Empowering AI in RNAi Therapeutics: A Foundational Dataset for siRNA Design and Optimization

Xin Guo
Shanghai Academy of Artificial Intelligence for Science
Artificial Intelligence Innovation and Incubation Institute, Fudan University
guoxin@sais.org.cn

Jiyang Li
Shanghai Academy of Artificial Intelligence for Science
lijiyang@sais.org.cn

1 Motivation and Bottleneck

RNA interference (RNAi) therapeutics represent an innovative class of gene-silencing medicines predicated upon the revolutionary, Nobel Prize-winning discovery of RNAi—a conserved cellular pathway that regulates gene expression via mechanisms such as post-transcriptional messenger RNA (mRNA) degradation, particularly through the mediation of small interfering RNA (siRNA)[1]. siRNA facilitates precise, sequence-dependent recognition and endonucleolytic cleavage of aberrant mRNAs or non-coding RNAs implicated in various diseases, thereby facilitating targeted therapeutic interventions. These siRNA-based modalities demonstrate superior target specificity, which supports the modulation of conventionally undruggable molecular entities underlying genetic pathologies while attenuating off-target perturbations. In addition, they exhibit favorable safety attributes and can augment conventional regimens for neoplasms, viral pathologies, and diverse indications[2].

Notwithstanding these merits, the rational engineering of siRNAs poses a principal obstacle in the progression of RNAi therapeutics, encompassing hurdles in delivery, stability, and mitigation of unintended effects. While early public repositories contained limited siRNA entries, contemporary innovations have substantially broadened siRNA utilization, culminating in clinical substantiation and extensive deployment of novel medicines. For instance, landmark approvals of drugs like Amvuttra (Patisiran) and Leqvio (Inclisiran) have dramatically augmented the therapeutic landscape for siRNAs and affirmed their clinical efficacy and market feasibility. Chemical modifications remain indispensable for optimizing pharmacokinetic profiles and curtailing immunostimulatory responses; however, the principles dictating their impacts are inadequately elucidated[3]. The current data landscape—composed of limited, publicly accessible datasets—prevents researchers from answering key scientific questions about how sequence, modification, and experimental context jointly determine silencing efficacy. Without a sufficiently rich and diverse foundational dataset, AI-guided siRNA design cannot achieve its full potential.

2 Dataset Description

To address pivotal deficiencies in siRNA research, we propose the development of a comprehensive, openly accessible dataset comprising chemically modified and experimentally validated siRNAs. Although extant public siRNA databases are constrained in scope—particularly lacking dedicated repositories for chemically modified variants—we aim to surmount these limitations through the systematic integration and enrichment of siRNA data across diverse dimensions. This cohesive

Published in the AI for Science workshop (NeurIPS 2025)

methodology will enable the assimilation of extensive information, encompassing core sequence details, intricate chemical modification profiles, and robust experimental metadata, including cell lines, transfection techniques, siRNA concentrations, and exposure durations. We will curate this multidimensional dataset from publicly available archives, such as patents, scholarly literature, and FDA documentation, while pursuing avenues for cross-company and cross-agency data collaboration. By linking these different data dimensions, we can provide a more complete picture of an siRNA's behavior. The dataset would include both in vitro assays and in vivo studies, providing a comprehensive representation of siRNA efficacy across different experimental contexts. A key aspect of our approach is the development of a specialized AI agent, an automated system designed to continuously and intelligently retrieve, process, and update siRNA data from disparate sources. This would transform a largely underutilized collection of information into a dynamic, living resource. The raw data we collect often contains significant noise and inconsistencies. A crucial part of our effort will be to dedicate time to data standardization and normalization, ensuring that all efficacy values are harmonized to enable cross-comparison and computational modeling. Furthermore, we will establish a data grading system to classify entries based on the quality and completeness of their annotations. This will allow us to build a set of robust rules and guidelines, ensuring the immediate and AI-ready usability of the dataset for all researchers, from computational biologists to experimental scientists.

3 AI Tasks and Acceleration Potential

The proposed dataset would enable diverse AI-driven research. It would support predictive modeling of siRNA efficacy, allowing models to learn the complex relationships among sequence, chemical modifications, and experimental context across multiple cell lines and protocols. The rich annotations would also facilitate generative design, allowing models to propose novel siRNAs optimized for potency, specificity, and stability. The dataset's rich annotations would support multimodal learning and exploration of combinatorial design spaces. Additionally, this resource would provide a benchmark for sequence-to-function learning, supporting model evaluation, transfer learning, and explainable AI approaches. Beyond computational tasks, the dataset would accelerate experimental planning and therapeutic design by guiding target selection, chemical modification optimization, and transfection strategy.

4 Feasibility & Rough Budget and Long-Term Impact

We conducted a Proof of Concept (POC) test, estimating the cost of curating 1,000 siRNA data entries (including sequence and experimental data) to be approximately 14$. Based on this, we estimate that over one million data entries exist across various public archives from past years, with new data accumulating at an exponential rate in the future. We propose a hybrid pipeline combining automated text and sequence extraction with expert human curation to ensure accuracy and data integrity. This AI agent-driven process would allow us to significantly reduce manual labor and scale our efforts. By adopting a human-in-the-loop development logic, we can dramatically lower the total cost of curation, despite the massive and growing data volume. With the acceleration of population aging and changes in lifestyle, the incidence of chronic diseases continues to increase. The World Health Organization estimates that chronic noncommunicable diseases are responsible for 74% of global deaths each year [4]. In this context, nucleic acid drugs offer significant advantages over traditional small molecule and antibody drugs, including a broader range of potential targets, shorter development cycles, and superior targeting and specificity. By providing a comprehensive, high-quality dataset linking sequence, chemical modification, and experimental efficacy, it would enable AI models to accurately predict and design potent siRNAs. This capability would directly accelerate therapeutic development, supporting target selection, optimization of chemical modifications, and design of effective RNAi-based drugs. Beyond therapeutic applications, the dataset establishes a generalizable framework for sequence-to-function learning, offering a model for integrating industrial-scale experimental data into open resources. Its structured, high-resolution annotations could catalyze fundamental discoveries in molecular biology and chemical biology, by providing a large-scale view of how subtle changes in an RNA molecule impact its function in different biological contexts. This project would unlock previously inaccessible knowledge, fostering reproducibility, innovation, and rapid translation from computational prediction to laboratory validation across diverse scientific domains.

Acknowledgement

This work was supported by AI for Science Program, Shanghai Municipal Commission of Economy and Information.

References

[1] Hu, B., Zhong, L., Weng, Y. et al. Therapeutic siRNA: state of the art. Sig Transduct Target Ther 5, 101 (2020). https://doi.org/10.1038/s41392-020-0207-x

[2] Ahn, I., Kang, C.S. & Han, J. Where should siRNAs go: applicable organs for siRNA drugs. Exp Mol Med 55, 1283–1292 (2023). https://doi.org/10.1038/s12276-023-00998-y

[3] Shukla, S., Sumaria, C. S., & Pradeepkumar, P. I. (2010). Exploring chemical modifications for siRNA therapeutics: a structural and functional outlook. ChemMedChem, 5(3), 328–349. https://doi.org/10.1002/cmdc.200900444

[4] *Noncommunicable diseases progress monitor 2025*. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240105775.