and hydrophobic regions of the ligand, such as alkyl chains or nonpolar aromatic moieties (Pratt, 2002).
- *Intercalation and base-stacking insertion*: Insertion of flat ligand molecules between DNA base pairs, stabilized by stacking and van der Waals forces (Zhou, 2014).

In addition, the prompt incorporated general principles from the aptamer design literature to guide sequence optimization, including:

- Hairpin loops are key regions for ligand-binding interactions (Luo et al., 2019; Jeddi & Saiz, 2017)
- Hairpin loop lengths of approximately 6–12 nucleotides provide structural flexibility and facilitate target recognition (Jeddi & Saiz, 2017).
- Hairpin stems of approximately 3–6 base pairs offer a balance between stability and conformational flexibility for effective ligand binding (Jeddi & Saiz, 2017).
- GC-rich stems increase structural stability but may limit flexibility (Matsunaga et al., 2016).
- AT-rich stems enhance structural flexibility, potentially improving ligand accessibility (Jeddi & Saiz, 2017; Matsunaga et al., 2016).
- Typical aptamer scaffolds are approximately 30–50 nucleotides in length, supporting stable yet adaptable folding for target interaction (Ramachandran & Slack, 2013; Keefe et al., 2013; Ellington & Szostak, 1990b).

Together, these interaction types and design principles provided the structural and chemical context for guiding sequence optimization within the prompt.

The task required the model to propose a defined number of optimal sequence modifications and to generate corresponding modified sequences. Each modification was to be accompanied by a rationale explaining how the change was expected to enhance sequence–ligand interactions, drawing on known principles of nucleic acid folding and molecular recognition. Additionally, the prompt incorporated example cases of known aptamer–ligand interactions to provide contextual guidance, following a few-shot learning setup, and explicitly encouraged stepwise, chain-of-thought reasoning to guide the model's design process. Through its design, the prompt fostered a logical and methodical approach in the model's outputs, supporting the targeted optimization of sequence variants for improved binding.

In this study, we use the small molecule Theophylline (Tobia et al., 2023). The corresponding experimental results are discussed in Section 5.

5. Results

Hairpin Loop length and composition: It is well known that in DNA aptamers, the hairpin loops are most significant for ligand-binding interactions. Aptamer hairpin loops are central to recognizing small molecules, because their 6–12 nucleotide length achieves an ideal balance: enough flexibility to fold and adapt around the target, yet structured enough to form a defined binding pocket. Beyond loop length, the specific nucleobase makeup drives precise interactions with the target. For e.g. Adenine and guanine residues provide extensive pi-surfaces for stacking with aromatic parts of the ligand, plus multiple N–H and carbonyl sites for hydrogen bonding and electrostatic complementarity. Our model, TxGemma, has successfully integrated these biochemical insights. Given information from theophylline's SMILES descriptor, TxGemma identifies the design principles underlying naturally evolved aptamers. By proposing edits that fine-tune both size and composition of loops and stems, TxGemma generates sequences that are more likely to fold into effective binding pockets and interact specifically with the small-molecule target. Some results for early-round theophylline SELEX sequences are shown below: 1) For early-round theophylline SELEX sequences 5'AAATGCAATGTCCTGAGAATTCTGAAGGCT 3' and 5'GAAGAGCATTATGCCGAGATTTAGCGAAAA 3' with loop shown in red, our model recognizes that the loop is too short and will decrease structural flexibility and well as reduce functional group diversity for target binding. The model suggests to 'increase hairpin loop size to 7 bases' and 'extend the loop by inserting several additional bases' respectively. It rationales these modifications saying 'The extended loop provides more space for interactions with theophylline, and adopt favorable conformations, potentially enhancing binding affinity'. 2) For sequences 5'CATGAGGTATGAAGTCTTCGTAAGAGTTTG 3' (loop G-C content 0.25) and 5'AAAGTGCAACGTTC-GAGCAATTCTCGACTT 3' (loop purine fraction 0.4), our model suggests substitution/insertion modifications such as 'increase the G-C content of the hairpin loop' and 'incorporate purine residues within the loop' or 'substitute a Thymine with Guanine' respectively. It rationales the modifications by suggesting 'it enhances hydrogen bonding and pi-pi stacking interactions with the purine-based aromatic theophylline' and 'added purine base can participate in pi-pi stacking interactions with aromatic system of theophylline, increasing binding stability’.

Hairpin Stem length and composition: In DNA aptamers, the stem region provides structural integrity and defines the orientation of the loop for effective target interaction. It has been frequently observed that the Optimal stem length for small-molecule aptamers typically falls between 4–8 base pairs. Stems shorter than 4bp may be too unstable to support loop structure, while stems longer than 8 bp can be overly rigid, preventing needed conformational changes upon ligand binding. The base composition of the stem is equally critical. A moderate GC content (0.4 - 0.6) balances duplex stability with dynamic flexibility. Too many GC pairs can overly stabilize the stem, locking the loop into unproductive conformations. Conversely, low GC content can undermine the structural scaffold altogether. By regulating both length and GC content, stems become robust frameworks that both present the loop in the correct spatial context and allow necessary local flexibility when a small molecule binds. Our model, TxGemma, incorporates these biochemical principles, suggesting adjustments to stem pairing strength and composition to optimize aptamer folding and loop access in early rounds. Some results for early-round theophylline SELEX sequences are shown below: 1) Here is a sequence 5’CAAAGGTGTCCAGTACGTATCAGGTCTATG 3’ with only 3 base-pairs interspersed with bulges and secondary structure: ’...........((..(......)..))...’ (the dots denote unpaired bases while the brackets denote base-pairs). Our model recognized that the due to the presence of interspered base-pairs, the stem is unstable and suggested modifications such as ’introduce a GC base pair in the hairpin stem to increase structural stability’ and gave the rationale as ’Increasing structural stability through GC base pairing can create a stronger scaffold for the aptamer, potentially enhancing its binding affinity to Theophylline’. Note that the model suggests the addition of a G-C base pair and not an A-T base pair, indicating that it understands that due to 3 hydrogen bonds and stronger stacking, a GC base pair enhances the stem stability. 2) In the sequence 5’AGGAGCGGT-CAACGTTTCAGTTGTCTTCTG 3’ there are 6 contiguous base-pairs with secondary structure:’.((((((.....))))))............’. Here, our model suggests to ’Shorten the stem by one base pair’ based on the understanding that 6 consecutive base pairs form a highly stable but possibly over-rigid structure and a delicate balance between rigidity and flexibility is key for aptamer functionality. It rationales the suggested modification as follows: ’reducing stem length increases loop flexibility, potentially allowing for more dynamic interactions with the target’.

Together, TxGemma’s suggested loop and stem edits attempt to create finely balanced hairpin scaffolds that are conformation-wise and interaction-wise optimized and for high-affinity theophylline binding.

6. Conclusion

In this work, we present a structure- and ligand-aware framework for aptamer sequence refinement that bridges early-round SELEX data with biochemical features of the target molecule. Our approach attempts to enable rational, context-driven sequence optimization by integrating detailed molecular descriptors e.g. ligand functional groups, ligand aromaticity, aptamer secondary structures, etc. Three layers of contextual information guide the generation strategy: 1) a functional group summary of the ligand and secondary structure characteristics of the DNA sequence to me modified, 2) molecular interactions generally observed between knwon aptamer-ligand pairs, as well as 3) generally observed known aptamer sequence features. This allows the model to suggest precise base-level modifications, particularly in hairpin loop and hairpin stem regions. Overall, this work contributes a scalable and chemically informed strategy for aptamer design that may reduce dependence on labour-intensive SELEX cycles. In future work, we aim to extend this framework to support cross-target generalization and integrate 3D structure prediction or docking-based feedback to refine sequence selection.

Impact Statement

This work contributes to advancing machine learning applications in aptamer discovery, with potential impact on diagnostics, biosensing, and targeted therapeutics. By enabling data-efficient refinement of early-stage aptamers, our approach may reduce experimental costs and accelerate the development of molecular tools for healthcare and research. While we do not foresee direct negative impacts, we acknowledge that misuse or over-reliance without experimental validation could lead to unintended consequences, as with all generative models in biomedical applications. We encourage responsible use in conjunction with experimental oversight.

Acknowledgements

We would like to thank Prof. Juewen Liu, University of Waterloo, Waterloo, ON Canada N3L 3G1 for providing Aptamer SELEX Next-Generation Sequencing (NGS) data for target Theophylline.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*. Garland Science, 6th edition, 2015.

Anslyn, E. V. and Dougherty, D. A. *Modern Physical Organic Chemistry*. University Science Books, 2005.

Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., Davis, G., Gong, Q., Armstrong, Z., Jang, J., Kang, H., Pawlosky, A., Scott, A., Dahl, G. E., Berndl, M., Dimon, M., and Ferguson, B. S. Machine learning guided aptamer refinement and discovery. *Nature Communications*, 12(2366), 2021. doi: https://doi.org/10.1038/s41467-021-22555-9.

Chen, X., Zhao, X., and Yang, Z. Aptasensors for the detection of infectious pathogens: design strategies and point-of-care testing. *Microchim Acta*, 189, 443, 2022. doi: https://doi.org/10.1007/s00604-022-05533-w.

Dougherty, D. A. The cation– interaction. *Chemical Reviews*, 97(5):1305–1326, 1997. doi: 10.1021/cr9603744.

Ellington, A. D. and Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346: 818–822, 1990a. doi: https://doi.org/10.1038/346818a0.

Ellington, A. D. and Szostak, J. W. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *Science*, 249(4968): 505–510, 1990b. doi: 10.1126/science.1696025.

Gioacchino, A. D., Procyk, J., Molari, M., Schreck, J. S., Zhou, Y., Liu, Y., Monasson, R., Cocco, S., and Šulc, P. Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. *PLoS Comput Biol*, 9(e1010561), 18. doi: https://doi.org/10.1371/journal.pcbi.1010561.

Hoon, S., Zhou, B., Janda, K. D., Brenner, S., and Scolnick, J. Aptamer Selection by High-Throughput Sequencing and Informatic Analysis. *BioTechniques*, 51(6):413–416, 2011. doi: https://doi.org/10.2144/000113786.

Jeddi, I. and Saiz, L. Three-dimensional modeling of single-stranded dna hairpins for aptamer-based biosensors. *Scientific Reports*, 7:1178, 2017. doi: 10.1038/s41598-017-01348-5.

Keefe, A., Pai, S., and Ellington, A. Aptamers as therapeutics. *Nat Rev Drug Discov*, 9:537–550, 2010. doi: https://doi.org/10.1038/nrd3141.

Keefe, A. D., Pai, S., and Ellington, A. Aptamers as therapeutics. *Nature Reviews Drug Discovery*, 12:13–27, 2013. doi: 10.1038/nrd3856.

Luo, F. n. et al. Free solution assay signal modulation in variable-stem-length dna aptamer for tenofovir. *ACS Omega*, 2019. doi: 10.1021/acsomega.9b04341.

Matsunaga, K., Kimoto, M., et al. Architecture of high-affinity unnatural-base dna aptamers toward pharmaceutical applications. *Scientific Reports*, 5:18478, 2016. doi: 10.1038/srep18478.

Meyer, E. A., Castellano, R. K., and Diederich, F. Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie International Edition*, 42(11):1210–1250, 2003. doi: 10.1002/anie.200390319.

Pratt, L. R. Molecular theory of hydrophobic effects: "she is too mean to have her name repeated.". *Annual Review of Physical Chemistry*, 53:409–436, 2002. doi: 10.1146/annurev.physchem.53.090401.093500.

Ramachandran, S. and Slack, F. J. Aptamers: Molecules of great potential. *Journal of Bioscience*, 38:507–517, 2013. doi: 10.1007/s12038-013-9323-5.

Reuter, J. S. and Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(129), 2010. doi: https://doi.org/10.1186/1471-2105-11-129.

Tobia, J. P., Huang, P.-J. J., Ding, Y., Narayan, R. S., Narayan, A., and Liu, J. Machine learning directed aptamer search from conserved primary sequences and secondary structures. *ACS Synthetic Biology*, 12 (1):186–195, 2023. doi: 10.1021/acssynbio.2c00462. URL `https://doi.org/10.1021/acssynbio.2c00462`. PMID: 36594697.

Tuerk, C. and Gold, L. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *Science*, 249(4968):505–510, 1990. doi: 10.1126/science.2200121. URL `https://www.science.org/doi/abs/10.1126/science.2200121`.

Wang, E., Schmidgall, S., Jaeger, P. F., Zhang, F., Pilgrim, R., Matias, Y., Barral, J., Fleet, D., and Azizi, S. TxGemma: Efficient and Agentic LLMs for Therapeutics. *ArXiv*, 2504.06196,, 2025.

Wang, Z., Liu, Z., Zhang, W., Li, Y., Feng, Y., Lv, S., Diao, H., Luo, Z., Yan, P., He, M., and Li, X. AptaDiff: de novo design and optimization of aptamers based on diffusion models. *Briefings in Bioinformatics*, 25(6):bbae517, 10 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae517. URL `https://doi.org/10.1093/bib/bbae517`.

Zhang, Y., Lai, S. B., and Juhas, M. Recent advances in aptamer discovery and applications. *Molecules*, 24(5), 2019. ISSN 1420-3049. doi: 10.3390/molecules24050941. URL `https://www.mdpi.com/1420-3049/24/5/941`.

Zhou, e. a. Dna intercalation optimized by two-step molecular lock mechanism. *Scientific Reports*, 4:37993, 2014. doi: 10.1038/srep37993.